CHAPTER 4

The Spatial Analyses of Data Fields

A fundamental problem in oceanography is how best to represent spatially distributed data (or statistical products computed from these data) in such a way that dynamical processes or their effects can best be visualized. As in most aspects of observational analysis, there has been a dramatic change in the approach to this problem due to the increased abundance of digital data and our ability to process them. Prior to the use of digital computers, data displays were constructed by hand and "contouring" was an acquired skill of the descriptive analyst. Hand contouring is still practiced today although, more frequently, the data points being contoured are averaged values produced by a computer. In other applications, the computer not only performs the averaging but also uses objective statistical techniques to produce both the gridded values and the associated contours.

The purpose of this section is to review data techniques and procedures designed to reduce spatially distributed data to a level that can be visualized easily by the analyst. We will discuss methods that address both spatial fields and time series of spatial fields since these are the primary modes of data distribution encountered by the oceanographer. Our focus is on the more widely used techniques which we present in a practical fashion, stressing the application of the method for interpretive applications.

4.1 TRADITIONAL BLOCK AND BULK AVERAGING

A common method for deriving a gridded set of data is simply to average the available data over an arbitrarily selected rectangular grid. This averaging grid can lie along any chosen surface but is most often constructed in the horizontal or vertical plane. Because the grid is often chosen for convenience, without any consideration to the sampling coverage, it can lead to an unequal distribution of samples per grid "box". For example, because distance in longitude varies as the cosine of the latitude, the practice of gridding data by 5 or 10° squares in latitude and longitude may lead to increasingly greater spatial coverage at low latitudes. Although this can be overcome somewhat by converting to distances using the central latitude of the box (Poulain and Niiler, 1989), it is easy to see that inhomogeneity in the sampling coverage can quickly nullify any of the useful assumptions made earlier about the Gaussian nature of sample populations or, at least, about the set of means computed from these samples. This is less of a problem with satellite-tracked drifter data since satellite ground tracks converge with increasing latitude, allowing the data density in boxes of fixed longitude length to remain nearly constant.

With markedly different data coverage between sample regions, we cannot always fairly compare the values computed in these squares. At best, one must be careful to consider properly the amount of data being included in such averages and be able to evaluate possible effects of the variable data coverage on the mapped results. Each value should be associated with a sample size indicating how many data points, N, went into the computed mean. This will not dictate the spatial or temporal distributions of the sample data field but will at least provide a sample size parameter which can be used to evaluate the mean and standard deviation at each point. While the standard deviation of each grid sample is composed of both spatial and temporal fluctuations (within the time period of the grid sample), it does give an estimate of the inherent variability associated with the computed mean value.

Despite the problems with nonuniform data coverage, it has proven worthwhile to produce maps or cross-sections with simple grid-averaging methods since they frequently represent the best spatial resolution possible with the existing data coverage. The approach is certainly simple and straightforward. Besides, the data coverage often does not justify more complex and computer-intensive data reduction techniques. Specialized block-averaging techniques have been designed to improve the resolution of the corresponding data by taking into account the nature of the overall observed global variability and by trying to maximize the coverage appropriately. For example, averaging areas are frequently selected which have narrow meridional extent and wide zonal extent, taking advantage of the stronger meridional gradients observed in the ocean. Thus, an averaging area covering 2° latitude by 10° longitude may be used to better resolve the meridional gradients which dominate the open ocean (Wyrtki and Meyers, 1975). This same idea may be adapted to more limited regions if the general oceanographic conditions are known. If so, the data can be averaged accordingly, providing improved resolution perpendicular to strong frontal features. A further extension of this type of grid selection would be to base the entire averaging area selection on the data coverage. This is difficult to formalize objectively since it requires the subjective selection of the averaging scheme by an individual. However, it is possible in this way to improve resolution without a substantial increase in sampling (Emery, 1983).

All of these bulk or block-averaging techniques make the assumption that the data being considered in each grid box are statistically homogeneous and isotropic over the region of study. Under these assumptions, area sample size can be based strictly on the amount of data coverage (number of data values) rather than having to know details about processes represented by the data. Statistical homogeneity does not require that all the data were collected by the same instrument having the same sampling characteristics. Thus, our grid-square averaging can include data from many different instruments which generally have the same error limits.

One must be careful when averaging different kinds of measurements, even if they are of the same parameter. It is very tempting, for example, to average mechanical bathythermograph (MBT) temperatures with newer expendable bathythermograph (XBT) temperatures to produce temperature maps at specific depths. Before doing so, it is worth remembering that XBT data are likely to be accurate to 0.1°C, as reported earlier, while MBT data are decidedly less accurate and less reliable. Another marked difference between the two instruments is their relative vertical coverage. While most MBTs stopped at 250 m depth, XBTs are good to 500–1800 m, depending on probe type. Thus, temperature profiles from MBTs can be expected to be different from those collected with XBTs. Any mix of the two will necessarily degrade the average to the quality of the MBT data and bias averages to shallow (< 300 m) depths. In some applications, the level of degraded accuracy will be more than adequate and it is only necessary to state clearly and be aware of the intended application when mixing the data from these instruments. Also, one can expect distinct discontinuities as the data make the transition from a mix of measurements at shallower levels to strictly XBT data at greater depth.

Other important practical concerns in forming block averages have to do with the usual geographic location of oceanographic measurements. Consider the global distribution of all autumn measurements up to 1970 of the most common oceanographic observation, temperature profiles (Figure 4.1.1). It is surprising how frequently these observations lie along meridians of latitude or parallels of longitude. This makes it difficult to assign the data to any particular 5 or 10° square when the border of the square coincides with integer values of latitude or longitude. When the latter occurs, one must decide to which square the borders will be assigned and be consistent in carrying this definition through the calculation of the mean values.

As illustrated by Figure 4.1.1, data coverage can be highly nonuniform. In this example, some areas were frequently sampled while others were seldom (or never) occupied. Such nonuniformity in data coverage is a primary factor in considering the representativeness of simple block averages. It certainly brings into question the assumptions of homogeneity (spatially uniform sampling distribution) and isotropy (uniform sampling regardless of direction) since the sample distribution varies greatly with location and may often have a preferred orientation. The situation becomes even more severe when one examines the quality of the data in the individual casts represented by the dots in Figure 4.1.1. In order to establish a truly consistent data set in terms of the quality of the observations (i.e. the depth of the cast, the number of samples, the availability of oxygen and nutrients, and so on), it is generally necessary to reject many of the available hydrographic casts.

The question of data coverage depends on the kind of scientific questions the data set is being asked to address. For problems not requiring high-quality hydrographic stations, a greater number of observations are available, while for more restrictive studies requiring a higher accuracy, far fewer casts would match the qualifications. This is also true for other types of historical data but is less true of newly collected data. However, even now, one must ensure that all observations have a similar level of accuracy and reliability. Variations in equipment performance, such as sensor response or failure, must be compensated for in order to keep the observations consistent. Also, changes in instrument calibration need to be taken into account over the duration of a sampling program. For example, transmissometer lenses frequently become matted with a biotic film that reduces the amount of light passing between the source and receiver lenses. A nonlinear, time-dependent calibration is needed to correct for this effect.

Despite the potential problems with the block-averaging approach to data presentation, much information can be provided by careful consideration of the data rather than the use of more objective statistical methods to judge data quality. The shift to statistical methods represents a transition from the traditional oceanographic efforts of the early part of the twentieth century when considerable importance was given to every measurement value. In those days, individual scientists were personally responsible for the collection, processing and quality of their data.

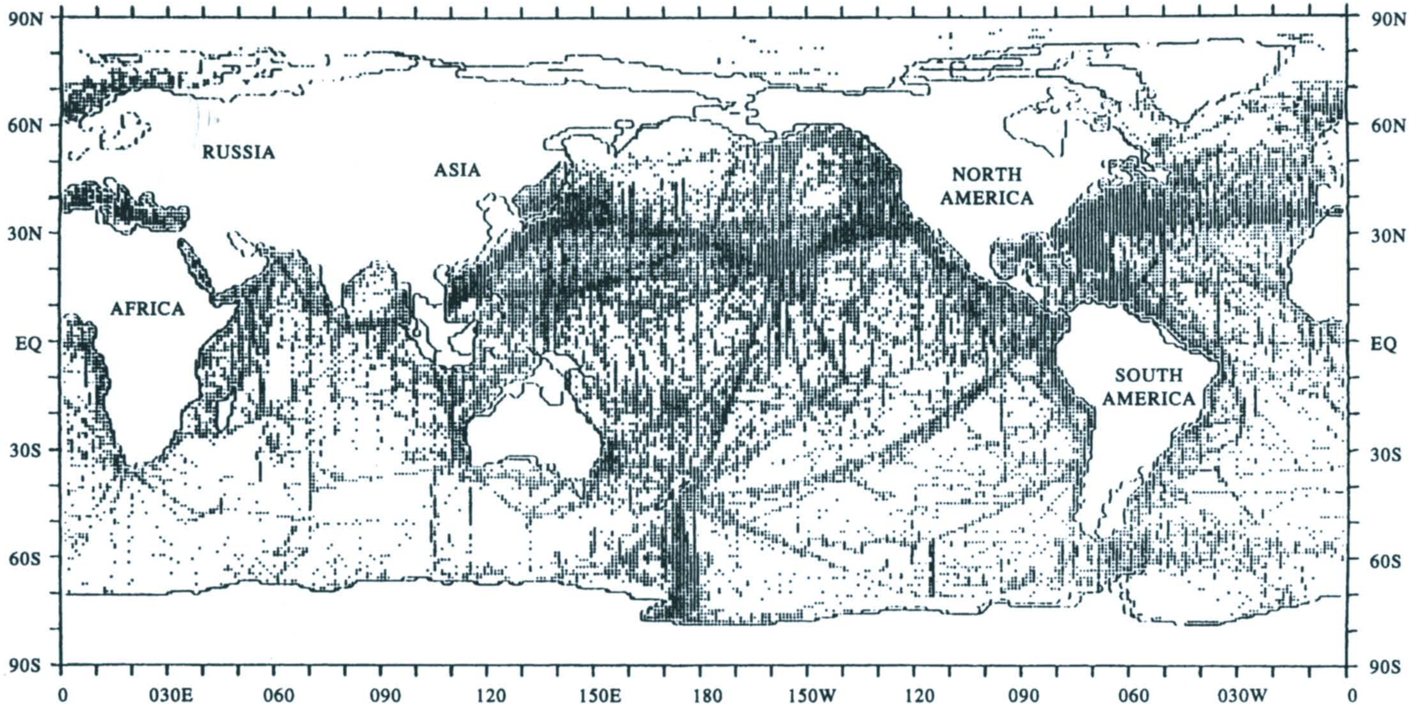

Figure 4.1.1. The global distribution of all temperature profiles collected during oceanographic surveys in the fall up to 1970. Sampling is most dense along major shipping routes.

Then, it was a simple task to differentiate between "correct" and "incorrect" samples without having to resort to statistical methods to indicate how well the environment had been observed. In addition, earlier investigations were primarily concerned with defining the mean state of the ocean. Temporal variability was sometimes estimated but was otherwise ignored in order to emphasize the mean spatial field. With today's large volumes of data, it is no longer possible to "hand check" each data value. A good example is provided by satellite-sensed information which generally consists of large groupings of data that are usually treated as individual data values.

In anticipation of our discussion of filtering in Chapter 5, we should point out that block averaging corresponds to the application of a box-car-shaped filter to the data series. This type of filter has several negative characteristics such as a slow filter roll off and large side lobes which distort the information in the original data series.

4.2 OBJECTIVE ANALYSIS

In a general sense, *objective analysis* is an estimation procedure which can be specified mathematically. The form of objective analysis most widely used in physical oceanography is that of least squares optimal interpolation, more appropriately referred to as *Gauss–Markov smoothing*, which is essentially an application of the linear estimation (smoothing) techniques discussed in Chapter 3. Since it is generally used to map spatially nonuniform data to a regularly spaced set of gridded values, Gauss–Markov smoothing might best be called "Gauss–Markov mapping". The basis for the technique is the Gauss–Markov theorem which was first introduced by Gandin (1965) to provide a systematic procedure for the production of gridded maps of meteorological parameters. If the covariance function used in the Gauss–Markov mapping is the covariance of the data field (as opposed to a more *ad hoc* covariance function, as is usually the case), then Gauss–Markov smoothing is optimal in the sense that it minimizes the mean square error of the objective estimates. A similar technique, called Kriging after a South African engineer H. G. Krige, was developed in mining engineering. Oceanographic applications of this method are provided by Bretherton *et al.* (1976), Freeland and Gould (1976), Bretherton and McWilliams (1980), Hiller and Käse (1983), Bennett (1992), and others.

The two fundamental assumptions in optimal interpolation are that the statistics of the subject data field are stationary (unchanging over the sample period of each map) and homogeneous (the same characteristics over the entire data field). A further assumption often made to simplify the analysis is that the statistics of the second moment, or covariance function, are isotropic (the same structure in all directions). Bretherton *et al.* (1976) point out that if these statistical characteristics are known, or can be estimated for some existing data field (such as a climatology based on historical data), they can be used to design optimum measurement arrays to sample the field. Since the optimal estimator is linear and consists of a weighted sum of all the observations within a specified range of each grid point, the objective mapping procedure produces a smoothed version of the original data field that will tend to underestimate the true field. In other words, if an observation point happens to coincide with an optimally interpolated grid point, the observed value and interpolated value will probably not be equal due to the presence of noise in the data. The degree of smoothing is determined by the characteristics of the signal and error covariance functions used in the mapping and increases with increasing spatial scales for a specified covariance function.

The general problem is to compute an estimate $\hat{D}(\mathbf{x}, t)$ of the scalar variable $D(\mathbf{x}, t)$ at a position $\mathbf{x} = (x, y)$ from irregularly spaced and inexact observations $d(\mathbf{x}_n, t)$ at a limited number of data positions $\mathbf{x}_n$ $(n = 1, 2, \ldots, N)$. Implementation of the procedure requires *a priori* knowledge of the variable's covariance function, $C(\mathbf{r})$, and uncorrelated error variance, ε, where $\mathbf{r}$ is the spatial separation between positions. For isotropic processes, $C(\mathbf{r}) \rightarrow C(r)$, where $r = |\mathbf{r}|$. Although specification of the covariance matrix should be founded on the observed structure of oceanic variables, selection of the mathematical form of the covariance matrix is hardly an "objective" process even with reliable data (cf. Denman and Freeland, 1985). In addition to the assumptions of stationarity, homogeneity, and isotropy, an important constraint on the chosen covariance matrix is that it must be positive definite (no negative eigenvalues). Bretherton *et al.* (1976) report that objective estimates computed from nonpositive definite matrices are not optimal and the mapping results are poor. In fact, nonpositive definite covariance functions can yield objective estimates with negative expected square error. One way to ensure that the covariance matrix is positive definite is to fit a function which results in a positive definite covariance matrix to the sample covariance matrix calculated from the data (Hiller and Käse, 1983). This results in a continuous mathematical expression to be used in the data weighting procedure. In attempting to specify a covariance function for data collected in continental shelf waters, Denman and Freeland (1985) further required that $\partial^2 C/\partial^2 x$ and $\partial^2 C/\partial^2 y$ be continuous at $r = 0$ (to ensure a continuously differentiable process) and that the variance spectrum, $S(k)$, derived from the transform of $C(\mathbf{r})$ be integrable and nonnegative for all wavenumbers, $\mathbf{k}$ (to ensure a realizable stochastic random process).

Calculation of the covariance matrix requires that the mean and "trend" be removed from the data (the trend is not necessarily linear). In three-dimensional space, this amounts to the removal of a planar or curvilinear surface. For example, the mean density structure in an upwelling domain is a curved surface which is shallow over the outer shelf and deepens seaward. Calculation of the density covariance matrix for such a region first involves removal of the curved mean density surface (Denman and Freeland, 1985). Failure to remove the mean and trend would not alter the fact that our estimates are optimal but it would redistribute variability from unresolved larger scales throughout the wavenumber space occupied by the data. We would then map features that have been influenced by the trend and mean.

As discussed later in the section on time series, there are many ways to estimate the trend. If ample good-quality historical data exist, the trend can be estimated from these data and then subtracted from the data being investigated. If historical data are not available, or the historical coverage is inadequate, then the trend must be computed from the sample data set itself. Numerous methods exist for calculating the trend and all require some type of functional fit to the existing data using a least-squares method. These functions can range from straight lines to complex higher-order polynomials and associated nonlinear functions. We note that, although many candidate oceanographic data fields do not satisfy the conditions of stationarity, homogeneity, and isotropy, their anomaly fields do. In the case of anomaly fields, the trend and mean have already been removed. Gandin (1965) reports that it may be possible to estimate the covariance matrix from existing historical data. This is more often the case in meteorology than in oceanography. In most oceanographic applications, the analyst must estimate the covariance matrix from the data set being studied.

In the following, we present a brief outline of objective mapping procedures. The interested reader is referred to Gandin (1965) and Bretherton *et al.* (1976) for further details. As noted previously, we consider the problem of constructing a gridded map of the scalar variable $D(\mathbf{x}, t)$ from an irregularly spaced set of scalar measurements $d(\mathbf{x}, t)$ at positions $\mathbf{x}$ and times t. The notation $\mathbf{x}$ refers to a suite of measurement sites, x_n ($n = 1, 2, \ldots$), each with distinct (x, y) coordinates. We use the term variable to mean directly measured oceanic variables as well as calculated variables such as the density or streamfunction derived from the observations. Thus, the data $d(\mathbf{x}, t)$ may consist of measurements of the particular variable we are trying to map or they may consist of some other variables that are related to D in a linear way. The former case gives

$$d(\mathbf{x}, t) = D(\mathbf{x}, t) + \varepsilon(\mathbf{x}) \tag{4.2.1}$$

where the ε are zero-mean measurement errors which are not correlated with the measurement D. In the latter case

$$d(\mathbf{x}, t) = F[D(\mathbf{x}, t)] + \varepsilon(\mathbf{x}) \tag{4.2.2}$$

in which F is a linear functional which acts on the function D in a linear fashion to give a scalar (Bennett 1992). For example, if $D(\mathbf{x}, t) = \Psi(\mathbf{x}, t)$ is the streamfunction, then the data could be current meter measurements of the zonal velocity field, $u(\mathbf{x}, t) = F[\Psi(\mathbf{x}, t)]$, where

$$d(\mathbf{x}, t) = u(\mathbf{x}, t) + \varepsilon(\mathbf{x}) = -\frac{\partial\Psi(\mathbf{x})}{\partial y} + \varepsilon(\mathbf{x}) \tag{4.2.3}$$

and $\partial\Psi/\partial y$ is the gradient of the streamfunction in the meridional direction.

To generalize the objective mapping problem, we assume that mean values have *not* been removed from the original data prior to the analysis. If we consider the objective mapping for a single "snap shot" in time (thereby dropping the time index, t), we can write linear estimates $\hat{D}(\mathbf{x})$ of $D(\mathbf{x})$ as the summation over a weighted set of the measurements d_i ($i = 1, \ldots, N$)

$$\hat{D}(\mathbf{x}) = \overline{D}(\mathbf{x}) + \sum_{i=1}^{N} b_i(d_i - \overline{\mathbf{d}}) \tag{4.2.4}$$

where the overbar denotes an expected value (mean), $d_i = d(\mathbf{x}) = d(x_i)$, $1 \leq i \leq N$ is shorthand notation for the data values, and the $b_i = b(\mathbf{x}) = b(x_i)$ are, as yet unspecified, weighting coefficients at the data points x_i. The selection of the N data values is made by restricting these values to some finite area about the grid point. The estimates of the parameters b_i in equation (4.2.4) are found in the usual way by minimizing the mean square variance of the error $e(\mathbf{x})^2$ between the measured variable, D, and the linear estimate, $\hat{D}$, at the data location. In particular,

$$\overline{e(\mathbf{x})^2} = \overline{\left[D(\mathbf{x}) - \hat{D}(\mathbf{x})\right]^2} \tag{4.2.5}$$

which on substitution of (4.2.4) yields

$$\overline{e(\mathbf{x})^2} = \overline{[D(\mathbf{x}) - \bar{D}(\mathbf{x})]^2} + \sum_{i=1}^{N}\sum_{j=1}^{N} b_i b_j \overline{(d_i - \bar{d})(d_j - \bar{d})} - 2\sum_{i=1}^{N} b_i \overline{(d_i - \bar{d})(D - \bar{D})} \tag{4.2.6}$$

Note, that if the mean has been removed, we can set $\bar{D}(\mathbf{x}) = \bar{d}(\mathbf{x}) = 0$ in (4.2.6). The mean square difference in equation (4.2.6) is minimized when

$$b_i = \sum_{j=1}^{N} \left\{ \left[\overline{(d_i - \bar{d})(d_j - \bar{d})}\right]^{-1} \overline{(d_j - \bar{d})(D - \bar{D})} \right\} \tag{4.2.7}$$

To calculate the weighting coefficients in (4.2.7), and therefore the grid-value estimates in (4.2.4), we need to compute the covariance matrix by averaging over all possible pairs of data taken at points x_i, x_j; the covariance matrix is

$$\overline{(d_i - \bar{d})(d_j - \bar{d})} = \overline{(d(x_i) - \bar{d})(d(x_j) - \bar{d})} \tag{4.2.8}$$

We do the same for the interpolated value

$$\overline{(d_i - \bar{d})(D_j - \bar{D})} = \overline{(d(x_i) - \bar{d})(d(x_k) - \bar{D})} \tag{4.2.9}$$

where x_k is the location vector for the grid point estimate $\hat{D}(x_k)$.

In general, we need a series of measurements at each location so that we can obtain statistically reliable expected values for the elements of the covariance matrices in (4.2.8) and (4.2.9). The expected values in the above relations could be computed as ensemble averages over spatially distributed sets of measurements. Typically, however, we have only one set of measurements for the specified locations x_i, x_j. As a consequence, we need to assume that, for the region of study, the data statistics are homogeneous, stationary and isotropic. If these conditions are met, the covariance matrix for the data distribution (for example, sea surface temperature) depends only on the distance r between data values, where $r = |x_j - x_i|$. Thus, we have elements i,j, of the covariance matrix given by

$$\begin{aligned} \overline{(d_i - \bar{d})(d_j - \bar{d})} &= C(|x_j - x_i|) + \overline{\varepsilon^2} \\ \overline{(d_i - \bar{d})(D_j - \bar{D})} &= C(|x_j - x_k|) + \overline{\varepsilon^2} \end{aligned} \tag{4.2.10}$$

where $C(|\mathbf{r}|) = \overline{d(\mathbf{x})d(\mathbf{x} + \mathbf{r})}$ is the covariance matrix and the mean square error $\varepsilon(\mathbf{x})^2$ implies that this estimate is not exact and there is some error in the estimation of the correlation function from the data. We note that this is not the same error in (4.2.6) that we minimize to solve for the weights in (4.2.7). The matrix can now be calculated by forming pairs of observed data values separated into bins according to the distance between sample sites, x_i. These are then averaged over the number of pairs that have the same separation distance to yield the product matrix

$$\overline{(d_i - \bar{d})(d_j - \bar{d})}$$

This computation requires us to define some "bin interval" for the separation distances so that we can group the product values together. To ensure that the resulting covariance matrix meets the condition of being positive definite, a smooth function satisfying this requirement can be fitted to the computed raw covariance function. This fitted covariance function is used for

$$\overline{(d_i - \bar{d})(D - \bar{D})}$$

and to calculate

$$\left[\overline{(d_i - \bar{d})(d_i - \bar{d})}\right]^{-1}$$

The weights b_i are then computed from (4.2.7). It is a simple process to then compute the optimal grid value estimates from (4.2.4). Note that, where the data provide no help in the estimate of D (that is, $\varepsilon(\mathbf{x}) \to \infty$), then $b_i = 0$ and the only reasonable estimate is $\hat{D}(\mathbf{x}) = \bar{D}$, the mean value. Similarly, if the data are error free (such that $\varepsilon(\mathbf{x})^2 \to 0$), then $\hat{D}(x_i) = D(x_i)$ for all x_i ($i = 1, \ldots, N$). In other words, the estimated value and the measured data are identical at the measurement sites (within the limits of the noise in the data) and the estimator interpolates between the observations.

The critical step in the objective mapping procedure is the computation of the covariance matrix. We have described a straightforward procedure to estimate the covariance matrix from the sample data. As with the estimate of the mean or overall trend, it is often possible to use an existing set of historical data to compute the covariance matrix. This is frequently the case in meteorological applications where long series of historical data are available. In oceanography, however, the covariance matrix typically must be computed from the sample data. Where historical data are available, it is important to recognize that using these data to estimate the covariance matrix for use with more recently collected data is tantamount to assuming that the statistics have remained stationary since the time that the historical data were collected.

Bretherton *et al.* (1976) suggest that objective analysis can be used to compute the covariance matrix. In this case, they start with an assumed covariance function, $\hat{F}$, which is then compared with a covariance function computed from data with a fixed distance x_o. The difference between the model $\hat{F}$ and the real F computed from the data is minimized by repeated iteration.

To this point, we have presented objective analysis as it applies to scalar fields. We can also apply optimal (Gauss–Markov) interpolation to vector fields. One approach is to examine each scalar velocity component separately so that for n velocity vectors we have $2n$ velocity components

$$d_r = u_1(\mathbf{x}_r); \quad d_{r+n} = u_2(\mathbf{x}_r) \tag{4.2.11}$$

where u_1 and u_2 are the x, y velocity components at x_r. If the velocity field is nondivergent, we can introduce a scalar streamfunction $\Psi(\mathbf{x})$ such that

$$u_1 = -\frac{\partial\Psi}{\partial y}; \quad u_2 = \frac{\partial\Psi}{\partial x} \tag{4.2.12}$$

and apply scalar methods to Ψ.

Once the optimal interpolation has been executed, there is a need to return to equation (4.2.6) to compute the actual error associated with each optimal interpolation. To this end, we note that we now have the interpolated data from (4.2.4). Thus, we can use $\hat{D}$ computed from (4.2.4) as the value for D in (4.2.6). The product in the last term of (4.2.6) is computed from the covariance in (4.2.9). In this way, it is possible to compute the error associated with each optimally interpolated value. Frequently, this error field is plotted for a specific threshold level, typically 50% of the interpolated values in the mapped field (see following examples). It is important to retain this error estimate as part of the optimal interpolation since it enables us to assess the statistical significance of individual gridded values.

4.2.1 Objective mapping: examples

An example of objective mapping applied to a single oceanographic survey is provided by the results of Hiller and Käse (1983). The data are from a CTD survey grid occupied in the North Atlantic about midway between the Azores and the Canary Islands (Figure 4.2.1). At each CTD station, the geopotential anomaly at 25 db (dBar) relative to the anomaly at 1500 db (written 25/1500 db) was calculated and selected as the variable to be mapped. The two-dimensional correlation function for these data is shown in three-dimensional perspective in Figure 4.2.2(a). A series of different correlation functions were examined and an isotropic, Gaussian function that was positive definite was selected as the best fit (Figure 4.2.2b). Using this covariance function, the authors obtained the objectively mapped 25/1500 db geopotential anomaly shown in Figure 4.2.3(a). Removal of a linear trend gives the objective map shown in Figure 4.2.3(b) and the associated RMS error field shown in Figure 4.2.3(c). Only near the outside boundaries of the data domain does the RMS error increase to around 50% of the geopotential anomaly field (Figure 4.2.3b).

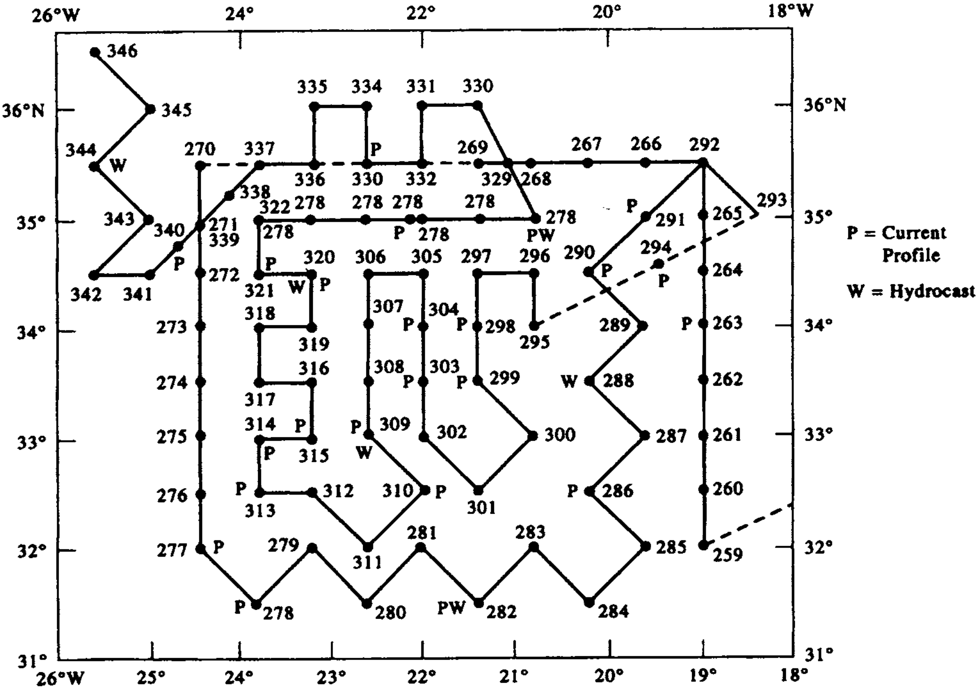

Figure 4.2.1. Locations of CTD stations taken in the North Atlantic between the Azores and the Canary Islands in spring 1982 (experiment POSEIDON 86, Hiller and Käse, 1983). Also shown are locations of current profile (P) and hydrocast (W) stations.

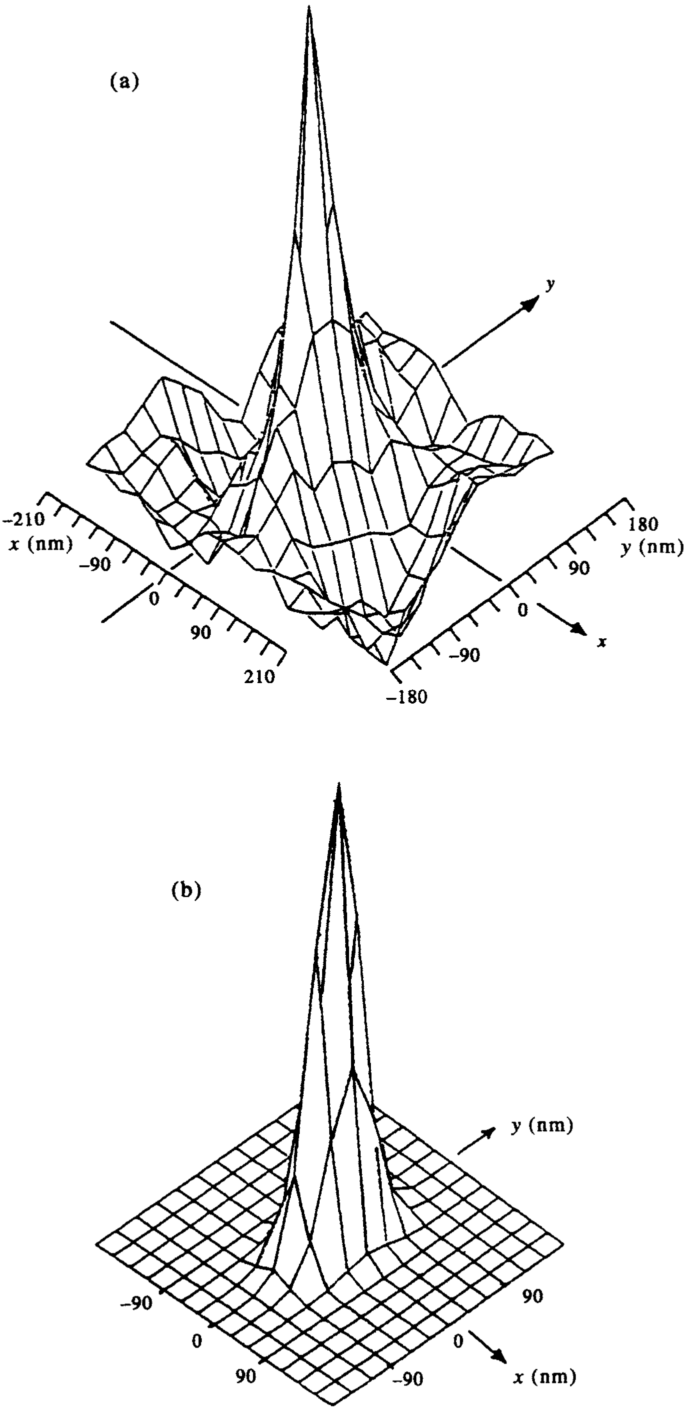

*Figure 4.2.2. The two-dimensional correlation function C(**r**) for the geopotential anomaly field at 25 db referenced to 1500 db (25/1500 dBar) for the data collected at stations shown in Figure 4.2.1 (1 db = 1 dBar = 1 m^2/s^2). Here, **r** = (x, y), where x, y are the eastward and northward coordinates, respectively. Distances are in nautical miles. (a) The "raw" values of C(**r**) based on the observations; (b) A model of the correlation function fitted to (a). (From Hiller and Käse, 1983).*

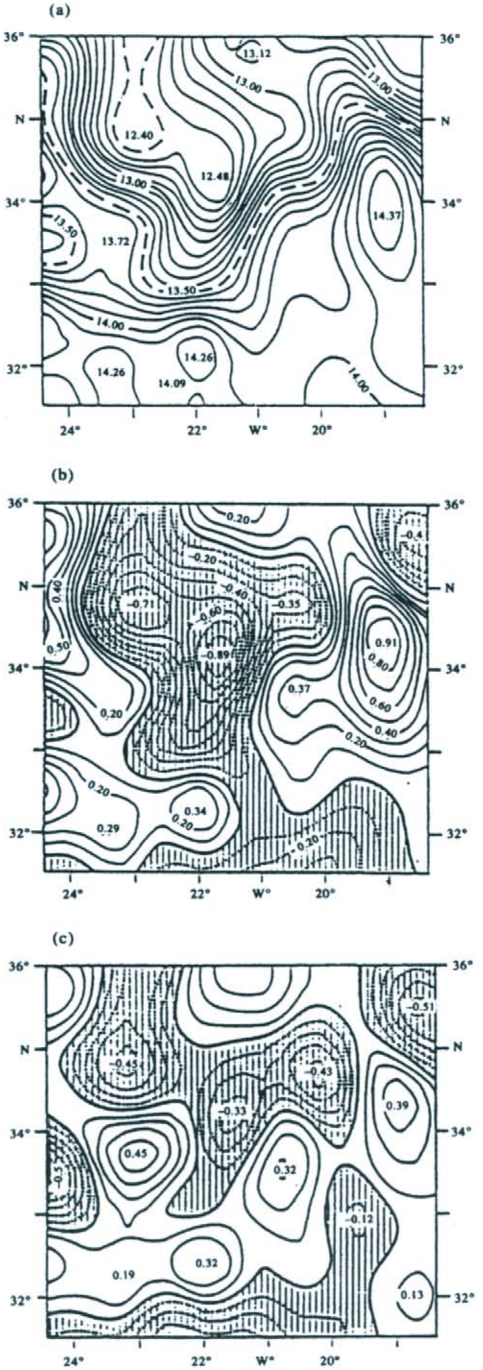

Figure 4.2.3. Objective analysis of the geopotential anomaly field 25/1500 db (m^2/s^2) using the correlation function in equation (4.2.2b). (a) The approximate center of the frontal band in this region of the ocean is marked by the 13.5 db isoline; (b) Same as (a) but after subtraction of the linear spatial trend; (c) Objective analysis of the residual mesoscale perturbation field 25/1500 dBar after removal of the composite mean field. (After Hiller and Käse, 1983.)

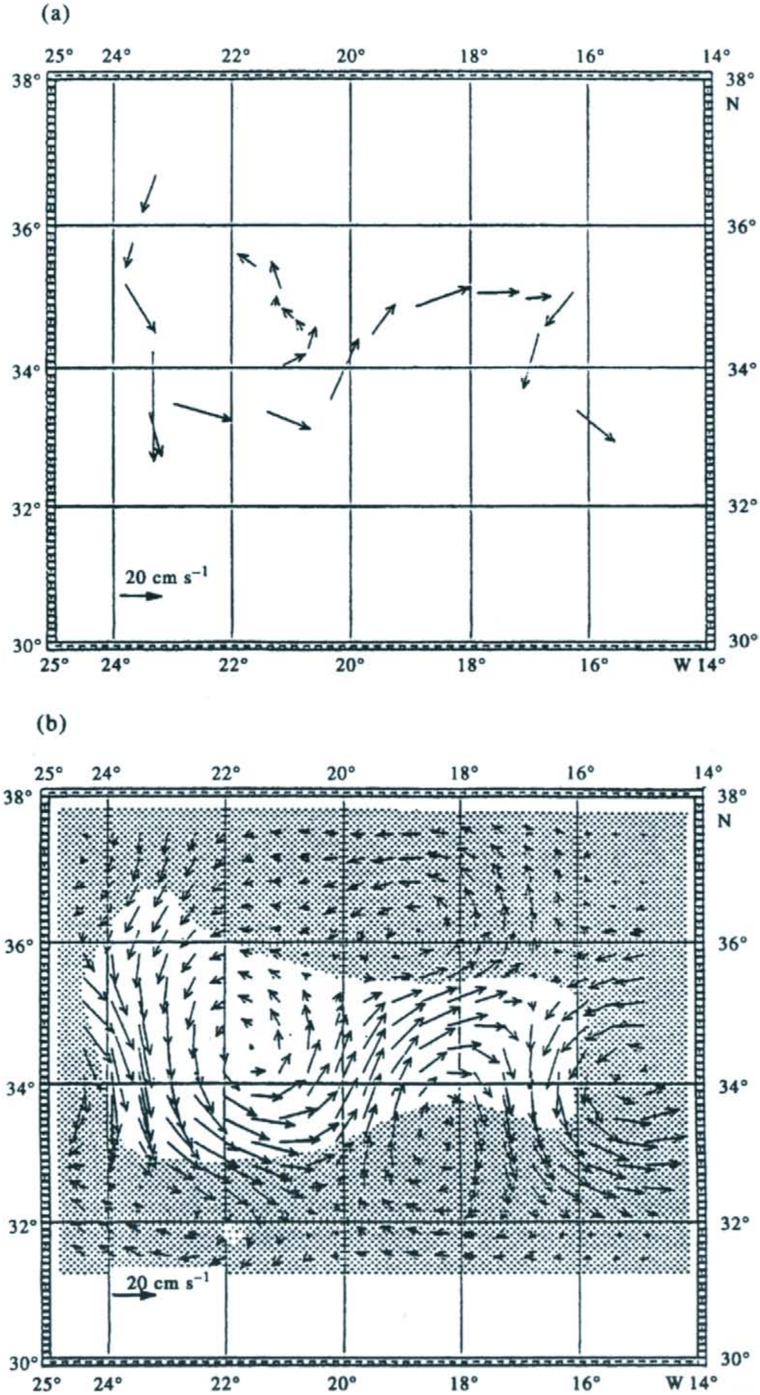

Figure 4.2.4. Analysis of the velocity field for the current profile collected on the grid in Figure 4.2.1. (a) The input velocity field; (b) Objective analysis of the input velocity field with correlation scale λ = 200 km and assumed noise variance of 30% of the total variance of the field. This approach treats mesoscale variability on scales less than 200 km as noise, which is smoothed out. In the shaded area, the error variance exceeds 50% of the total variance.

As an example of objective mapping applied to a vector field, Hiller and Käse (1983) examined a limited number of satellite-tracked drifter trajectories that coincided with the CTD survey in space and time. Velocity vectors based on daily averages of low-passed finite difference velocities are shown in Figure 4.2.4(a). Rather than compute a covariance function for this relatively small sample, the covariance function from the analysis of the 25/1500 db geopotential anomaly was used. Also, an assumed error level, $\overline{\varepsilon^2}$, was used rather than a computed estimate from the small sample. With the isotropic correlation scale estimated to be 75 km, the objective mapping produces the vector field in Figure 4.2.4(b). The stippled area in this figure corresponds to the region where the error variance exceeds 50% of the total variance. Due to the paucity of data, the area of statistically significant vector mapping is quite limited. Nevertheless, the resulting vectors are consistent with the geopotential height map in Figure 4.2.3(a).

Another example is provided by McWilliams (1976) who used dynamic height relative to 1500 m depth plus deep float velocities at 1500 m to estimate the streamfunction field. The isotropic covariance function for the random fluctuations in streamfunction $\Psi' = \Psi - \overline{\Psi}$ at 1500 m depth was

$$\begin{aligned} C(r) &= \overline{\Psi'(\mathbf{x}, z, t)\Psi'(\mathbf{x}+\mathbf{r}, z, t)} \\ &= \overline{\Psi'^2}(1-\varepsilon^2)(1-\gamma^2 r^2)\exp\left(-\tfrac{1}{2}\delta^2 r^2\right) \end{aligned} \tag{4.2.13}$$

where $\mathbf{r}$ is a horizontal separation vector, $r = |\mathbf{r}|$, ε is an estimate of relative measurement noise ($0 \leq \varepsilon \leq 1$), and γ^{-1}, δ^{-1} are decorrelation length scales found by fitting equation (4.2.13) to prior data. Denman and Freeland (1985) discuss the merits of five different covariance functions fitted to geopotential height data collected over a period of three years off the west coast of Vancouver Island. For other examples, the reader is referred to Bennett (1992).

As a final point, we remark that the requirement of isotropy is easily relaxed by using direction-dependent covariance matrices, $C(r_1, r_2)$ whose spatial structure depends on two orthogonal spatial coordinates, r_1 and r_2 (with $r_2 \geq r_1$). For example, the map of light attenuation coefficient at 20 m depth obtained from transmissometer profiles off the west coast of Vancouver Island (Figure 4.2.5) uses an exponentially decaying, elliptically shaped covariance matrix

$$C(r_1, r_2) = \exp\left[-a\Delta x^2 - b\Delta y^2 - c\Delta x\Delta y\right] \tag{4.2.14a}$$

where

$$\begin{aligned} a &= \tfrac{1}{2}\left\{[\cos(\pi\phi/180)/r_1]^2 + [\sin(\pi\phi/180)/r_2]^2\right\} \\ b &= \tfrac{1}{2}\left\{[\sin(\pi\phi/180)/r_1]^2 + [\cos(\pi\phi/180)/r_2]^2\right\} \\ c &= \cos(\pi\phi/180)\sin(\pi\phi/180)\left[r_2^2 - r_1^2\right]/(r_1 r_2)^2 \end{aligned} \tag{4.2.14b}$$

Here, Δx and Δy are, respectively, the eastward and northward distances from the grid point to the data point, and ϕ is the orientation angle (in degrees) of the coastline measured counterclockwise from north. In this case, it is assumed that the alongshore correlation scale, r_2, is twice the across-shore correlation scale, r_1. The idea here is that, like water-depth changes, alongshore variations in coastal water properties such as temperature, salinity, geopotential height, and log-transformed phytoplankton chlorophyll-*a* pigment concentration occur over longer length scales than across-shore variations.

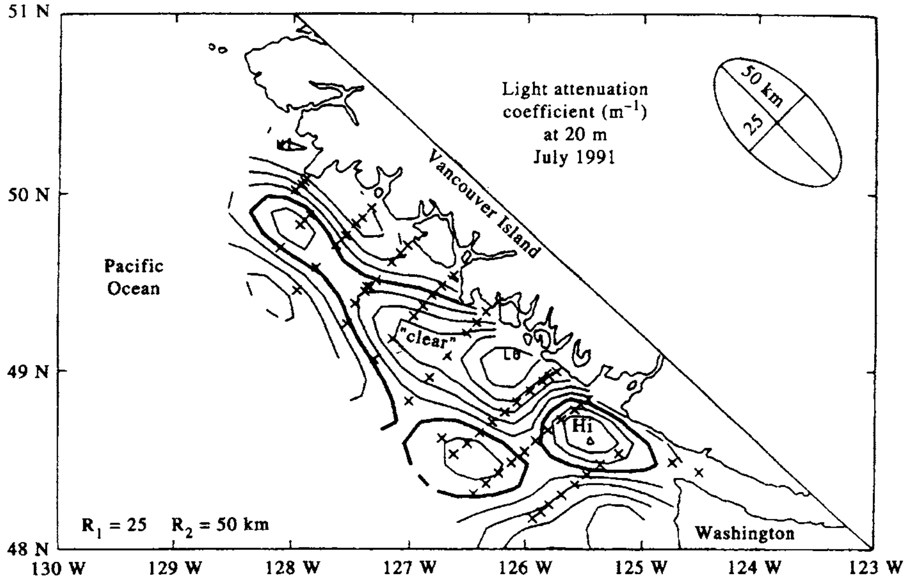

Figure 4.2.5. Objective analysis map of light attenuation coefficient (per meter) at 20 m depth on the west coast of Vancouver Island obtained from transmissometer profiles. The covariance function $C(r_1, r_2)$ given by the ellipse is assumed to decay exponentially with distance with the longshore correlation scale r_2 = 50 km and cross-shore correlation scale r_1 = 25 km.

4.3 EMPIRICAL ORTHOGONAL FUNCTIONS

The previous section dealt with the optimal smoothing of irregularly spaced data onto a gridded map. In other studies of oceanic variability, we may be presented with a large data set from a grid of time-series stations which we wish to compress into a smaller number of independent pieces of information. For example, in studies of climate change, it is necessary to deal with time series of spatial maps, such as surface temperature. A useful obvious choice would involve a linear combination of orthogonal spatial "predictors", or modes, whose net response as a function of time would account for the combined variance in all of the observations. The signals we wish to examine may all consist of the same variable, such as temperature, or they may be a mixture of variables such as temperature and wind velocity, or current and sea level. The data may be in the form of concurrent time-series records from a grid (regular or irregular) of stations $x_i(t), y_i(t)$ on a horizontal plane or time-series records at a selection of depths on an $x_i(t)$, $z_i(t)$ cross-section. Examples of time series from cross-sectional data include those from a single current meter string or from along-channel moorings of thermistor chains.

A useful technique for compressing the variability in this type of time-series data is *principal component analysis* (PCA). In oceanography, the method is commonly known as *empirical orthogonal function* (EOF) analysis. The EOF procedure is one of a larger class of inverse techniques and is equivalent to a data reduction method widely used in the social sciences known as *factor analysis*. The first reference we could find to the application of EOF analysis to geophysical fluid dynamics is a report by Edward Lorenz (1956) in which he develops the technique for statistical weather prediction and coins the term "EOF".

The advantage of EOF analysis is that it provides a compact description of the spatial and temporal variability of data series in terms of orthogonal functions, or statistical "modes." Usually, most of the variance of a spatially distributed series is in the first few orthogonal functions whose patterns may then be linked to possible dynamical mechanisms. It should be emphasized that no direct physical or mathematical relationship necessarily exists between the statistical EOFs and any related dynamical modes. Dynamical modes conform to physical constraints through the governing equations and associated boundary conditions (LeBlond and Mysak, 1979); empirical orthogonal functions are simply a method for partitioning the variance of a spatially distributed group of concurrent time series. They are called "empirical" to reflect the fact that they are defined by the covariance structure of the specific data set being analyzed (as shown below).

In oceanography and meteorology, EOF analysis has found wide application in both the time and frequency domains. Conventional EOF analysis can be used to detect standing oscillations only. To study propagating wave phenomena, we need to use lagged covariance matrix (Weare and Nasstrom, 1982), or complex principal component analysis in the frequency domain (Wallace and Dickinson, 1972; Horel, 1984). Our discussion, in this section, will focus on space/time domain applications. Readers seeking more detailed descriptions of both the procedural aspects and their applications are referred to Lorenz (1956), Davis (1976), and Preisendorfer (1988).

The best analogy to describe the advantages of EOF analysis is the classical vibrating drum problem. Using mathematical concepts presented in most undergraduate texts, we know that we can describe the eigenmodes of drumhead oscillations through a series of two-dimensional orthogonal patterns. These modes are defined by the eigenvectors and eigenfunctions of the drumhead. Generally, the lowest modes have the largest spatial scales and represent the most dominant (most prevalent) modes of variability. Typically, the drumhead has as its largest mode an oscillation in which the whole drumhead moves up and down, with the greatest amplitude in the center and zero motion at the rim where the drum is clamped. The next highest mode has the drumhead separated in the center with one side 180° out of phase with other side (one side is up when the other is down). Higher modes have more complex patterns with additional maxima and minima. Now, suppose we had no mathematical theory, and were required to describe the drumhead oscillations in terms of a set of observations. We would look for the kinds of eigenvalues in our data that we obtain from our mathematical analysis. Instead of the analytical or dynamical solutions that can be derived for the drum, we wish to examine "empirical" solutions based strictly on a measured data set. Since we are ignorant of the actual dynamical analysis, we call the resulting modes of oscillation, empirical orthogonal functions.

EOFs can be used in both the time and frequency domains. For now, we will restrict ourselves to the time domain application and consider a series of N maps at times $t = t_i$ $(1 \leq i \leq N)$, each map consisting of scalar variables $\psi_m(t)$ collected at M locations, $\mathbf{x}_m(1 \leq m \leq M)$. One could think of N weather maps available every 6 h over a total period of $6N$ h, with each map showing the sea surface pressure $\psi_m(t) = P_m(t)(1 \leq m \leq M)$ recorded at M weather buoys located at mooring sites $\mathbf{x}_m = (x_m, y_m)$. Clearly, the subscript m refers to the spatial grid locations in each map.

Alternatively, the N maps might consist of pressure data $P(t)$ from $M - K$ weather buoys plus velocity component records $u(t)$, $v(t)$ from $K/2$ current meter sites. Or, again the time series could be from $M/2$ current meters on a moored string. Any combination of scalars is permitted (remember, this is a statistical analysis not a dynamical analysis). The goal of this procedure is to write the data series $\psi_m(t)$ at any given location $\mathbf{x}_m$ as the sum of M orthogonal spatial functions $\phi_i(\mathbf{x}_m) = \phi_{im}$ such that

$$\psi(\mathbf{x}_m, t) = \psi_m(t) = \sum_{i=1}^{M} [a_i(t)\phi_{im}] \tag{4.3.1}$$

where $a_i(t)$ is the amplitude of the ith orthogonal mode at time $t = t_n$ $(1 \leq n \leq N)$. Simply put, equation (4.3.1) says that the time variation of the dependent scalar variable $\psi(\mathbf{x}_m, t)$ at each location $\mathbf{x}_m$ results from the linear combination of M spatial functions, ϕ_i, whose amplitudes are weighted by M time-dependent coefficients, $a_i(t)$, $(1 \leq i \leq M)$. The weights $a_i(t)$ tell us how the spatial modes ϕ_{im} vary with time. There are as many (M) basis functions as there are stations for which we have data. Put another way, we need as many modes as we have time-series stations so that we can account for the combined variance in the original time series at each time, t. We can also formulate the problem as M temporal functions whose amplitudes are weighted by M spatially variable coefficients. Whether we partition the data as spatial or temporal orthogonal functions the results are identical.

Since we want the $\phi_i(\mathbf{x}_m)$ to be orthogonal, so that they form a set of basis functions, we require

$$\sum_{m=1}^{M} \left[\phi_{im}\phi_{jm}\right] = \delta_{ij} \text{ (orthogonality condition)} \tag{4.3.2}$$

where the summation is over all observation locations and δ_{ij} is the Kronecker delta

$$\delta_{ij} = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{4.3.3}$$

It is worth remarking that two functions are said to be orthogonal when the sum (or integral) of their product over a certain defined space (or time) is zero. Orthogonality in equation (4.3.2) does not mean $\phi_{im}\phi_{jm} = 0$ for each m. For example, in the case of continuous sines and cosines, $\int \sin\theta \cos\theta \, d\theta = 0$ when the integral is over a complete phase cycle, $0 \leq \theta \leq 2\pi$. By itself, the product $\sin\theta \cdot \cos\theta = 0$ only if the sine or cosine term happens to be zero.

There is a multitude of basis functions, ϕ_i, that can satisfy equations (4.3.1) and (4.3.2). Sine, cosine, and Bessel functions come to mind. The EOFs are determined uniquely among the many possible choices by the constraint that the time amplitudes $a_i(t)$ are uncorrelated over the sample data. This requirement means that the time-averaged covariance of the amplitudes satisfies

$$\overline{a_i(t)a_j(t)} = \lambda_i\delta_{ij} \text{ (uncorrelated time variability)} \tag{4.3.4}$$

in which the overbar denotes the time-averaged value and

$$\lambda_i = \overline{a_i(t)^2} = \frac{1}{N}\sum_{n=1}^{N} \left[a_i(t_n)^2\right] \tag{4.3.5}$$

is the variance in each orthogonal mode. If we then form the covariance matrix

$\psi_m(t)\psi_k(t)$ for the known data and use (4.3.4), we find

$$\begin{aligned}\overline{\psi_m(t)\psi_k(t)} &= \overline{\sum_{i=1}^{M}\sum_{j=1}^{M}\left[a_i(t)a_j(t)\phi_{im}\phi_{jk}\right]} \\ &= \sum_{i=1}^{M}\left[\lambda_i\phi_{im}\phi_{ik}\right]\end{aligned} \tag{4.3.6}$$

Multiplying both sides of (4.3.6) by ϕ_{ik}, summing over all k and using the orthogonality condition (4.3.2), yields

$$\sum_{k=1}^{M}\overline{\psi_m(t)\psi_k(t)}\phi_{ik} = \lambda_i\phi_{im} \quad (i\text{th mode at the } m\text{th location}; i, m = 1, \ldots, M) \tag{4.3.7}$$

Equation (4.3.7) is the canonical form for the *eigenvalue problem*. Here, the EOFs, ϕ_{im}, are the ith *eigenvectors* at locations $\mathbf{x}_m$, and the mean-square time amplitudes

$$\lambda_i = \overline{a_i(t)^2}$$

are the corresponding *eigenvalues* of the mean product, **R**, which has elements

$$R_{mk} = \overline{\psi_m(t)\psi_k(t)}$$

This is equal to the covariance matrix, **C**, if the mean values of the time series $\psi_m(t)$ have been removed at each site $\mathbf{x}_m$. The total of M empirical orthogonal functions corresponding to the M eigenvalues of (4.3.7) forms a complete basis set of linearly independent (orthogonal) functions such that the EOFs are uncorrelated modes of variability. Assuming that the record means $\overline{\psi_m(t)}$ have been removed from each of the M time series, equation (4.3.7) can be written more concisely in matrix notation as

$$\mathbf{C}\boldsymbol{\phi} - \lambda\mathbf{I}\boldsymbol{\phi} = 0 \tag{4.3.8}$$

where the covariance matrix, **C**, consists of M data series of length N with elements

$$C_{mk} = \overline{\psi_m(t)\psi_k(t)}$$

I is the unity matrix, and $\boldsymbol{\phi}$ are the EOFs. Expanding (4.3.8) yields the eigenvalue problem

$$\begin{pmatrix} \overline{\psi_1(t)\psi_1(t)} & \overline{\psi_1(t)\psi_2(t)} & \ldots & \overline{\psi_1(t)\psi_M(t)} \\ \psi_2(t)\psi_1(t) & \psi_2(t)\psi_2(t) & \ldots & \psi_2(t)\psi_M(t) \\ \ldots & \ldots & \ldots & \ldots \\ \overline{\psi_M(t)\psi_1(t)} & \overline{\psi_M(t)\psi_2(t)} & \ldots & \overline{\psi_M(t)\psi_M(t)} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \ldots \\ \phi_M \end{pmatrix} = \begin{pmatrix} \lambda\ 0 \ldots 0 \\ 0\ \lambda \ldots 0 \\ \ldots \\ 0 \ldots \lambda \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \ldots \\ \phi_M \end{pmatrix} \tag{4.3.9a}$$

corresponding to the series of linear system of equations

$$\begin{aligned} \left[\overline{\psi_1(t)\psi_1(t)} - \lambda\right]\phi_1 + \overline{\psi_1(t)\psi_2(t)}\,\phi_2 + \ldots + \overline{\psi_1(t)\psi_M(t)}\,\phi_M &= 0 \\ \overline{\psi_2(t)\psi_1(t)}\,\phi_1 + \left[\overline{\psi_2(t)\psi_2(t)} - \lambda\right]\phi_2 + \ldots + \overline{\psi_2(t)\psi_M(t)}\,\phi_M &= 0 \\ &\ldots \\ \overline{\psi_M(t)\psi_1(t)}\,\phi_1 + \overline{\psi_M(t)\psi_2(t)}\,\phi_2 + \ldots + \left[\overline{\psi_M(t)\psi_M(t)} - \lambda\right]\phi_M &= 0 \end{aligned} \tag{4.3.9b}$$

The eigenvalue problem involves diagonalization of a matrix, which in turn amounts to finding an axis orientation in M-space for which there are no off-diagonal terms in the matrix. When this occurs, the different modes of the system are orthogonal. Since each $\mathbf{C}$ is a real symmetric matrix, the eigenvalues λ_i are real. Similarly, the eigenvectors (EOFs) of a real symmetric matrix are real. Because $\overline{C(x_m, x_k)}$ is positive, the real eigenvalues are all positive.

If equation (4.3.8) is to have a nontrivial solution, the determinant of the coefficients must vanish; that is

$$\det\begin{vmatrix} C_{11}-\lambda & C_{12} & \dots & C_{1M} \\ C_{21} & C_{22}-\lambda & \dots & \dots \\ \dots & \dots & \dots & \dots \\ C_{m1} & \dots & \dots & C_{MM}-\lambda \end{vmatrix} = 0 \qquad (4.3.10)$$

which yields an Mth order polynomial, $\lambda^M + \alpha\lambda^{M-1} + \dots$, whose M eigenvalues satisfy

$$\lambda_1 > \lambda_2 > \dots > \lambda_M \qquad (4.3.11)$$

Thus, the "energy" (variance) associated with each statistical mode is ordered according to its corresponding eigenvector. The first mode contains the highest percentage of the total variance, λ_1; of the remaining variance, the greatest percentage is in the second mode, λ_2, and so on. If we add up the total variance in all the time series, we get

$$\sum_{m=1}^{M}\left\{\frac{1}{N}\sum_{n=1}^{N}[\psi_m(t_n)]^2\right\} = \sum_{j=1}^{M}\lambda_j$$

Sum of variances in data = sum of variance in eigenvalues (4.3.12)

The total variance in the M time series equals the total variance contained in the M statistical modes. The final piece of the puzzle is to derive the time-dependent *amplitudes* of the ith statistical mode

$$a_i(t) = \sum_{m=1}^{M}\psi_m(t)\phi_{im} \qquad (4.3.13)$$

Equation (4.3.7) provides a computational procedure for finding the EOFs. By computing the mean product matrix, $\overline{\psi_m(t)\psi_k(t)}$ $(m, k = 1, \dots, M)$ or "scatter matrix" $\mathbf{S}$ in the terminology of Preisendorfer (1988), the eigenvalues and eigenvectors can be determined using standard computer algorithms. From these, we obtain the variance associated with each mode, λ_j, and its time-dependent variability, $a_i(t)$.

As outlined by Davis (1976), two advantages of a statistical EOF description of the data are: (1) the EOFs provide the most efficient method of compressing the data; and (2) the EOFs may be regarded as uncorrelated (i.e. orthogonal) modes of variability of the data field. The EOFs are the most efficient data representation in the sense that, for a fixed number of functions (trigonometric or other), no other approximate expansion of the data field in terms of $K < M$ functions

$$\hat{\psi}_m(t) = \sum_{m=1}^{K}a_i(t)\hat{\phi}_{im} \qquad (4.3.14)$$

can produce a lower total mean-square error

$$\sum_{m=1}^{K} \overline{\left[\psi_m(t) - \hat{\psi}_m(t)\right]^2} \tag{4.3.15}$$

than would be obtained when the $\hat{\phi}_i$ are the EOFs. A proof of this is given in Davis (1976). Also, as we will discuss later in this section, we could just as easily have written our data $\psi(\mathbf{x}_m, t)$ as a combination of orthogonal temporal modes $\phi_i(t)$ whose amplitudes vary spatially as $a_i(\mathbf{x}_m)$. Since this is a statistical technique, it doesn't matter whether we use time or space to form the basis functions. However, it might be easier to think in terms of spatial orthogonal modes that oscillate with time.

As noted above, EOFs are ordered by decreasing eigenvalue so that, among the EOFs, the first mode, having the largest eigenvalue, accounts for most of the variance of the data. Thus, with the inherent efficiency of this statistical description a very few empirical modes generally can be used to describe the fundamental variability in a very large data set. Often it may prove useful to employ the EOFs as a filter to eliminate unwanted scales of variability. A limited number of the first few EOFs (those with the largest eigenvalues) can be used to reconstruct the data field, thereby eliminating those scales of variability not coherent over the data grid and therefore less energetic in their contribution to the data variance. An EOF analysis can then be made of the filtered data set to provide a new apportionment of the variance for those scales associated with most of the variability in the original data set. In this application, EOF analysis is much like standard Fourier analysis used to filter out scales of unwanted variability. In fact, for homogeneous time series sampled at evenly spaced increments, it can be shown that the EOFs are Fourier trigonometric functions.

The computation of the eigenfunctions $a_i(t)$ in equation (4.3.13) requires the data values $\psi_m(t)$ for all of the time series. Often these time series contain gaps which make it impossible to compute $a_i(t)$ at those times for which the data are missing. One solution to this problem is to fill the gaps in the original data records using one of the procedures discussed in the previous chapter on interpolation. Most consistent with the present approach is to use objective analysis as discussed in the preceding section. While this will provide an interpolation consistent with the covariance of the subject data set, these optimally estimated values of $\psi_m(t)$ often result in large expected errors if the gaps are large or the scales of coherent variability are small.

An alternative method, suggested by Davis (1976), that can lead to a smaller expected error is to estimate the EOF amplitude at time, t, directly from the existing values of $\psi_m(t)$ thus eliminating the need for the interpolation of the original data. Conditions for this procedure are that the available number of sample data pairs is reasonably large (gaps do not dominate) and that the data time series are stationary. Under these conditions, the mean product matrix $\overline{\psi_m(t)\psi_k(t)}$ $(m, k = 1, \ldots, M)$ will be approximately the same as it would have been for a data set without gaps. For times when none of the $\psi_m(t)$ values are missing, the coefficients $a_i(t)$ can be computed from equation (4.3.13). For times t when data values are missing, $a_i(t)$ can be estimated from the available values of $\psi_m(t)$

$$\hat{a}_i(t) = b_i(t) \sum_{j=1}^{M'} \psi_j(t)\phi_{ij} \tag{4.3.16}$$

where the summation over j includes only the available data points, $M' \leq M$. From equations (4.3.8), (4.3.14), and (4.3.16), the expected square error of this estimate is

$$\overline{[a_i(t) - \hat{a}_{(}t)]^2} = b_i^2(t)\sum_{j=1}^{M'}\left(\lambda_j\gamma_{ji}^2\right) + \lambda_i[1 + b_i(t)(\gamma_{ii}-1)]^2 \qquad (4.3.17)$$

where

$$\gamma_{ji} = \sum_k \phi_j(k)\phi_i(k) \qquad (4.3.18)$$

and the summation over k applies only to those variables with missing data. Taking the derivative of the right-hand side of (4.3.17) with respect to b_i, we find that the expected square error is minimized when

$$b_i(t) = (1 - \gamma_{ii})\lambda_j / \left[(1 - \gamma_{ii})^2\lambda_j + \sum_j \lambda_j\gamma_{ji}^2\right] \qquad (4.3.19)$$

Applications of this procedure (Davis, 1976, 1978; Chelton and Davis, 1982, Chelton *et al.*, 1982), have shown that the expected errors are surprisingly small even when the number of missing data is relatively large. This is because the dominant EOFs in geophysical systems generally exhibit large spatial scales of variability, leading to a high coherence between grid values. As a consequence, contributions to the spatial pattern from the most dominant EOFs at any particular time, t, can be reliably estimated using a relatively small number of sample grid points.

4.3.1 Principal axes of a single vector time series (scatter plot)

A common technique for improving the EOF analysis for a set of vector time series is to first rotate each data series along its own customized principal axes. In this new coordinate system, most of the variance is associated with a major axis and the remaining variance with a minor axis. The technique also provides a useful application of principal component analysis. The problem consists of finding the principal axes of variance along which the variance in the observed velocity fluctuations $\mathbf{u}'(t) = [u_1'(t), u_2'(t)]$ is maximized for a given location; here u_1' and u_2' are the respective east–west and north–south components of the wind or current velocity obtained by removing the respective means $\overline{u_1}$ and $\overline{u_2}$ from each record; i.e. $u_1' = u_1 - \overline{u_1}$, $u_2' = u_2 - \overline{u_2}$. The amount of data "scatter" is a maximum along the major axis and a minimum along the minor axis (Figure 4.3.1). We also note that principal axes are defined in such a way that the velocity components along the principal axes are uncorrelated.

The eigenvalue problem (4.3.8) for a two-dimensional scatter plot has the form

$$\begin{vmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{vmatrix} \begin{vmatrix} \phi_1 \\ \phi_2 \end{vmatrix} = \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} \begin{vmatrix} \phi_1 \\ \phi_2 \end{vmatrix} \qquad (4.3.20)$$

where the C_{ij} are components of the covariance matrix, $\mathbf{C}$, and (ϕ_1, ϕ_2) are the eigenvectors associated with the two possible values of the eigenvalues, λ. To find the principal axes for the scatter plot of u_2' versus u_1', we set the determinant of the covariance matrix equation (4.3.20) to zero

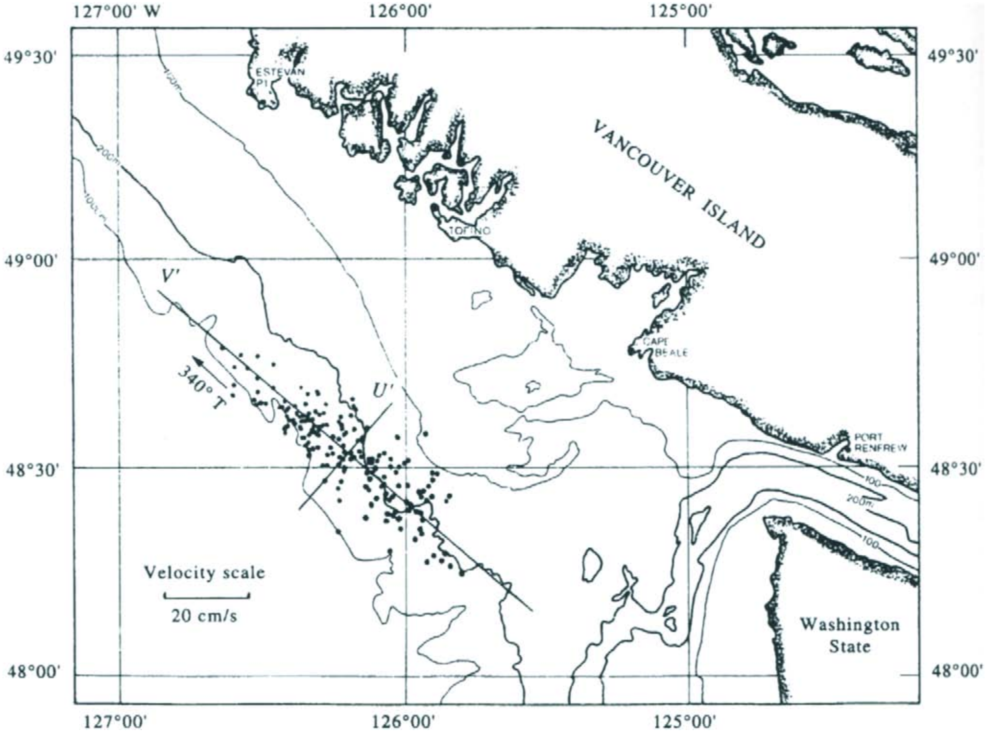

Figure 4.3.1. The principal component axes for daily averaged velocity components u, v measured by a current meter moored at 175 m depth on the west coast of Canada. Here, the north–south component oj velocity, v(t), is plotted as a scatter diagram against the east–west component of current velocity, u(t). Data cover the period 21 October 1992 to 25 May 1993. The major axis along 340°T can be used to define the longshore direction, **v**'.

$$\det|\mathbf{C}-\lambda\mathbf{I}| = \det\begin{vmatrix} C_{11}-\lambda & C_{12} \\ C_{21} & C_{22}-\lambda \end{vmatrix} = \det\begin{vmatrix} \overline{u_1'^2}-\lambda & \overline{u_1'u_2'} \\ \overline{u_2'u_1'} & \overline{u_2'^2}-\lambda \end{vmatrix} = 0 \tag{4.3.21a}$$

where (for $i = 1, 2$) the elements of the determinant are given by

$$C_{ii} = \overline{u_i'^2} = \frac{1}{N}\sum_{n=1}^{N}\left[u_i'(t_n)\right]^2 \tag{4.3.21b}$$

$$C_{ij} = \overline{u_i'u_j'} = \frac{1}{N}\sum_{n=1}^{N}\left[u_i'u_j'(t_n)\right] \tag{4.3.21c}$$

Solution of (4.3.21) yields the quadratic equation

$$\lambda^2 - \left[\overline{u_1'^2}+\overline{u_2'^2}\right]\lambda + \overline{u_1'^2}\,\overline{u_2'^2} - \overline{u_1'u_2'}^2 = 0 \tag{4.3.22}$$

whose two roots $\lambda_1 > \lambda_2$ are the eigenvalues, corresponding to the variances of the velocity fluctuations along the major and minor principal axes. The orientations of the two axes differ by 90° and the principal angles θ_p (those along which the sum of the squares of the normal distances to the data points u_1', u_2' are extremum) are found from the transcendental relation

$$\tan(2\theta_p) = \frac{2\overline{u_1'u_2'}}{\overline{u_1'^2} - \overline{u_2'^2}} \tag{4.3.23a}$$

or

$$\theta_p = \frac{1}{2}\tan^{-1}\left[\frac{2\overline{u_1'u_2'}}{\overline{u_1'^2} - \overline{u_2'^2}}\right] \tag{4.3.23b}$$

where the principal angle is defined for the range $-\pi/2 \le \theta_p \le \pi/2$ (Freeland *et al.*, 1975; Kundu and Allen, 1976; Preisendorfer, 1988). As usual, the multiple $n\pi/2$ ambiguities in the angle that one obtains from the arctangent function must be addressed by considering the quantrants of the numerator and denominator in equation (4.3.23). Preisendorfer (1988; Figure 2.3) outlines the nine different possible cases. Proof of (4.3.23) is given in Section 4.3.5.

The principal variances (λ_1, λ_2) of the data set are found from the determinant relations (4.3.21a) and (4.3.22) as

$$\left.\begin{matrix}\lambda_1\\ \lambda_2\end{matrix}\right\} = \frac{1}{2}\left\{\left(\overline{u_1'^2} + \overline{u_2'^2}\right) \pm \left[\left(\overline{u_1'^2} - \overline{u_2'^2}\right)^2 + 4\left(\overline{u_1'u_2'}\right)^2\right]^{1/2}\right\} \tag{4.3.24}$$

in which the + sign is used for λ_1 and the − sign for λ_2. In the case of current velocity records, λ_1 gives the variance of the flow along the major axis and λ_2 the variance along the minor axis. The slope, $s_1 = \phi_2/\phi_1$, of the eigenvector associated with the variance λ_1 is found from the matrix relation

$$\begin{vmatrix} \overline{u_1'^2} - \lambda & \overline{u_1'u_2'} \\ \overline{u_2'u_1'} & \overline{u_2'^2} - \lambda \end{vmatrix} \begin{vmatrix} \phi_1 \\ \phi_2 \end{vmatrix} = 0 \tag{4.3.25a}$$

Solving (4.3.25a) for $\lambda = \lambda_1$, gives

$$\begin{aligned} \left[\overline{u_1'^2} - \lambda_1\right]\phi_1 + \overline{u_1'u_2'}\,\phi_2 &= 0 \\ \left[\overline{u_2'u_1'}\,\phi_1\right] + \left[\overline{u_2'^2} - \lambda_1\right]\phi_2 &= 0 \end{aligned} \tag{4.3.25b}$$

so that

$$s_1 = \left[\lambda_1 - \overline{u_1'^2}\right] / \overline{u_1'u_2'} \tag{4.3.25c}$$

with a similar expression for the slope s_2 associated with the variance $\lambda = \lambda_2$. If $\lambda_1 \gg \lambda_2$, then $\lambda_1 \approx \overline{u_1'^2} + \overline{u_2'^2}$, and $s_1 \approx \overline{u_2'^2}/\overline{u_1'u_2'}$. The usefulness of principal component analysis is that it can be used to find the main orientation of fluid flow at any current meter or anemometer site, or within a "box" containing velocity variances derived from Lagrangian drifter trajectories (Figure 4.3.2). Since the mean and low frequency currents in relatively shallow waters are generally "steered" parallel to the

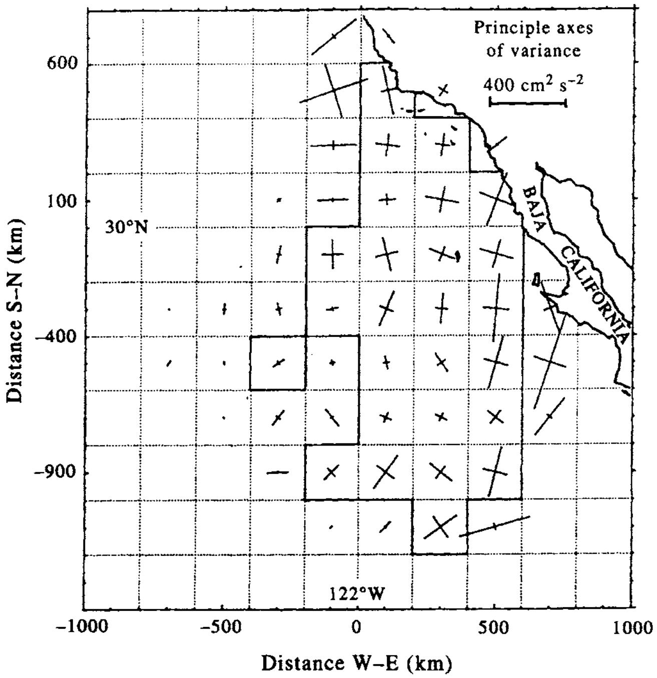

Figure 4.3.2. Principal axes of current velocity variance (kinetic energy) obtained from surface satellite-tracked drifter measurements off the coast of southern California during 1985–86. For this analysis, data have been binned into 200 × 200 km^2 boxes Solid border denotes the region for which there were more than 50 drifter-days and more than two different drifter tracks. (From Poulain and Niiler, 1989).

coastline or local bottom contours, the major principal axis is often used to define the "longshore" direction while the minor axis defines the "cross-shore" direction of the flow. It is this type of information that is vital to estimates of cross-shore flux estimates. In the case of prevailing coastal winds, the major axis usually parallels the mean orientation of the coastline or coastal mountain range that steers the surface winds.

4.3.2 EOF computation using the scatter matrix method

There are two primary methods for computing the EOFs for a grid of time series of observations. These are: (1) The scatter matrix method which uses a "brute force" computational technique to obtain a symmetric covariance matrix **C** which is then decomposed into eigenvalues and eigenvectors using standard computer algorithms (Preisendorfer, 1988); and (2) the computationally efficient singular value decomposition (SVD) method which derives all the components of the EOF analysis (eigenvectors, eigenvalues, *and* time-varying amplitudes) without computation of the covariance matrix (Kelly, 1988). The EOFs determined by the two methods are identical. The differences are mainly the greater degree of sophistication, computational speed, and computational stability of the SVD approach.

Details of the covariance matrix approach can be found in Preisendorfer (1988). This recipe, which is only one of several possible procedures that can be applied, involves the preparation of the data and the solution of equation (4.3.8) as follows:

(1) Ensure that the start and end times for all M time series of length N are identical. Typically, $N > M$.

(2) Remove the record mean and linear trend from each time-series record $\psi_m(t)$, $1 \leq m \leq M$, such that the fluctuations of $\psi_m(t)$ are given by $\psi'_m(t) = \psi_m(t) - [\overline{\psi_m(t)} + b_m(t - \bar{t})]$ where b_m is the slope of the least-squares regression line for each location. Other types of trend can also be removed.

(3) Normalize each de-meaned, de-trended time series by dividing each data series by its standard deviation $s = [1/(N-1)\sum(\psi_{m'})^2]^{1/2}$ where the summation is over all time, t $(t_n : 1 \leq n \leq N)$. This ensures that the variance from no one station dominates the analysis (all stations get an equal chance to contribute). The M normalized time-series fluctuations, ψ'_m, are the data series we use for the EOF analysis. The total variance for each of the M eigenvalues = 1; thus, the total variance for all modes, $\sum \lambda_i = M$.

(4) Rotate any vector time series to its principal axes. Although this operation is not imperative, it helps maximize the signal-to-noise ratio for the preferred direction. For future reference, keep track of the means, trends and standard deviations derived from the M time series records.

(5) Construct the $M \times N$ data matrix, $\mathbf{D}$, using the M rows (locations $\mathbf{x}_m$) and N columns (times t_n) of the normalized data series

$$D = \begin{array}{ll} \text{Time} \rightarrow & \\ \begin{pmatrix} \psi'_1(t_1) & \psi'_1(t_2) & \ldots & \psi'_1(t_N) \\ \psi'_2(t_1) & \psi'_2(t_2) & \ldots & \psi'_2(t_N) \\ \ldots & \ldots & \ldots & \ldots \\ \psi'_M(t_1) & \psi'_M(t_2) & \ldots & \psi'_M(t_N) \end{pmatrix} & \text{Location} \downarrow \end{array} \tag{4.3.26}$$

and from this derive the symmetric covariance matrix, $\mathbf{C}$, by multiplying $\mathbf{D}$ by its transpose $\mathbf{D}^T$

$$\mathbf{C} = \frac{1}{N\phi\phi - 1}\mathbf{D}\mathbf{D}^T \tag{4.3.27}$$

where $\mathbf{S} = (N-1)\mathbf{C}$ is the scatter matrix defined by Preisendorfer (1988), and

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1M} \\ C_{21} & C_{22} & \ldots & C_{2M} \\ \ldots & \ldots & \ldots & \ldots \\ C_{M1} & \ldots & \ldots & C_{MM} \end{pmatrix} \tag{4.3.28}$$

The elements of the real symmetric matrix $\mathbf{C}$ are

$$C_{ij} = C_{ji} = \frac{1}{N-1}\sum_{n=1}^{N}\left[\psi'_i(t_n)\psi'_j(t_n)\right] \tag{4.3.29}$$

The eigenvalue problem then becomes

$$\mathbf{C}\boldsymbol{\phi} = \lambda\boldsymbol{\phi} \tag{4.3.30}$$

where λ are the eigenvalues and $\boldsymbol{\phi}$ the eigenvectors.

At this point, we remark that we have formulated the EOF decomposition in terms of an $M \times M$ "spatial" covariance matrix whose time-averaged elements are given by the product $(N - 1)^{-1}\,\mathbf{DD}^T$ (4.3.27). We could just as easily have formed an $N \times N$ "temporal" covariance matrix whose spatially averaged elements are given by the product $(M - 1)^{-1}\,\mathbf{D}^T\mathbf{D}$. The mean values we remove in preparing the two data sets are slightly different since preparation of $\mathbf{D}$ involves time averages while preparation of $\mathbf{D}^T$ involves spatial averages. However, in principle, the two problems are identical, and the percentage of the total time-series variance in each mode depends on whether one computes the spatial EOFs or temporal EOFs. As we further point out in the following section, another difference between the two problems is how the singular values are grouped and which is identified with the spatial function and which with the temporal function (Kelly, 1988). The designation of one set of orthogonal vectors as EOFs and the other as amplitudes is quite arbitrary.

Once the matrix $\mathbf{C}$ has been calculated from the data, the problem can be solved using "canned" programs from one of the standard statistical or mathematical computer libraries for the eigenvalues and eigenvectors of a real symmetric matrix. In deriving the values listed in Tables 4.3.1–4.3.6, we have used the double-precision program DEVLSF of the International Math and Science Library (IMSL). The program outputs the eigenvalues λ in increasing order. To obtain λ in decreasing order of importance, we have had to invert the eigenvalue output. For each eigenvector or mode, the program normalizes all values to the maximum value for that mode. The amplitude of the maximum value is unity (= 1). Since there are M eigenvalues, the data normalization process gives a total EOF variance of $M(\sum \lambda_i = M)$. The canned programs also allow for calculation of a "performance index" (PI) which measures the error of the eigenvalue problem (4.3.30) relative to the various components of the problem and the machine precision. The performance of the eigenvalue routine is considered "excellent" if PI < 1, "good" if $1 \leq \mathrm{PI} \leq 100$, and "poor" if PI > 100. As a final analysis, we can conduct an *orthogonality check* on the EOFs by using the relation (4.3.2). Here we look for significant departures from zero in the products of different modes; if any of the products

$$\sum_{m=1}^{M} [\phi_{im}\phi_{jm}]$$

Table 4.3.1. Data matrix $\mathbf{D}^T$. *Components of velocity (cm/s) at three different sites at 1700 m depth in the northeast Pacific. Records start 29 September 1985 and are located near 48°N, 129°W. For each of the three stations we list the east–west (u) and north–south component (v). The means and trends have not yet been removed*

Time (days)	Site 1 (u_1)	Site 1 (v_1)	Site 2 (u_2)	Site 2 (v_2)	Site 3 (u_3)	Site 3 (v_3)
1	−0.3	0.0	0.4	−0.4	−0.8	−1.4
2	−0.1	0.3	0.4	−0.3	−1.1	0.0
3	−0.1	−0.4	0.0	−0.5	0.0	−2.5
4	0.2	0.6	0.0	−0.6	−0.7	0.4
5	0.3	−0.1	−0.6	−0.3	0.0	−0.3
6	0.5	0.0	0.9	−0.6	0.6	0.3
7	0.2	0.2	−0.1	−0.7	1.2	−2.8
8	−0.5	−0.9	0.0	−0.6	0.0	−1.8

Table 4.3.2. Means, standard deviations and trends for each of the time-series components for each of the three current meter sites listed in Table 4.3.1. Means and trends have been removed from the time series prior to calculation of the standard deviations

Component	Mean (cm/s)	Standard deviation (cm/s)	Trend (cm/s/day)
u_1 (east–west)	0.025	0.328	0.024
v_1 (north–south)	– 0.037	0.418	– 0.075
u_2 (east–west)	0.125	0.433	– 0.038
v_2 (north–south)	– 0.500	0.114	– 0.040
u_3 (east–west)	– 0.100	0.503	0.233
v_3 (north–south)	– 1.012	1.250	– 0.108

Table 4.3.3. Principal axes for the current velocity at each site in Table 4.3.1. The angle θ is measured counterclockwise from east. Axes (half) lengths are in cm/s

Station ID	Angle θ (°)	Major axis	Minor axis
Site 1	54.7	0.461	0.185
Site 2	– 6.2	0.408	0.098
Site 3	– 77.7	1.193	0.406

Table 4.3.4. Eigenvalues and percentage of variance in each statistical mode derived from the data in Table 4.3.1

Eigenvalue No.	Eigenvalue	Percentage
1	2.2218	37.0
2	1.7495	29.2
3	1.1787	19.6
4	0.6953	11.6
5	0.1498	2.5
6	0.0048	0.1
Total	6.0000	100.0

Table 4.3.5. Eigenvectors (EOFs) ϕ_i for the data matrix in Table 4.3.1. Modes are normalized to the maximum value for each mode

Station ID	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
Site 1 u_1	1.000	– 0.032	– 0.430	0.479	– 0.599	– 0.969
Site 1 v_1	0.958	– 0.078	– 0.162	– 0.966	1.000	0.085
Site 2 u_2	0.405	0.230	1.000	0.910	0.517	– 0.295
Site 2 v_2	– 0.329	– 0.898	– 0.525	1.000	0.784	– 0.111
Site 3 u_3	0.349	1.000	– 0.474	0.812	0.124	0.907
Site 3 v_3	0.654	– 0.964	0.263	0.190	– 0.539	1.000

are significantly different from zero for $i \neq j$, then the EOFs are not orthogonal and there are errors in the computation. A computational example is given in Section 4.3.4.

Table 4.3.6. Time series of the amplitudes, $a_i(t)$, for each of the statistical modes

Time	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6
Day 1	0.798	−0.773	0.488	0.089	0.091	0.124
Day 2	−0.076	1.258	0.402	0.126	0.595	−0.089
Day 3	1.153	−1.582	−0.458	0.275	−0.492	−0.094
Day 4	−1.531	0.759	0.363	−1.585	−0.382	0.000
Day 5	0.097	1.647	−2.099	0.509	−0.128	0.039
Day 6	−2.169	−0.142	1.084	1.296	−0.171	0.008
Day 7	−0.721	−1.921	−0.866	−0.534	0.503	0.004
Day 8	2.450	0.754	1.085	−0.176	−0.017	0.008

4.3.3 EOF computation using singular value decomposition

The above method of computing EOFs requires use of covariance matrix, **C**. This becomes computationally impractical for large, regularly spaced data fields such as a sequence of infrared satellite images (Kelly, 1988). In this case, for a data matrix **D** over N time periods (N satellite images, for example), the covariance or mean product matrix is given by (4.3.27)

$$\mathbf{C} = \frac{1}{N-1}\mathbf{D}\mathbf{D}^T \tag{4.3.31}$$

where $\mathbf{D}^T$ is the transpose of the data matrix **D**. If we assume that all of the spatial data fields (i.e. satellite images) are independent samples, then the mean product matrix is the covariance matrix and the EOFs are again found by solving the eigenvalue problem

$$\mathbf{C}\boldsymbol{\phi} = \boldsymbol{\phi}\boldsymbol{\Lambda} \tag{4.3.32}$$

where $\boldsymbol{\phi}$ is the square matrix whose columns are eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. For satellite images, there may be $M = 5000$ spatial points sampled $N = 50$ times, making the covariance matrix a 5000 × 5000 matrix. Solving the eigenvalue problem for $\boldsymbol{\phi}$ would take $\max\{O(M^3), O(MN^2)\}$ operations. As pointed out by Kelly (1988), the operation count for the SVD method is $O(MN^2)$ which represents a considerable savings in computations over the traditional EOF approach if M is large. This is primarily true for those cases where M, the number of locations in the spatial data matrix, **D**, are far greater than the number of temporal samples (i.e. images).

There are two computational reasons for using the singular value decomposition method instead of the covariance matrix approach (Kelly, 1988): (1) The SVD formulation provides a one-step method for computing the various components of the eigenvalue problem; and (2) it is not necessary to compute or store a covariance matrix or other intermediate quantities. This greatly simplifies the computational requirements and provides for the use of canned analysis programs for the EOFs. Our analysis is based on the double-precision program DLSVRR in the IMSL. The SVD method is based on the concept in linear algebra (Press *et al.*, 1992) that any $M \times N$ matrix, **D**, whose number of rows M is greater than or equal to its number of columns, N, can be written as the product of three matrices: an $M \times N$ column-orthogonal matrix, **U**, an $N \times N$ diagonal matrix, **S**, with positive or zero elements, and the transpose ($\mathbf{V}^T$) of an $N \times N$ orthogonal matrix, $\mathbf{V}$. In matrix notation, the SVD becomes:

$$\mathbf{D} = \mathbf{U}\begin{pmatrix} s_1 & & & \\ & s_2 & & \\ & & \cdots & \\ & & & s_N \end{pmatrix}\mathbf{V}^T \tag{4.3.33}$$

For oceanographic applications, the data matrix, $\mathbf{D}$, consists of M rows (spatial points) and N columns (temporal samples). The scalars $s_1 \geq s_2 \geq \ldots \geq s_N \geq 0$ of the matrix $\mathbf{S}$, called the *singular values* of $\mathbf{D}$, appear in descending order of magnitude in the first N positions of the matrix. The columns of the matrix $\mathbf{V}$ are called the left singular vectors of $\mathbf{D}$ and the columns of the matrix $\mathbf{U}$ are called the right singular vectors of $\mathbf{D}$. The matrix $\mathbf{S}$ has a diagonal upper $N \times N$ part, $\mathbf{S}'$, and a lower part of all zeros in the case when $M > N$. We can express these aspects of $\mathbf{D}$ in matrix notation by rewriting equation (4.3.33) in the form

$$\mathbf{D} = [\mathbf{U}|\mathbf{0}]\left|\frac{\mathbf{S}'}{\mathbf{0}}\right|\mathbf{V}^T \tag{4.3.34}$$

where $[\mathbf{U}\,|\,\mathbf{0}]$ denotes a left singular matrix and $\mathbf{S}'$ denotes the nonzero part of $\mathbf{S}$ which has zeros in the lower part of the matrix (Kelly, 1988).

The matrix $\mathbf{U}$ is orthogonal, and the matrix $\mathbf{V}$ has only N significant columns which are mutually orthogonal such that,

$$\begin{aligned} \mathbf{V}^T\mathbf{V} &= \mathbf{I} \\ \mathbf{U}^T\mathbf{U} &= \mathbf{I} \end{aligned} \tag{4.3.35}$$

Returning to equation (4.3.33), we can compute the eigenvectors, eigenvalues and eigenfunctions of the principal component analysis in one single step. To do this, we prepare the data as before following steps 1–5 in Section 4.3.2. We then use commercially available programs such as the double-precision program DLSVRR in the IMSL. The elements of matrix $\mathbf{U}$ are the eigenvectors while those of matrix $\mathbf{S}$ are related to the eigenvalues $s_1 \geq s_2 \geq \ldots \geq s_N \geq 0$. To obtain the time-dependent amplitudes (eigenfunctions), we require a matrix $\mathbf{A}$ such that

$$\mathbf{D} = \mathbf{U}\mathbf{A}^T \tag{4.3.36}$$

which, by comparison with equation (4.3.33), requires

$$\mathbf{A} = \mathbf{V}\mathbf{S} \tag{4.3.37}$$

Hence, the amplitudes are simply the eigenvectors of the transposed problem multiplied by the singular values, $\mathbf{S}$. Solutions of (4.3.33) are identical (within round-off errors) to those obtained using the covariance matrix of the data, $\mathbf{C}$. We again remark that the only difference between the matrices $\mathbf{U}$ and $\mathbf{V}$ is how the singular values are grouped and which is identified with the spatial function and which with the temporal function. The designation of $\mathbf{U}$ as EOFs and $\mathbf{V}$ as amplitudes is quite arbitrary.

The decomposition of the data matrix $\mathbf{D}$ through singular value decomposition is possible since we can write it as a linear combination of functions $F_i(x)$, $i = 1, M$ so that

$$\mathbf{D} = \mathbf{F}\boldsymbol{\alpha} \tag{4.3.38a}$$

or

$$\begin{pmatrix} D(x_1, t_j) \\ D(x_2, t_j) \\ \dots \\ \dots \\ D(x_N, t_j) \end{pmatrix} = \begin{pmatrix} F_1(x_1) \dots F_N(x_1) \\ F_1(x_2) \dots F_N(x_2) \\ \dots \\ \dots \\ F_1(x_N) \dots F_N(x_N) \end{pmatrix} \begin{pmatrix} \alpha_1(t_j) \\ \alpha_2(t_j) \\ \dots \\ \dots \\ \alpha_N(t_j) \end{pmatrix} \tag{4.3.38b}$$

where the α_i are functions of time only. The functions F are chosen to satisfy the orthogonality relationship

$$\mathbf{F}\mathbf{F}^T = \mathbf{I} \tag{4.3.39}$$

so that the data matrix **D** is divided into orthogonal modes

$$\mathbf{D}\mathbf{D}^T = \mathbf{F}\mathbf{a}\mathbf{a}^T\mathbf{F}^T = \mathbf{F}\mathbf{L}\mathbf{F}^T \tag{4.3.40}$$

where $\mathbf{L} = \mathbf{a}\mathbf{a}^T$ is a diagonal matrix. The separation of the modes arises from the diagonality of the **L** matrix, which occurs because $\mathbf{D}\mathbf{D}^T$ is a real and symmetric matrix and **F** a unitary matrix. To reduce sampling noise in the data matrix **D**, one would like to describe it with fewer than M functions. If **D** is approximated by $\tilde{\mathbf{D}}$, which uses only K functions ($K < M$), then the K functions which best describe the **D** matrix in the sense that

$$(\tilde{\mathbf{D}} - \mathbf{D})^T(\tilde{\mathbf{D}} - \mathbf{D})$$

is a minimum are the empirical orthogonal functions which correspond to the largest valued elements of the traditional EOFs found earlier.

4.3.4 An example: deep currents near a mid-ocean ridge

As an example of the different concepts presented in this section, we again consider the eight days of daily averaged currents ($N = 8$) at three deep current meter sites in the northeast Pacific near the Juan de Fuca Ridge (Table 4.3.1). Since each site has two components of velocity, $M = 6$. The data all start on the same day and have the same number of records. Following the five steps outlined in Section 4.3.2, we first removed the average value from each time series. We then calculated the standard deviation for each time series and used this to normalize the time series so that each normalized series has a variance of unity. For convenience, we write the transpose of the data matrix, $\mathbf{D}^T$,where columns are the pairs of components of velocity (u, v) and rows are the time in days.

Time-series plots of the first three eigenmodes are presented in Figure 4.3.3. The performance index (PI) for the scatter matrix method was 0.026, which suggests that the matrix inversion in the eigenvalue solutions was well defined. A check on the orthogonality of the eigenvectors suggests that the singular value decomposition gave vectors which were slightly more orthogonal than the scatter matrix approach. For each combination (i, j) of the orthogonality condition (4.3.2), the products $\sum_{i,j}[\phi_{im}\phi_{jm}]$ were typically of order 10^{-7} for the SVD method and 10^{-6} for the

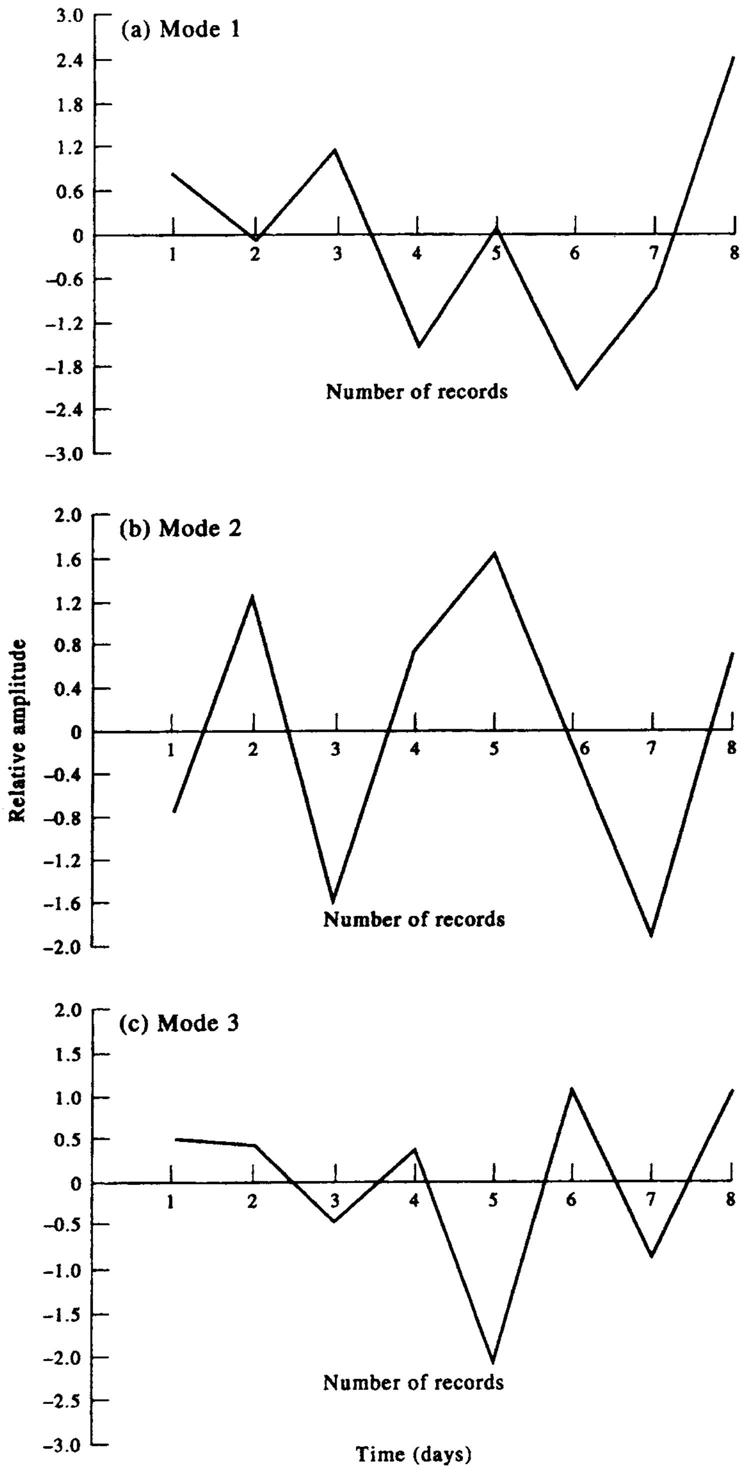

Figure 4.3.3. Eight-day time series for the first three EOFs for current meter data collected simultaneously at three sites at 1700 m depth in the northeast Pacific in the vicinity of Juan de Fuca Ridge, 1985. Modes 1, 2, 3 account for 37.0, 29.2, and 19.6 % of the variance, respectively.

scatter matrix method. Similar results apply to the orthogonality of the eigenmodes given by equation (4.3.4).

Before closing this section, we remark that we also could have performed the above analysis using complex EOFs of the form

$$\psi_m(t) = u_m(t) + \mathrm{i}v_m(t)$$

(where $\mathrm{i} = \sqrt{-1}$) in which case $M = 3$. This formulation not only allows the EOF vectors to change amplitude with time, as in our previous decomposition using $2M$ real EOFs, but also to rotate in time.

4.3.5 Interpretation of EOFs

In interpreting the meaning of EOFs, we need to keep in mind that, while EOFs offer the most efficient statistical compression of the data field, empirical modes do not necessarily correspond to true dynamical modes or modes of physical behavior. Often, a single physical process may be spread over more than one EOF. In other cases, more than one physical process may be contributing to the variance contained in a single EOF. The statistical construct derived from this procedure must be considered in light of accepted physical mechanisms rather than as physical modes themselves. It often is likely that the strong variability associated with the dominant modes is attributable to several identifiable physical mechanisms. Another possible clue to the physical mechanisms associated with the EOF patterns can be found in the time-series coefficients $a_i(t)$. Something may be known about the temporal variability of a process that might resemble the time series of the EOF coefficients, which would then suggest a causal relationship not readily apparent in the spatial structure of the EOF.

One way to interpret EOFs is to imagine that we have displayed the data as a scatter diagram in an effort to discover if there is any inherent correlation among the values. For example, consider two parameters such as sea surface temperature (SST) and sea-level pressure (SLP) measured at a number of points over the North Pacific. This is the problem studied by Davis (1976) where he analyzed sets of monthly SST and SLP over a period of 30 years for a grid in the North Pacific. If we plot x = SST against y = SLP in a scatter diagram, any correlation between the two would appear as an elliptical cluster of points. A more common example is that of Figure 4.3.1 where we plotted the north–south (y) component of daily mean current against the corresponding east–west (x) component for a continental shelf region. Here, the mean flow tends to parallel the coastline, so that the scatter plot again has an elliptical distribution. To take advantage of this correlation, we want to redefine our coordinate system by rotating x and y through the angle θ to the principal axes representation x', y' discussed in Section 4.3.2. This transformation is given by

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta \end{aligned} \tag{4.3.41}$$

What we have done in this rotation is to formulate a new set of axes that explains most of the variance, subject to the assumption that the variance does not change with time. Since the axes are orthogonal, the total variance will not change with rotation. Let $V = \overline{x'^2} = N^{-1}\sum x'^2$ be the particular variance we want to maximize (as usual, the summation is over all time). Note that we have focused on x' whereas the total variance is actually determined by r^2, where r is the distance of each point from the origin. However, we can expand $r^2 = x^2 + y^2$ and associate the variance with a given coordinate. In other words, if we maximize the variance associated with x', we will minimize the variance associated with y'. Using our summation convention, we can write

$$V = \overline{x'^2} = \overline{x^2}\cos^2\theta + 2\overline{xy}\sin\theta\cos\theta + \overline{y^2}\sin^2\theta \tag{4.3.42}$$

and

$$\frac{\partial V}{\partial\theta} = 2\left(\overline{y^2} - \overline{x^2}\right)\sin\theta\cos\theta + 2\overline{xy}\cos 2\theta \tag{4.3.43}$$

We maximize (4.3.43) by setting $\partial V/\partial\theta = 0$, giving (4.3.24), which we previously quoted without proof

$$\tan(2\theta_p) = \frac{2\overline{xy}}{\overline{x^2} - \overline{y^2}} \tag{4.3.44}$$

From (4.3.44), we see that if

$$\overline{xy} \ll \max\left(\overline{x^2}, \overline{y^2}\right)$$

then $\tan(2\theta_p) \to 0$ and $\theta_p = 0$, or $\pm 90°$, and we are left with the original axes. If $\overline{x^2} = \overline{y^2}$ and $\overline{xy} \neq 0$, then $\tan(2\theta_p) \to \pm\infty$ and the new axes are rotated $\pm 45°$ from the original axes.

We now find the expression for V. Since $\sec^2(2\theta) = 1 + \tan^2(2\theta)$

$$\begin{aligned} \cos 2\theta &= \left(\overline{x^2} - \overline{y^2}\right)/\pm D \\ \sin 2\theta &= \left[1 - \cos^2(2\theta)\right]^{1/2} = 2\overline{xy}/\pm D \end{aligned} \tag{4.3.45}$$

where

$$D = \left[\left(\overline{x^2} - \overline{y^2}\right)^2 + 4\overline{xy}^2\right]^{1/2} \tag{4.3.46}$$

Then, using the identities

$$\cos^2\theta = \tfrac{1}{2}(1 + \cos 2\theta), \quad \sin^2\theta = \tfrac{1}{2}(1 - \cos 2\theta) \tag{4.3.47}$$

we can write the variance as

$$\begin{aligned} V &= \overline{x^2}\frac{(1 + \cos 2\theta_p)}{2} + \overline{y^2}\frac{(1 - \cos 2\theta_p)}{2} + \overline{xy}\sin 2\theta_p \\ &= \frac{1}{2}\left\{\left(\overline{x^2} + \overline{y^2}\right) \pm \left[\left(\overline{x^2} - \overline{y^2}\right)^2 + 4\overline{xy}^2\right]^{1/2}\right\} \end{aligned} \tag{4.3.48}$$

The two roots of this equation correspond to a maximum and a minimum of V. For a new axis for which $\overline{x'^2}$ is a maximum, we will find $\overline{y'^2}$ a minimum. This follows automatically from the fact that the total variance is conserved. However, we can confirm this mathematically by computing $\partial^2 V/\partial\theta^2 = 0$. From equation (4.3.43) we find

$$\begin{aligned}\partial^2 V/\partial\theta^2 &= 2\left(\overline{y^2} - \overline{x^2}\right)\cos(2\theta_p) - 4\overline{xy}\sin(2\theta_p)\\ &= -2\left[\left(\overline{x^2} - \overline{y^2}\right) + 4\overline{xy}\right]/\pm D = \pm 2D\end{aligned} \tag{4.3.49}$$

The positive sign in equation (4.3.49) corresponds to a maximum (since (4.3.48) is negative); the negative sign corresponds to a minimum. It so happens that the variance solutions given by (4.3.49) are also the eigenvalues of the covariance matrix. Thus, we can return to our previous methods where we used the covariance matrix to compute the EOFs.

A published example of EOF analysis is presented by Davis (1976) who examined monthly maps of SST and SLP for the years 1947–74. The SLP data were originally obtained from the Long-Range Prediction Group of the U.S. National Meteorological Center (NMC) as one-month averages on a 5° diamond-shaped grid (i.e. 20°N–140°W, 20°N–150°W, ... , 25°N–145°W, 25°N–155°W, etc.). The data were transferred to a regular 5°-square grid using linear interpolation from the four nearest diamond grid points to fill in the square grid. The SST data were obtained from the U.S. National Marine Fisheries Service in the form of monthly averages over 2° squares. Because this grid spacing is not a submultiple of 5°, and because sometimes data were missing, the following data analysis scheme was employed. The 2° data were subjectively analyzed to produce maps contoured with a 1°F contour interval. During this stage, missing values were filled in where feasible. The corrected values were then linearly interpolated onto a 1° grid and 25 values were averaged to formulate area averages on the chosen 5° grid coincident with the SLP data. The ship data originated as ship injection temperatures and are subject to all of the problems discussed earlier in the section on SST.

Before carrying out the EOF analysis, the SST and SLP data sets were further averaged onto a grid with a 5° latitude spacing and a 10° longitude spacing (Figure 4.3.4). In those cases where some SST values were missing, the available observations were used to compute the grid average. Even then there were some 5° × 10° regions with missing data in the SST fields. Both fields were then converted to anomalies using the mean of the 28-year data set as the reference field. Thus, each of the individual monthly maps were transformed into anomaly maps, corresponding to the deviation of local values from the long-term mean.

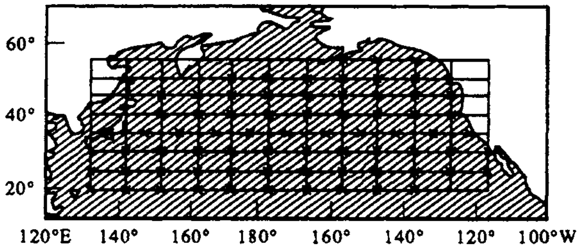

Figure 4.3.4. The grid of sea surface temperature (SST) and sea-level pressure (SLP). The 10° longitude by 5° latitude SLP averages are centered at grid intersections and SST averages are centered at crosses. (From Davis, 1976.)

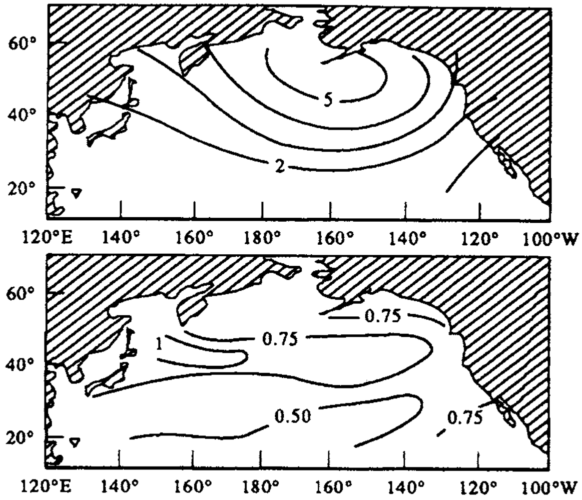

Figure 4.3.5. Standard deviation of: (a) Sea level pressure anomaly (mb); and (b) Sea surface temperature anomaly (°C) for the North Pacific. The anomalies are departures from monthly normal values. Variances are averaged over all months of the 28-year record (1947–1974). (From Davis, 1976.)

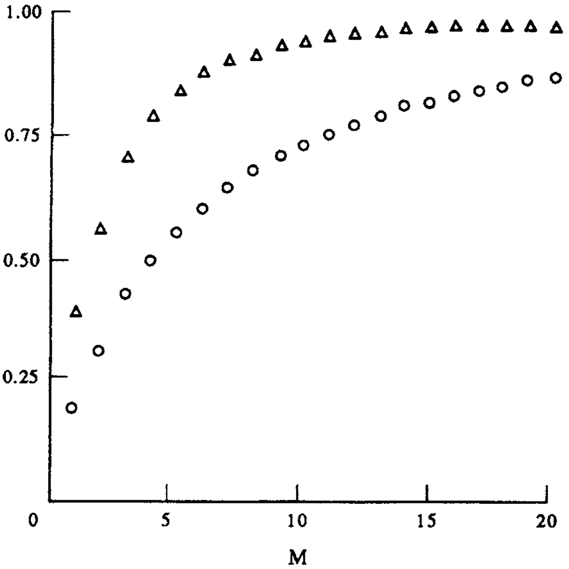

Figure 4.3.6. The fraction of total sea surface temperature (circles, ○) and sea-level pressure (triangles, △) anomaly variance accounted for by the first M empirical orthogonal functions. (From Davis, 1976.)

The standard deviations of both the SLP and SST anomaly fields are shown in Figure 4.3.5. It is interesting to note some of the basic differences between the variability of these two fields. The SLP field has its primary variability in the central northern part of the field just off the tip of the Aleutian Islands. Here, the Aleutian Low dominates the pressure field in winter and becomes the source of the main variability in the SLP data. In contrast, the SST field has near-uniform variance levels except in the Kuroshio Extension region off of northeast Japan where a maximum associated with advection from the Kuroshio is clearly evident.

To compute the EOFs from the anomaly fields, Davis (1976) used the covariance (scatter) matrix method presented in Section 4.3.2. The fraction of total variance accounted for by the EOFs for both the SST and SLP data is presented in Figure 4.3.6 as a function of the number of EOFs. The steep slope of the SLP curve means that fewer SLP EOFs are needed to express the variance. The SST EOF level is consistently below that for the SLP EOF series. As a consequence, Davis presented only the first six SLP EOFs (labeled P_1–P_6 in Figure 4.3.7) but presented the first eight SST EOFs (labeled T_1–T_8 in Figure 4.3.8). The SLP EOFs exhibited fairly simple, large-scale patterns with P_1 having the same basic shape as the SLP standard deviation (Figure 4.3.5). The structural sequence for the first three SLP EOFs was: For P_1, a single maximum; for P_2, two meridionally separated maxima; and for P_3, two zonally separated maxima. Higher modes appear to be combinations of these first three with an increasing number of smaller maxima.

The SST maps obtained by Davis were considerably more complicated than the SLP maps, with large-scale patterns dominating only the first three modes of the temperature field. As with the SLP modes, the sequence seems to be from a central maximum (T_1), to meridionally separated maxima (T_2), and then to zonally separated maxima (T_3). The higher-order EOFs have a number of smaller maxima with no simple structures. The overall scales are much shorter than those for the SLP EOFs. This turns out to be true for the time scales of the EOFs, with the SLP time scales being much shorter than those computed for the SST EOFs.

The goal of the EOF analysis by Davis (1976) was to determine if there is some direct statistical connection between the SLP and SST anomaly fields. By using the EOF procedure he was able to present the primary modes of variability for both fields in the most compact form possible. This is the real advantage of the EOF procedure. In terms of the two anomaly fields, Davis found that there were connections between the variables. First, he found that SST anomalies could be predicted from earlier SST anomaly fields. This is a consequence of the persistence of individual SST patterns as well as the fact that some patterns appear to evolve from earlier patterns through advective processes. Davis also concluded that it was possible to specify the SLP anomaly on the basis of the coincident SST anomaly field. Finally, it was not possible to statistically predict the SST field from the simultaneous SLP field. These conclusions would have been difficult to arrive at without using the EOF procedure.

4.3.6 Variations on conventional EOF analysis

Conventional principal component (EOF) analysis is limited by a number of factors including the dependence of the solution on the domain of analysis, the requirement for orthogonal spatial modes, and the lumping together of variability over all frequency bands. In addition, the method can detect standing waves but not progressive waves. Over the years, several authors have developed what might be called "variations" on the standard EOF theme. For the most part, the methods differ in the types of variances they insert into the algorithms used to determine the empirical orthogonal functions (principal components). Given that EOF analysis is a strictly statistical method, it is irrelevant how the variance is derived, provided that the type of variance used in the analysis is the same for all spatial locations. All that is required is that the matrix **D**, derived from statistical averages (such as the covariance, correlation and cross-covariance functions) of the gridded time series is a Hermitian matrix.

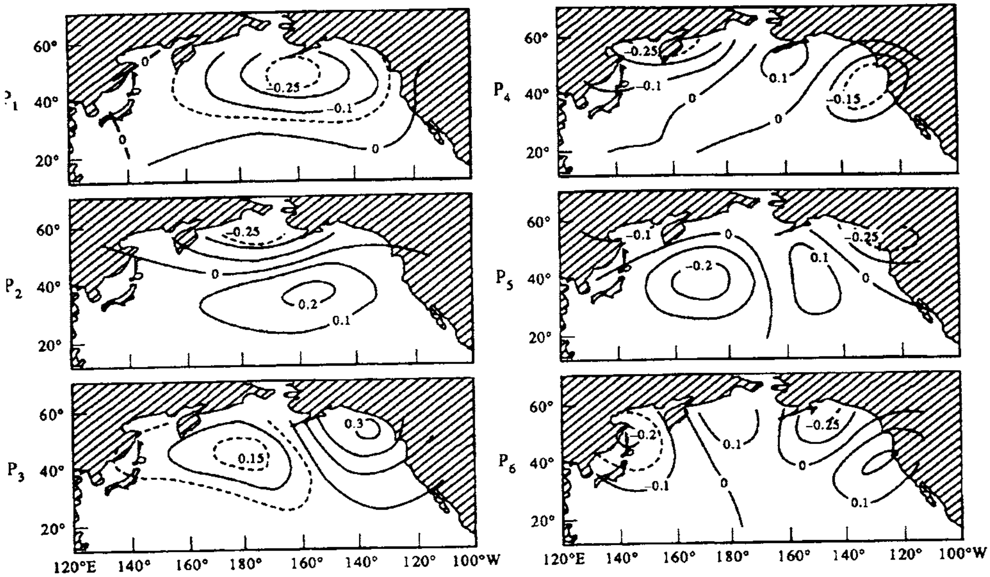

Figure 4.3.7. The six principal empirical orthogonal functions P_1–P_6 describing the sea level pressure anomalies. (From Davis, 1976.)

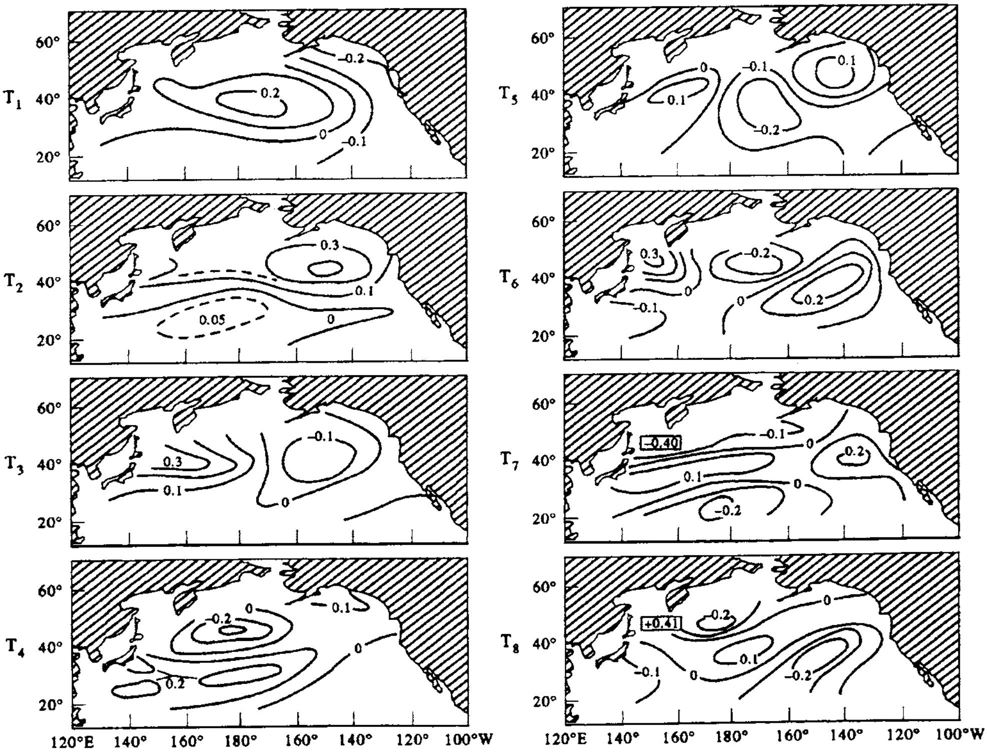

Figure 4.3.8. The eight principal empirical orthogonal functions T_1–T_8 describing the sea surface temperature anomalies. (From Davis, 1976.)

Departure from standard EOF analysis can have numerous forms. For example, one may choose to work in the frequency domain instead of the time domain by using spectral analysis to calculate the spectral "energy" density for specific frequency bands. In this case, the matrix **D** is complex, consisting of the cross-spectra between the gridded time series over a specific frequency band. The spectral densities represent the data variances which are used to determine the empirical orthogonal functions. Thus, the method is equally at home with variances obtained in the time or frequency domains. Regardless of variance-type, principal component methods are simply techniques for compressing the variability of the data set into the fewest possible number of modes.

Returning to the time domain, suppose that we are examining the statistical structure of longshore wind and current fluctuations over the continental shelf and that we have reason to believe that current response to wind forcing is delayed by one or more time steps in the combined data series. A delay of half a pendulum day (≈12 h at mid-latitudes) is not unreasonable. From a causal point of view, the best way to examine the EOF modes for the combined wind and current data is to first create new time series in which the wind records are lagged (shifted forward in time) relative to the current records. Suppose we want a delay of one time step. Then, longshore wind velocity values $V_k(t_j)$ at site k at times t_j ($j = 2, 3, ...$) get replaced with the earlier records at times t_{j-1}. That is, $V_k(t_j) \rightarrow V_k(t_{j-1}) = V_k^*(t_j)$, while the current record remains unchanged, $v_k(t_j) = v_k^*(t_j)$. The asterisk (*) denotes the new time series. Optimal empirical modes are those for which the wind and current records are properly "tuned" with the correct time lags. For large spatial regions with variable wind response times, this can get a little tricky so caution is advised.

A departure from conventional EOF analysis was presented by Kundu and Allen (1976) who combined the zonal (u) and meridional (v) time series of currents into complex time series $w = u + iv$, where each scalar series is defined for times t_j and locations x_k. The method was applied to current data collected during the Coastal Upwelling Experiment (CUE-II) off the Oregon coast in the summer of 1973. The complex covariance matrix obtained from these time series were then decomposed into complex eigenvectors by solving a standard complex eigenvalue problem. Unlike the scalar approach to the problem, this complex EOF technique can be used to describe rotary current variability within selected frequency bands. A further variation on conventional EOF analysis, which is related to complex EOF analysis, was provided by Denbo and Allen (1984). Using a technique we describe in Chapter 5, the current fluctuations in each of the time series (u, v) records collected during CUE-II were decomposed into clockwise (S^+) and counterclockwise (S^-) rotary spectra. The spectra (or variance per unit frequency range) for the dominant spectral components, which is typically $S-$ in the ocean, were then decomposed into empirical orthogonal functions by solving the standard complex eigenvalue problem. This *rotary empirical orthogonal function analysis* is best suited to flows with strong rotary signals such as continental shelf waves and near-inertial motions, but is not well suited to highly rectilinear flows such as those in tidal channels for which S^+ and S^- are of comparable amplitude (see Hsieh, 1986; Denbo and Allen, 1986).

The first use of *complex empirical orthogonal functions* in the frequency domain was described by Wallace and Dickinson (1972) and subsequently used by Wallace (1972) to study long-wave propagation in the tropical atmosphere. Early oceanographic applications are provided by Hogg (1977) for long waves trapped along a continental rise and by Wang and Mooers (1977) for long, coastal trapped waves along a continental margin. In this approach, complex eigenvectors are computed from the cross-spectral matrices for specified frequency bands. This is the most general technique for studying propagating wave phenomena. As noted by Horel (1984), however, EOF analysis in the frequency domain can be cumbersome if applied to time series in which the power of a principal component is spread over a wide range of frequencies as a result of nonstationarity in the data. Horel presents a version of complex EOF analysis in the time domain in which complex time series of a scalar variable are formed from the original time series and their Hilbert transforms. The complex eigenvectors are then determined from the cross-correlation or cross-covariance matrices derived from the complex time series. The Hilbert transform $u_m^H(t)$ of the original time series $u_m(t)$ represents a filtering operation in which the amplitude of each spectral component remains unchanged but the phase of each component is shifted by $\pi/2$. Expanding the scalar time series

$$u_m(t) = \sum_{\omega} [a_m(\omega)\cos(\omega t) + b_m(\omega)\sin(\omega t)] \tag{4.3.50}$$

as a Fourier series over all frequencies, ω, the Hilbert transform $u_m^H(t)$ is

$$u_m^H(t) = \sum_{\omega} [b_m(\omega)\cos(\omega t) - a_m(\omega)\sin(\omega t)] \tag{4.3.51}$$

In practice, the Hilbert transform can be derived directly from the coefficients of the Fourier transform of $u_m(t)$, although with the usual problems caused by aliasing and truncations effects. The complex covariance matrix $r_{mk} = \overline{U_m(t)U_k(t)^*}$ obtained for the series $U_m(t) = u_m(t) + iv_m(t)$ and its complex conjugate, $U_k(t)^*$, are shown to be useful for identifying traveling and standing wave modes; here, (u, v) are the zonal and meridional components of velocity. In the extreme case where the data set is dominated by a single frequency, the frequency domain EOF technique and complex time domain EOF technique are identical. According to Merrifield and Guza (1990), the Hilbert transform complex EOF only makes sense if the frequency distribution in the original time series $u(t)$ is narrow band.

In summary, conventional EOF analysis in the time domain works best when the variance is dominated by standing waves and spread over a wide range of frequencies and wavenumbers. Frequency domain EOF analysis should be used when the dominant variability within the data set is concentrated into narrow frequency bands. Rotary spectral EOF analysis is best used for data sets in which the variance is in narrow frequency bands and dominated by either the clockwise or counterclockwise rotating component of velocity. Complex time domain principal component analysis allows for the detection of propagating wave features (if the process has a narrow frequency band) and the identification of these motions in terms of their spatial and temporal behavior. However, regardless of which method is applied, the best test of a method's validity is whether the results make sense physically and whether the variability is readily visible in the raw time series.

4.4 NORMAL MODE ANALYSIS

In the previous sections, we were concerned with the partition of data variance into an ordered set of spatial and temporal statistical modes. The eigenvalue problem associated with these EOF modes was solved without any consideration given to the underlying physics of the oceanic system. In contrast, normal mode decomposition takes into account the physics and associated boundary conditions of the fluid motion. A common approach is to separate the vertical and horizontal components of the motion and to isolate the forced component of the response from the freely propagating response. As illustrations of these techniques, we consider two basic types of normal mode, eigenvalue problem:

(1) The calculation of vertical normal modes (eigenfunctions), $\psi_k(z)$, for a stratified, hydrostatic fluid with specified top and bottom boundary conditions; and
(2) the derivation of the cross-shore orthogonal modes (eigenfunctions), $\phi_k(x, z)$, for coastal-trapped waves over a variable depth, stratified ocean with or without a coastal boundary.

The first problem can be solved without including the earth's rotation, f, while the second problem requires specification of f. Both eigenvalue problems yield solutions only for certain eigenvalues, λ_k, of the parameter, λ.

4.4.1 Vertical normal modes

A common oceanographic problem is to find the amplitudes (a_k) and phases (θ_k) of a set of K orthogonal basis functions, or modes, by fitting them to a profile of M ($> K$) observed values of amplitude and phase. For instance, one might have observations from $M = 5$ depths and want to find the modal parameters (a_k, θ_k) for the first three theoretical modes, $k = 1, 2, 3$, derived from an analysis of the equations of motion. Once the set of theoretical modes are derived, they can be fitted using a least-squares technique to observations of the along-channel current amplitude and phase. This yields the required estimates, (a_k, θ_k), for $k = 1, 2, 3$.

To obtain the vertical normal modes for a nonrotating fluid ($f = 0$), we assume that the pressure, p, density, ρ , and horizontal and vertical components of velocity (u, v) and w, respectively, can be separated into vertical and horizontal components. This separation of variables has the form

$$[u(\mathbf{x}, t), v(\mathbf{x}, t), p(\mathbf{x}, t)/\rho_o] = \sum_{k=0}^{\infty} p_k(x, y, t)\psi_k(z) \tag{4.4.1a}$$

$$w = \sum_{k=0}^{\infty} w_k \int_{-H}^{z} \psi_k(z)\, \mathrm{d}z \tag{4.4.1b}$$

$$\rho = \sum_{k=0}^{\infty} \rho_k \frac{\mathrm{d}\psi_k(z)}{\mathrm{d}z} \tag{4.4.1c}$$

where $k = 0, 1, 2, \ldots$ is the vertical mode number and the variables without subscripts are functions of ($\mathbf{x}$, t) = (x, y, t). Substituting these expressions into the usual equations of motion (see LeBlond and Mysak, 1979; Kundu, 1990), we obtain the *Sturm–Liouville equation*

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{1}{N^2}\frac{\mathrm{d}\psi_k}{\mathrm{d}z}\right)+\frac{1}{c_k^2}\psi_k=0 \tag{4.4.2}$$

where $N(z) = [-(g/\rho)\ \mathrm{d}\rho/\mathrm{d}z]^{1/2}$ is the Brunt–Väisälä frequency, ${c_k}^2$ is the separation constant and $1/{c_k}^2$ the eigenvalues, λ_k. For a rotating fluid ($f \neq 0$), we assume N(z) is uniform with depth and replace $N^2/{c_k}^2$ in equation (4.4.2) as follows:

$$N^2/c_k^2 \rightarrow (N^2 - \omega^2)/gh_k, \quad k = 1, 2, \ldots \tag{4.4.3a}$$

where h_k is an "equivalent depth", ω is the wave frequency

$$gh_k = (\omega^2 - f^2)/(l^2 + q^2) = c_k^2 - f^2/l^2 \tag{4.4.3b}$$

and (l, q) are the wavenumbers in the horizontal (x, y) directions. Wave-like solutions are possible provided that $f^2 < \omega^2 < N^2$. For a rectangular channel of width L, the cross-channel wavenumber $q \rightarrow q_m = m\pi/L$ and solutions must be considered for both $k, m = 1, 2, \ldots$ (Thomson and Huggett, 1980). For both the rotating and nonrotating case, solutions to the eigenvalue problem (4.4.2) are subject to specified boundary conditions at the seafloor ($z = -H$) and the upper free surface ($z = 0$) of the fluid. These end-point boundary conditions are:

$$\frac{\mathrm{d}\psi_k}{\mathrm{d}z} = 0 \text{ (i.e. } w = 0\text{) at } z = -H \tag{4.4.4a}$$

$$\frac{\mathrm{d}\psi_k}{\mathrm{d}z} + \frac{N^2}{g}\psi_k = 0 \text{ (i.e.} \frac{\partial p}{\partial t} = \rho g w\text{) at } z = 0 \tag{4.4.4b}$$

Modal analysis of the type described by (4.4.2)–(4.4.4) is valid only for an inviscid hydrostatic fluid in which oscillations occur at frequencies much lower than the local buoyancy frequency, N, and for which the vertical length scale is much smaller than the horizontal length scale. In addition, the ocean must be of uniform depth and have no mean current shear. (For sloping bottoms, the horizontal cross-slope velocity component, u, is linked to the vertical boundary, w, through the bottom boundary condition $u = -w\ \mathrm{d}H/\mathrm{d}x$ and separation of variables is not possible.) The method can be applied to an ocean with zero rotation or with rotation that changes linearly with latitude, y. Solutions to (4.4.2) are obtained for specified values of $N(z)$ subject to the surface and bottom boundary conditions. Although the individual orthogonal modes propagate horizontally, the sum of a group of modes can propagate vertically if some of the modes are out of phase.

Analytical solutions: Simple analytical solutions to the Sturm–Liouville equation are obtained with and without rotation when N = constant (density gradient constant with depth). Assuming the rigid lid condition (i.e. no surface gravity waves so that $w = 0$ at $z = 0$), the vertical shapes of the orthogonal eigenfunctions $\psi_k(z)$ in (4.4.2) are given by

$$\psi_k(z) = \cos(k\pi z/H), \quad k = 0, 1, 2, \ldots \tag{4.4.5}$$

where $k = 0$ is the depth-independent barotropic mode, and $k = 1, 2, \ldots$ are the depth-dependent baroclinic modes. The kth mode has k zero crossings over the depth range

$-H \leq z \leq 0$ and satisfies the boundary conditions $w = 0$ (cf. 4.4.1b). Phase speeds (eigenvalues) of the modes are given by

$$c_o = (gH)^{1/2}, \quad k = 0 \text{ (barotropic mode)} \tag{4.4.6a}$$

$$c_k = NH/k\pi, \quad k = 1, 2, \ldots \text{(baroclinic modes)} \tag{4.4.6b}$$

In general, $N(z)$ is nonuniform with depth and, for a given k, the solutions will have the form

$$c_k = (gh_k)^{1/2} \tag{4.4.7}$$

where the "equivalent depth" h_k is used in analogy with H in (4.4.6a). For an ocean of depth $H \approx 2500$ m and buoyancy frequency $N \approx 2 \times 10^{-3}$/s, the eigenvalue for the first baroclinic mode has a phase speed $c_1 \approx 1.6$ m/s and the equivalent depth $h_k = c_1^2/g \approx 0.26$ m. For the 400-m deep tidal channel, we find $N \approx 5 \times 10^{-3}$ m/s, $c_1 \approx 0.8$ m/s and $h_k \approx 0.06$ m.

General solutions: To solve the general eigenvalue problem (4.4.2)–(4.4.4) for variable buoyancy frequency, $N(z)$, we resort to numerical integration techniques for ordinary differential equations with two-point boundary conditions. That is, given the start and end values of the function $\psi_k(z)$, and variable coefficient $N(z)$ we seek values at all points within the domain $(-H \leq z \leq 0)$. Fortunately, there exist numerous packaged programs for finding the eigenvectors and eigenvalues of the Sturm-Liouville equation for specified boundary conditions. The NAG routine D02KEF (Nag Library Routines, 1986) finds the eigenvalues and eigenfunctions (and their derivatives) of a regular singular second-order Sturm–Liouville system of the form

$$\frac{\mathrm{d}}{\mathrm{d}z}\left[F(z)\frac{\mathrm{d}\psi_k}{\mathrm{d}z}\right] + G(z;\,\lambda)\psi_k = 0 \tag{4.4.8}$$

together with boundary conditions

$$z_{a2}\psi_k(z_a) = z_{a1}F(z_a)\,\mathrm{d}\psi_k(z_a)/\mathrm{d}z \tag{4.4.9a}$$

$$z_{b2}\psi_k(z_b) = z_{b1}F(z_b)\,\mathrm{d}\psi_k(z_b)/\mathrm{d}z \tag{4.4.9b}$$

for real-valued functional coefficients F and G on a finite or infinite range, $z_a < z < z_b$. Provision is made for discontinuities in F and G and their derivatives. The following conditions hold on the function coefficients:

(1) The function $F(z)$, which equals $1/N^2(z)$ in the case of (4.4.2), must be nonzero and of one sign throughout the closed interval $z_a < z < z_b$. This is certainly true in a stable oceanic environment where $N^2 > 0$; for $N^2 < 0$, the fluid is gravitationally unstable and vertical modes are not possible;

(2) $\partial G/\partial\lambda$ must be of constant sign and nonzero throughout the interval $z_a < z < z_b$ and for all relevant values λ, and must not be identically zero as z varies for any relevant value of λ.

Numerical solutions to the Sturm–Liouville equation are obtained through a Pruefer transformation of the differential equations and a shooting method. (The shooting method and relaxation methods for the solution of two-point boundary value problems are described in *Numerical Methods* (Press *et al.*, 1992)). The computed eigenvalues are correct to a certain error tolerance specified by the user. Eigenfunctions $\psi_k(z)$ for the problem have increasing numbers of inflection points and zero crossings within the domain $z_a < z < z_b$ as the eigenvalue increases. When the final estimate of λ_k is found by the shooting method, the routine D02KEF integrates the differential equation once more using that value of λ_k and with initial conditions chosen such that the integral

$$I_k = \int_{z_a}^{z_b} [\psi_k(z)]^2 \partial G/\partial\lambda(z;\, \lambda)\, dz \tag{4.4.10}$$

is roughly unity. When $G(z;\, \lambda)$ is of the form $\lambda w(z) + \psi(z)$, which is the most common case, I_k represents the square of the norm of ψ_k induced by the inner product

$$\overline{\psi_k(z)\psi_m(z)} = \int_{z_a}^{z_b} \psi_k(z)\psi_m(z)w(z)\, dz \tag{4.4.11}$$

with respect to which the eigenfunctions are mutually orthogonal if $k \neq m$. This normalization of ψ for $k = m$ is only approximate but typically differs from unity by only a few percent.

If one is working with observed density (σ_t) profiles for the region of interest, a useful approach is to solve the Sturm–Liouville equation using an analytical expression for $N(z)$ by fitting a curve of the type $\sigma_t(z) = [\rho(z) - 1]10^3 = \sigma_o \exp[a/(z+b)]$ or other exponential form, to the data. The eigen (modal) analysis is fairly insensitive to small changes in density so that, even though changes in $N(z)$ are large in the upper oceanic layer, we usually can get away with a simple analytical curve fit. Alternatively, we can specify the actual density on a numerical grid for which modes are to be calculated. Once $N(z)$ is available, we can use numerical methods to solve (4.4.2) subject to the boundary conditions (4.4.4), allowing for specified error bounds or degree of convergence on the final boundary estimate. Based on the analytical solutions (4.4.5), we can expect solutions ψ_k to resemble cosine functions whose vertical structure has been distorted by the nonuniform distribution of density along the vertical profile. There is a direct analogy here with the modes of oscillation of a taut string clamped at either end and having a nonuniform mass distribution along its length.

The normal modes are normalized relative to their maximum value and then fitted to the data in a least-squares sense (Table 4.4.1). If there are M current meters on a mooring string, the maximum possible number of normal baroclinic modes is $M - 1$. By comparing the normal modes with the data, we can derive the absolute values of the barotropic mode and a maximum of $M - 1$ baroclinic modes. Solutions to the least-squares fitting are described in (Press *et al.*, 1992).

4.4.2 An example: normal modes of semidiurnal frequency

Suppose that the along-axis semidiurnal currents, u, in a tidal channel have the form $u_m = a_m \cos(\omega_t + \theta_m)$, where a_m, θ_m $(m = 1, \ldots, M)$ are the observed current amplitude and phase, respectively. In terms of tidal current ellipses, we can think of u as the major axis of the current ellipse for each current meter on the mooring line. The oscillations have frequency $\omega = \omega_{M_2}$ corresponding to M_2 semidiurnal tidal currents and the phase θ is referenced to some time zone or meridian of longitude so that we can intercompare values for different current meters and the surface tides. The values a_m, θ_m for the different current meter records can be determined using harmonic analysis techniques (Foreman, 1976) provided the measured data are at hourly intervals over a period of seven days or longer so that the M_2 and K_1 constituents are separable. We next rewrite the above expression for u in the usual way as $u_m = A_m \cos(\omega_t) + B_m \sin(\omega_t)$, where $\tan\theta_m = A_m/B_m$ and $a_m^2 = (A_m^2 + B_m^2)$. This allows us to examine the sine and cosine components separately. The observed magnitudes A_m and B_m at each current meter depth z_m, $m = 1, \ldots, M$ are then used to compute the amplitudes and phases of the basis functions $\psi_k(z_m)$, for a maximum of K different modes ($K < M$). At best, we can obtain the amplitudes and phases of the barotropic mode ($k = 0$) and up to $M - 1$ baroclinic modes.

Details of the modal analysis at semidiurnal frequency using current meter data from a tidal channel are presented by Thomson and Huggett (1980). The first step is to obtain an exponential functional fit (Figure 4.4.1a) to the observed mean density structure, $N(z)$. This structure is then used with the local water depth H (assuming a flat bottom), the Coriolis parameter, f, and the wave frequency, ω, to calculate the theoretical dynamic modes (Figure 4.4.1b). A finite sum of these theoretical modes $\sum \psi_k(z)$ is then least-squares fitted to the observed cosine component $A_m(z)$ to obtain estimates of the contributions A_k from each mode, k. This operation is repeated for the sine component B_k. (Recall that the maximum total of barotropic plus baroclinic modes allowed in the summation is fewer than the number of current meter records per mooring string and that the vertical structure of each mode is found through the products $(A_k, B_k)\psi_k(z)$ where the coefficients are constant.) Using the relationships $\tan\theta_k = A_k/B_k$ and $a_k^2 = (A_k^2 + B_k^2)$, we get the amplitudes and phases of the various modes. In their analysis, Thomson and Huggett (1980) typically had only three reliable current meter records per mooring string. Normally, this would be enough to obtain the first two baroclinic modes. However, the bottom current meter in most instances was within a few meters of the bottom and therefore strongly affected by benthic boundary layer effects. To include a mode-2 solution in the estimates, the observed phase and amplitude of the bottom current meter record had to be adjusted for frictional effects via the added term $\exp(-z')\cos(\omega t + \theta - z')$, where

Table 4.4.1. Modal amplitudes (cm/s) and phases (degrees relative to 120°W longitude) for Johnstone Strait M_2 tidal currents computed from nine-day current meter records. Column 2 gives the number of current meters (M) on the string. The first column for the barotropic mode (a_0, θ_0) and each of the two baroclinic modes (a_k, θ_k), $k = 1, 2$, gives the amplitude and phase (a, θ) before and after the bottom current meter is included in the analysis. The bottom current is included after its amplitude and phase are corrected for bottom boundary layer friction. The vertical eddy viscosity K_v is that value which gives the minimum ratio between the first and second baroclinic modes when the frictionally corrected bottom current meter is included. NC means "no change", implying perfect modal fit for all depths with and without the bottom current meter record. At CM04, no near-surface current meter was deployed and the records were only five days long and therefore suspect

Site	M	K_v (cm^2/s)	Before (a_0, θ_0)	After (a_0, θ_0)	Before (a_1, θ_1)	After (a_1, θ_1)	Before (a_2, θ_2)	After (a_2, θ_2)
CM13	3	15	42, 55°	42, 55°	12, 172°	25, 171°	–	19, –10°
CM14	3	8	35, 51°	35, 51°	11, 169°	15, 171°	–	8, –4°
CM15	3	13	32, 35°	32, 36°	18, 175°	12, 166°	–	7, –31°
CM02	5	0	36, 42°	NC	21, 220°	NC	9, 13°	NC
CM04	4	7	29, 45°	50, 24°	13, 215°	79, 174°	2, –34°	70, 0°

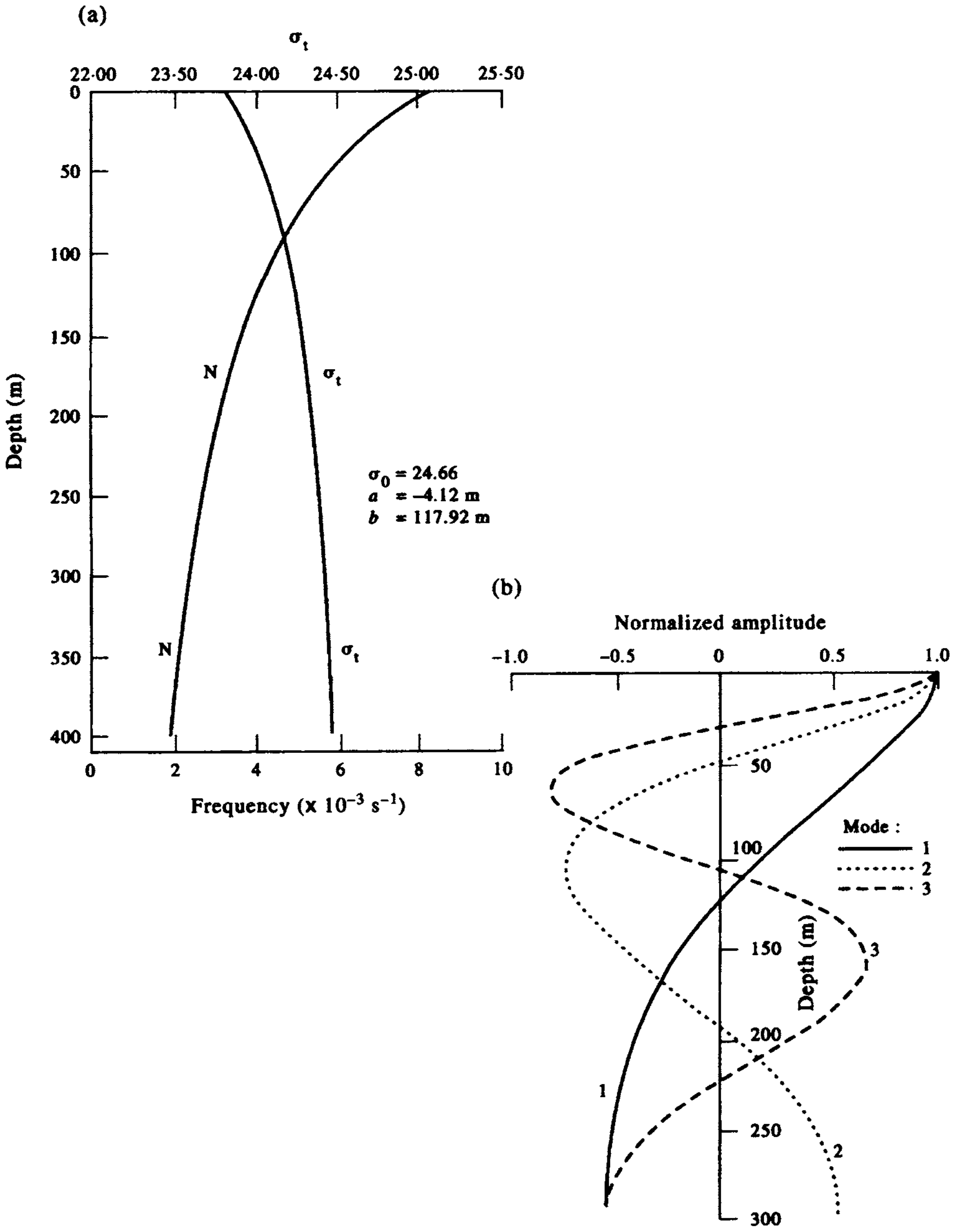

Figure 4.4.1. Baroclinic modes for semidiurnal frequency (ω_{M_2}) in a uniformly rotating, uniform depth channel. (a) The mean density structure (σ_t) and corresponding buoyancy frequency N(z) used to calculate the eigenvalues; (b) Eigenvectors for the first three baroclinic modes. The barotropic mode (not plotted) has a magnitude of unity at all depths. Phase speeds for the modes fitted to the current meter data are $c_1 \approx 34$ cm/s; $c_2 \approx 20$ cm/s. (From Thomson and Huggett, 1980.)

$z' = (z + H)/\delta$, and $\delta \approx (2K_v/\omega)^{1/2}$ is the boundary layer thickness for eddy viscosity K_v. Since K_v is not known *a priori*, the final solution required finding that value of K_v which minimized the ratio formed by the first mode calculated with and without the bottom current meter included in the analysis (Table 4.4.1). In the case where five current meters were available, Thomson and Huggett found that there was no difference in the value of the second mode estimate with and without inclusion of the bottom current meter record in the analysis, suggesting that the three-mode decomposition was representative of the actual current variability with depth.

4.4.3 Coastal-trapped waves (CTWs)

Stratified or nonstratified oceanic regions characterized by abrupt bottom topography adjacent to deeper regions of uniform depth support the propagation of trapped ocean waves with frequencies, ω, which are lower than the local inertial frequency, f. Trapped sub-inertial motions ($\omega < f$) typically are found along continental margins where the coastal boundary is bordered by a marked change in water depth consisting of a shallow (< 200 m) continental shelf, a steep continental slope, and a deep (> 2000 m) weakly sloping continental rise. The longshore wavelengths vary from tens to thousands of kilometers while the cross-shore trapping scale is determined by the density structure and length scale of the topography. For baroclinic waves, the *internal deformation radius* $r = NH/f$ provides an estimate of the cross-shelf trapping scale while the *stratification parameter* $S = (N^2{}_{\rm max}\ H^2{}_{\rm max})/f^2L^2$ measures the importance of stratification for a shelf-slope region of width L. For a mid-latitude ocean of depth $H \approx 2500m$ and buoyancy frequency $N \approx 2 \times 10^{-3}$/s, we find $r \approx 50\,$km. For wide shelves (L > 100 km), the motions are confined mainly to the continental slope, while for narrower shelf regions the motions extend to the coast where they "lean" up against the coastal boundary. For $S \gg 1$ the CTWs are strongly baroclinic, while for $S \ll 1$, they are mainly barotropic (Chapman, 1983). The case $S \approx 1$ corresponds to barotropic shelf waves modified by stratification.

In addition to continental shelf regions, coastal-trapped waves can occur along mid-ocean ridges and in oceanic trenches (where they are known as *trench waves*), as well as around isolated seamounts. Phase propagation, in all cases, is with the coastal boundary to the right of the direction of propagation in the Northern Hemisphere and to the left of the direction of propagation in the Southern Hemisphere. For strongly baroclinic waves, energy propagation is always in the direction of phase propagation; for barotropic motions, short waves can propagate energy in the opposite direction to phase propagation.

The general coastal-trapped wave solution consists of a Kelvin wave mode ($k = 0$), for which the cross-shore velocity component is identically zero at the coast ($U \equiv 0$ at $x = 0$), together with a hierarchy of higher mode shelf waves ($k = 1, 2, \ldots$) whose cross-shore velocity structures have increasing numbers of zero crossings (sign changes) normal to coast. The first shelf wave mode will have one zero crossing in elevation ζ over the continental margin, the second mode will have two crossings, and so on. For the current component, U, the first mode shelf wave will have no zero crossing, the second mode will have one crossing, and so on. The condition of no normal flow through the coastal boundary requires $U = 0$ at $x = 0$.

Computer programs that calculate the frequencies and cross-shore modal structure of coastal-trapped waves of specified wavelength are available in reports written by Brink and Chapman (1987) and Wilkin (1987). We confine ourselves to a general outline of the programs for the interested reader. Practical difficulties with the numerical solutions to the equations are provided in these comprehensive reports. The programs of Brink and Chapman use linear wave dynamics in which the water depth, $h(x)$, is assumed to be a function of the cross-shore coordinate, x, alone. Similarly, the buoyancy frequency, $N(z)$, is a function of depth alone. The one profile that can be used in the analysis is best obtained by least-squares fitting a function (such as a polynomial or exponential) to a series of observed profiles. The wave parameters such as velocity, pressure and density are assumed to be sinusoidal in time (t) and longshore direction (y) such that for any particular wave parameter, ξ, we have

$$\xi(x, y, t) = \xi_o(x)\exp[\mathrm{i}(\omega t + ly)] \tag{4.4.12}$$

where ω is the wave frequency and l is the alongshore wavenumber. This gives rise to a two-dimensional eigenvalue problem in (ω, l) of the form

$$L[\xi_o(x; \omega, l)] = 0 \tag{4.4.13}$$

where L is a linear operator. The problem is solved for arbitrary forcing and a fixed l. In particular, for a given wavenumer, k, the frequency ω is varied until the algorithm finds the free-wave mode resonance. Resonance is defined as the frequency at which the square of the spatially integrated wave variable

$$I_v = \int_0^{\infty} \xi_o^2 \, \mathrm{d}x, \quad \text{or } I_p = \int_0^{\infty} \int_{-h}^{0} p^2 \, \mathrm{d}z \, \mathrm{d}x \tag{4.4.14}$$

is at a maximum. The suite of programs tackle the following problems for which the user provides the bottom profile $h(x)$, a mean flow profile (if needed) and a selection of boundary conditions:

(1) The program BTCSW yields the dispersion curves $\omega = \omega(l)$ (the frequency as a function of wavenumber), the cross-shore modal structure fo\ velocity $U(x)$ and/or surface elevation $\zeta(x)$, and wind coupling coefficients fo, oarotropic coastal-trapped waves—including continental shelf waves and trench waves—for arbitrary topography and mean longshore current. Options for the long-wave and rigid-lid approximations are included in the program. The user can specify one of two geometries corresponding to topography with and without a coastal boundary. The outer boundary $x = x_{\max}$ is set as $-2L$, where L is the width of the typographically varying domain in the cross-shore direction. Thus, about half the domain has a flat bottom. The outer boundary condition is specified as $\partial U/\partial x = 0$. To obtain solutions for both ζ and U, the depth at the coast should be given a nonzero value $h(0) \geq 1$ m.

(2) For wave frequencies $\omega \leq 0.9f$, the program BIGLOAD2 yields dispersion curves $\omega = \omega(l)$, the horizontal modal structure, and wind-coupling coefficients for an ocean with continuous, horizontally uniform stratification and arbitrary topography. Density in the model has the form $\rho^*(x, y, z, t) = \rho_o(z) + \rho(x, y, z, t)$, where ρ_o is background density and ρ is the density perturbation. Since $\rho \ll \rho_o$, the Boussinesq approximation is assumed throughout (i.e. density perturbations are ignored except where they multiply gravity, g, the acceleration due to gravity). The program allows for the component of the β-effect normal to the coast and for both the free surface and rigid lid boundary conditions at the

ocean surface. Solutions are obtained using the coordinate transformation $\theta = z/h(x)$ and assuming a linear bottom friction drag. A total of 17 vertical and 25 horizontal grids (rectangles) are generated so that the vertical resolution is much better near shore than in deep water. Problems with singularities are avoided by setting $h(x) \geq 1$ m at the coast, $x = 0$. The program does not work well when the shelf-slope width (or width of a trench at the base of the shelf) is small relative to the internal deformation radius for the first mode in the deep ocean. Spurious features appear in unexpected places and force the user to increase the density of horizontal grids over regions of rapidly varying topography. In addition, a spurious mode occurs in the pressure equation for $\beta = 0$ at the local inertial frequency $\omega = f$, making the overall solution suspect. As noted by the authors, the user will have difficulty finding the barotropic Kelvin wave parameters.

(3) The program CROSS is used to find baroclinic coastal-trapped modes for $\omega \leq f$ for arbitrary stratification and uniform depth.

(4) The program BIGDRV2 is used to obtain the velocity, pressure, and density fluctuations over a continental shelf-slope region of arbitrary depth, stratification, and bottom friction and is driven by a longshore wind stress of the form $\tau(x) = \tau_0 \exp[i(\omega t + ly)]$. Specification of a linear friction coefficient of zero (r = 0) results in a divide by zero error. As a result, inviscid solutions should not be attempted. As with (2), solutions are obtained on a 25 × 17 stretched grid. In practice, it is generally best to start a study of coastally trapped waves using BTCSW since it gives first-order insight into the type of modal structure one can expect. However, if the barotropic dispersion curves do not fit the data (e.g. observations reveal strong diurnal-period shelf waves but the first-mode dispersion curves consistently remain below the diurnal frequency band for realistic topography), then density and mean currents should be introduced using BIGLOAD2 and CROSS.

The Brink and Chapman programs have been used by Crawford and Thomson (1984) to examine free wave propagation along the west coast of Canada and by Church *et al.* (1986) and Freeland *et al.* (1986) to examine wind-forced coastal-trapped waves along the southeast coast of Australia (Figure 4.4.2). In all cases, model results are compared with longshore sea-level records and current meter observations from cross-shore mooring lines. The cross-shore depth profiles $h(x)$ and associated buoyancy frequencies $N^2(z)$ used in the Australian model are presented in Figures 4.4.3(a, b). From these input parameters, the program was used to generate eigenvalues and eigenfunctions for the first three CTW wave modes (Figure 4.4.4) and the theoretical dispersion curves (Figure 4.4.5) relating wave frequency, ω, to longshore wavenumber, l. The slopes of the (ω, l) curves give the phase speeds c_k for the given modes (k = 1, 2, 3) listed on the figure.

Wilkin (1987) presents a series of FORTRAN programs for computing the frequencies and cross-shore modal structure of free coastal-trapped waves in a stratified, rotating channel with arbitrary bottom topography. The programs solve the linearized, inviscid, hydrostatic equations of motion using the Boussinesq approximation. The Brunt-Väisälä frequency $N(z)$ is a function of the vertical coordinate only. As with Brink and Chapman (1987), the eigenvalue problem is solved using resonance iteration and finite difference equations. The cross-shore perturbation fields returned by the model include velocity, pressure, and density. The difference with Wilkin's model

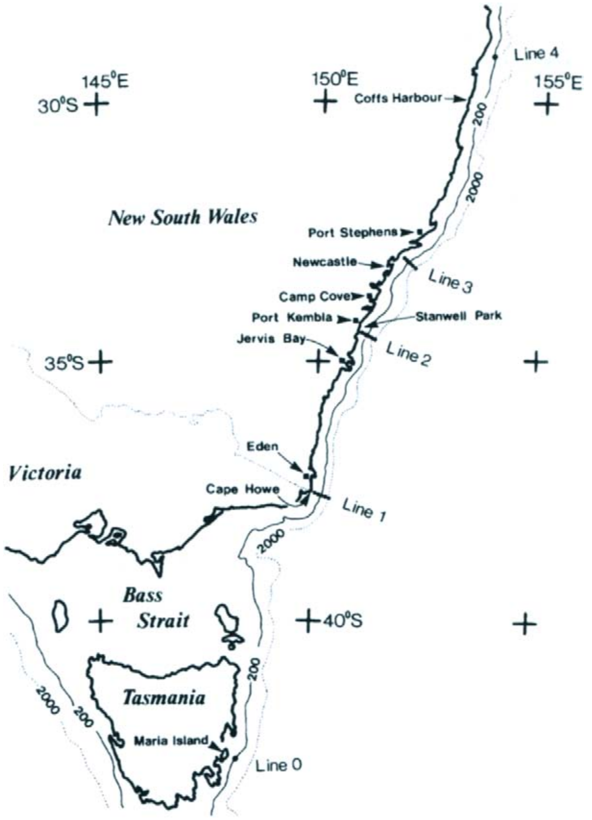

Figure 4.4.2. Southwest coast of Australia showing the locations of the tide gauge stations (■) and current meter lines (0, 1, 2, 3) occupied during the Australian Coastal Experiment (ACE). (From Freeland et al., 1986.)

is that it uses a staggered horizontal (Arakawa "C") grid for which the usual horizontal Cartesian coordinates (x, y) have been mapped to orthogonal curvilinear co-ordinates (ξ, η). Instead of using finite differencing, the vertical structures of the modes are determined through modified sigma coordinates with expansion of the field variables in terms of Chebyshev polynomials of the first kind. The program has the option of specifying wavenumber, l, and searching for the corresponding free wave frequency, $\omega(l)$, as in Brink and Chapman, or specifying ω and searching for l. For reasons explained by Wilkin, the model is designed to be compatible with the primitive equation ocean circulation model developed by Haidvogel *et al.* (1988).

In the curvilinear coordinate system, a line element of length ds in the Wilkin model satisfies

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 = \mathrm{d}\xi^2/\mathrm{d}m^2 + \mathrm{d}\eta^2/\mathrm{d}n^2 \tag{4.4.15}$$

and the metric coefficients m, n are defined by

$$m = \left[(\partial x/\partial\xi)^2 + (\partial y/\partial\xi)^2\right]^{-1/2} \tag{4.4.16a}$$

$$n = \left[(\partial x/\partial\eta)^2 + (\partial y/\partial\eta)^2\right]^{-1/2} \tag{4.4.16b}$$

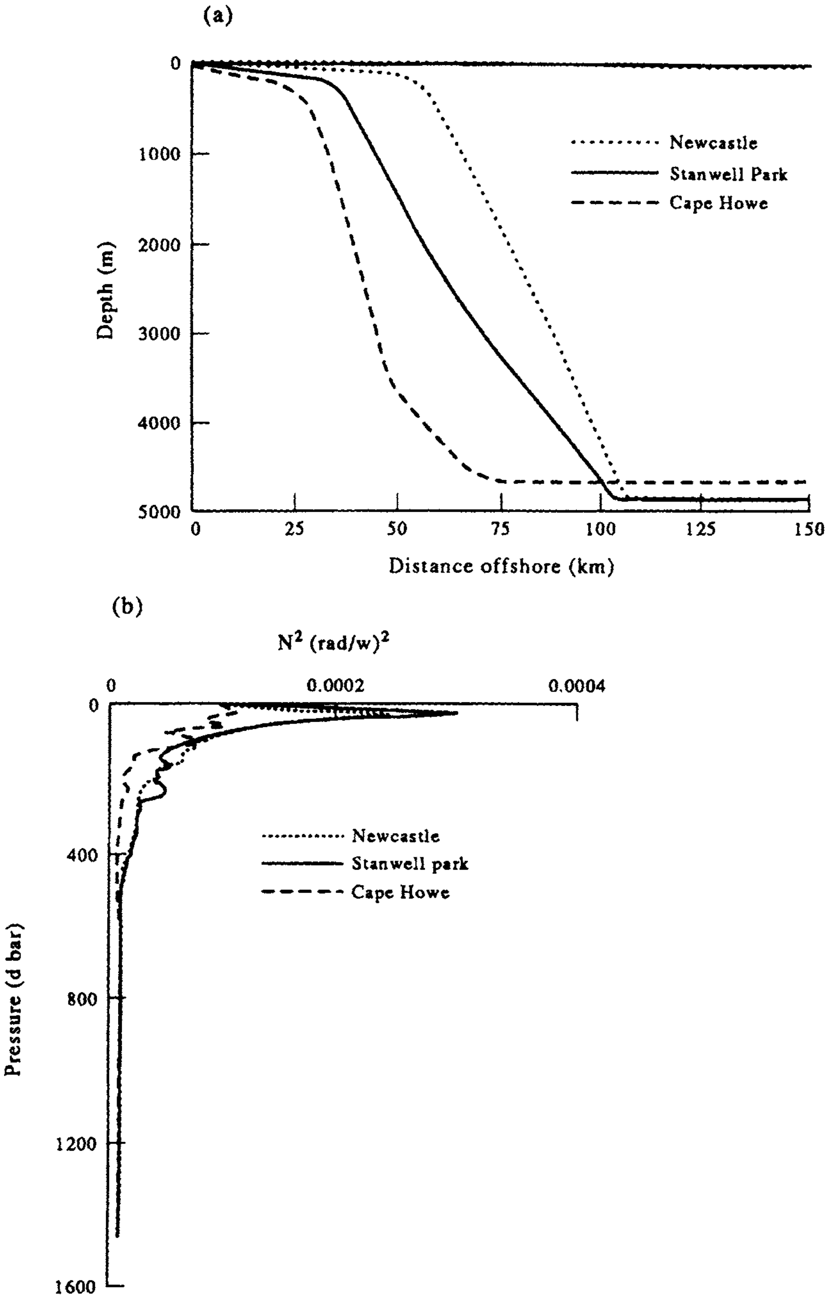

Figure 4.4.3. Parameters used in determining the coastal-trapped wave eigenfunctions at Cape Howe Stanwell Park and Newcastle; (a) The cross-shore depth profiles h(x); (b) The N(z)^2 profiles. Below 600 db (≈ 590 m) all curves are similar so that only one is drawn. (From Church et al., 1986.)

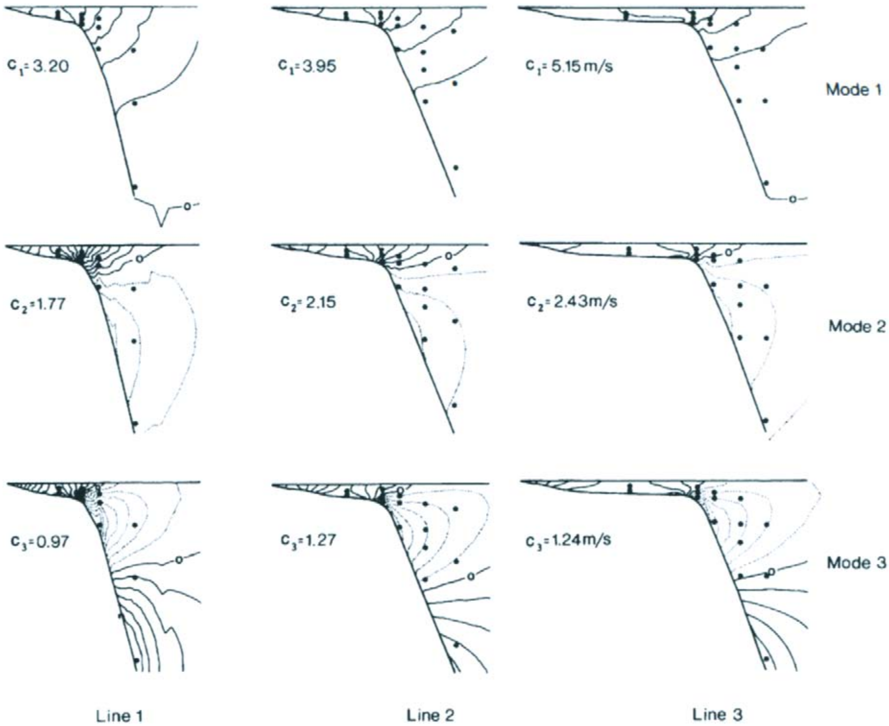

Figure 4.4.4. The eigenfunctions $U(x; z)$ for the first three baroclinic longshore current modes for the three lines in Figures 4.4.2 and parameters in Figure 4.4.3. The contouring is in arbitrary units. Phase speeds c_k (eigenvalues) of each mode for each of the three lines also are shown. (From Church et al., 1986.)

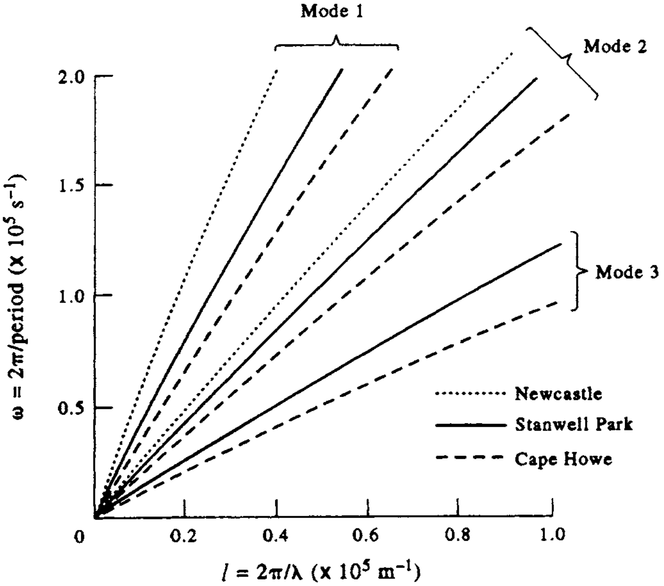

Figure 4.4.5. The theoretical dispersion curves $\omega = \omega(l)$ relating the longshore wavenumber, l, to the wave frequency, ω (here, λ is the wavelength). Curves correspond to the first three baroclinic modes for each mooring location. For mode 3, the dispersion curve at Stanwell Park and Newcastle are almost identical. The slopes of the lines are the theoretical phase speeds, c_k. (From Church et al., 1986.)

The velocity perturbations for time-dependent solutions of the form $\exp(-\mathrm{i}\omega t)$ are then

$$u = \frac{1}{f^2 - \omega^2}\left(\mathrm{i}\omega m \frac{\partial \phi}{\partial \xi} - fn \frac{\partial \phi}{\partial \eta}\right) \tag{4.4.17a}$$

$$v = \frac{1}{f^2 - \omega^2}\left(\mathrm{i}\omega n \frac{\partial \phi}{\partial \eta} - fm \frac{\partial \phi}{\partial \xi}\right) \tag{4.4.17b}$$

$$w = \frac{\mathrm{i}\omega}{N^2}\frac{\partial \phi}{\partial z} \tag{4.4.17c}$$

where (u, v, w) are the usual velocity components and $\phi = p/\rho_o$ is the perturbation pressure. Solutions are then sought for the resulting pressure equation

$$mn\frac{\partial}{\partial \eta}\left(\frac{n}{m}\frac{\partial \phi}{\partial \eta}\right) + (f^2 - \omega^2)\frac{\partial}{\partial z}\left(\frac{1}{N^2}\frac{\partial \phi}{\partial z}\right) + mn\frac{\partial}{\partial \xi}\left(\frac{m}{n}\frac{\partial \phi}{\partial \xi}\right) = 0 \tag{4.4.18}$$

For a straight coastline, $m\partial/\partial\xi = \partial/\partial x$ and we arrive at the usual solutions for longshore (x-direction) propagation of progressive waves of the form $F(y)\exp[\mathrm{i}(lx - \omega t)]$.

The Wilkin model is less general than the Brink and Chapman model in that application of the rigid-lid approximation does not allow for the barotropic (longwave) Kelvin wave solution and a "slippery" solid wall is placed at the offshore boundary. The new vertical coordinate variable, σ , is defined by

$$\sigma = 1 + 2z/h(\eta) \tag{4.4.19}$$

so that the ocean surface is located at $\sigma = 1$ and the (now flattened) seafloor at $\sigma = -1$. Application of this model to the west coast of New Zealand (South Island) is presented by Cahill *et al.* (1991). Modes 1 and 2 of the longshore current for the northern portion of this region based on Wilkin's program CTWEIG are reproduced in Figure 4.4.6. Similar results for the southern region are presented in Figure 4.4.7. Notice that the coastal-trapped waves are nearly barotropic over the shallow shelf immediately seaward of the coast in both sections but are more baroclinic in the offshore region off the southwest coast.

4.5 INVERSE METHODS

4.5.1 General inverse theory

General inverse methods have become a sophisticated analysis tool in the earth sciences. For example, in the field of geophysics, a goal of this technique is to infer the internal structure of the earth from the measurement of seismic waves. The essence of the geophysical *inverse problem* is to find an earth structure model which could have generated the observed acoustic travel-time data. This is in contrast to the *forward problem* which uses a known input and an understood physical system to predict the output. In the inverse problem, the input and output are known and the result is the *model* required to translate one set of data into the other.

In oceanography, inverse methods are used for a variety of applications, including the inference of absolute ocean currents using known tracer distributions and

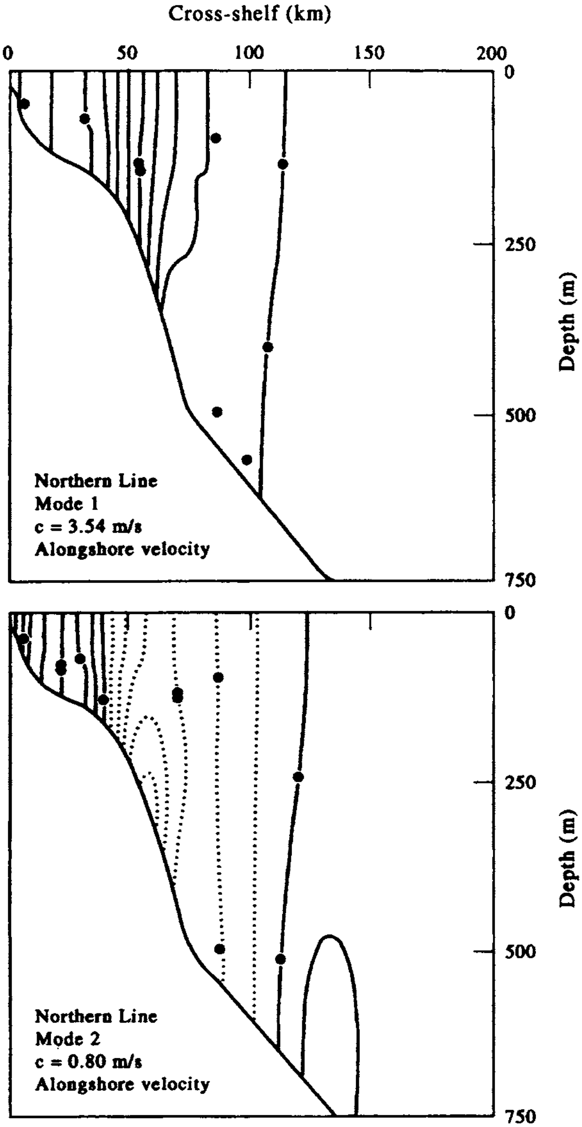

Figure 4.4.6. The longshore velocity structure of coastal-trapped waves for the northwestern shelf-slope region of South Island, New Zealand. (a) Mode 1; (b) Mode 2. Contour lines when multiplied by 10^{-7} correspond to the longshore velocities in m/s for unit energy flux in watts. Negative values are dashed. Current meter locations are given by the dots. (From Cahill et al., 1991.)

geostrophic flow dynamics (Wunsch, 1978, 1988). Another application uses underwater acoustic travel times to determine the average temperature of the global ocean for long-term climate studies (Worchester *et al.*, 1988). A study by Mackas *et al.* (1987) used inverse techniques to determine the origins and mixing of water masses off the coast of British Columbia. In these oceanographic applications, the "solutions" are what we previously called the "models" in the geophysical problem. The kernel functions are formulated from the physics of the problem in question and the result is found by matching the "solution" to the input data. A cursory look at the problem is

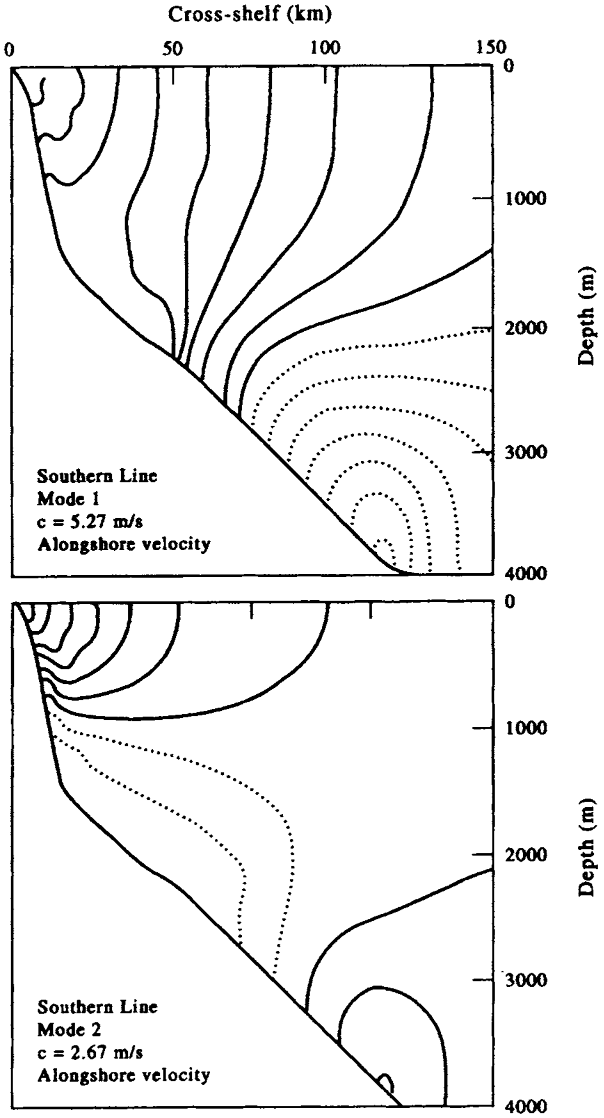

Figure 4.4.7. As for Figure 4.4.6 but for the shelf-slope region off the southwestern tip of South Island. Note the change in depth and offshore distance scale in the two figures. This line is roughly 500 km to the south of the line in Figure 4.4.6.

provided in this section. The interested reader is referred to Bennett (1992) for detailed insight into the theory and application of inverse methods in oceanography.

In general, the inverse problem takes the form

$$e(t) = \int_a^b C(t, \xi) m(\xi) \, d\xi \tag{4.5.1}$$

where $e(t)$ are the input data, $m(\xi)$ is the model and $C(t, \xi)$ is the kernel function for the variable ξ. The kernel functions are determined from the relevant physical equations for the problem and are assumed to be known (Oldenburg, 1984). It is the judicious selection of these kernel functions that makes the inverse problem a complex exercise requiring physical insight from the oceanographer. In order to extract information about the model, $m(\xi)$, we will restrict our consideration of inverse theory to linear inverse methods applied to a set of observations. This is referred to as "finite dimensional inverse theory" by Bennett (1992). In his discussion of this form of inverse theory, Bennett suggests that it applies to:

(1) An incomplete ocean model, based on physical laws but possessing multiple solutions.
(2) Measurements of quantities not included in the original model but related to the model by additional physical laws.
(3) Inequality constraints on the model fields or the data.
(4) Prior estimates of errors in the physical laws and the data.
(5) Analysis of the level of information in the system of physical laws, measurements, and inequalities.

Equation (4.5.1) is a *Fredholm equation* of the first kind. Inverse theory is centered around solving this equation in such a way as to extract information about the model, $m(\xi)$, when information is available for the data, $e(t)$. It is important to realize that the inverse problem cannot be solved unless the physics and the geometry of the problem are known (i.e. equation (4.5.1) has been set up). It is, therefore, impossible to consider a solution to the inverse problem unless the forward problem can be solved. The physics of the forward problem may be ill-posed, in which case not all of the solutions will match or, if they do, it is a coincidence and not a solution to (4.5.1). Thus, the basic questions to ask regarding a solution of the inverse problem are: (1) Does a solution exist? In other words, is there an $m(\xi)$ which produces $e(t)$?; (2) How does one construct a solution?; (3) Is the solution unique?; and (4) How is the nonuniqueness appraised?

The answers to the above questions will depend on the data, $e(t)$. In theory, there exist three types of data:

(1) An infinite amount of accurate data;
(2) a finite amount of accurate data;
(3) a finite amount of inaccurate data.

In reality, only option (3) occurs as we are forced to work with observations which contain a variety of measurement and sampling errors. While perfect data are limited to the realm of the mathematical, it is often instructive to consider analytic "inverses". For example, the analytical inverse to

$$x(f) = \int_{-\infty}^{\infty} x(t)\mathrm{e}^{-i2\pi ft}\,\mathrm{d}t \qquad (4.5.2)$$

is

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} x(f)\mathrm{e}^{i2\pi ft}\,\mathrm{d}t \qquad (4.5.3)$$

Similarly, the inverse of

$$\phi(x) = 2/\lambda \int_{x}^{a} \left[r\varepsilon(r)/(r^2 - x^2)^{1/2}\right] \mathrm{d}r \tag{4.5.4}$$

is

$$\varepsilon(r) = -\lambda/\pi \int_{r}^{a} \left[(\mathrm{d}\phi/\mathrm{d}x)/(x^2 - r^2)^{1/2}\right] \mathrm{d}x \tag{4.5.5}$$

In the second case, we require knowledge of $\mathrm{d}\phi/\mathrm{d}x$ to find $\varepsilon(r)$, which is easy to do for ideal continuous data (Figure 4.5.1a), or even for a finite sample of accurate data (Figure 4.5.1b). If, however, we have a finite sample of inaccurate data (Figure 4.5.1c), we have difficulty estimating $\mathrm{d}\phi/\mathrm{d}x$.

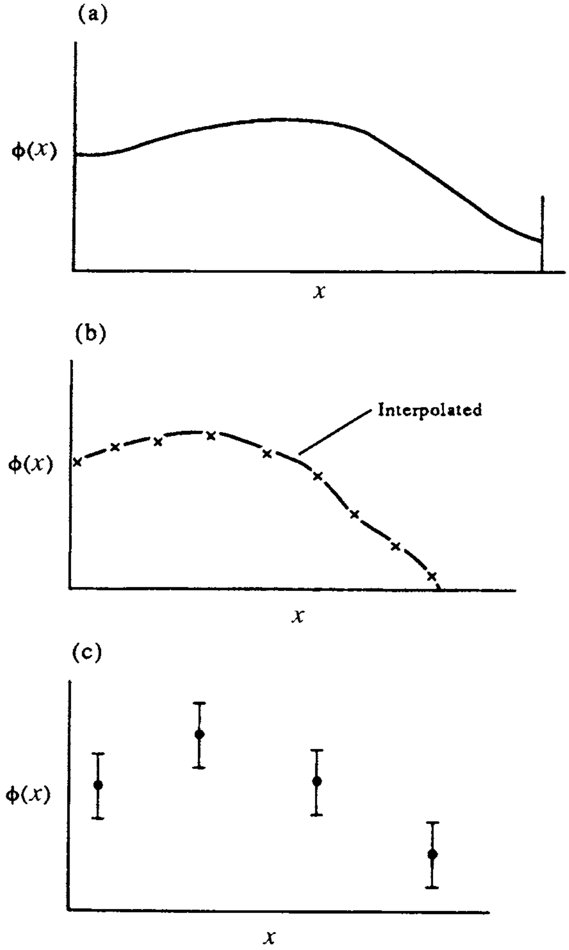

Figure 4.5.1. Three examples of the function $\phi(x)$ required for the inverse solution, $\varepsilon(r)$, of equation (4.5.5). Analytical (a) and digital (b) versions of $\phi(x)$ for which inversion is readily possible. (c) A typical "observed" version of $\phi(x)$, consisting of four mean values (plus standard deviations) for which inversion is considerably less accurate.

The problem of dealing with a limited sample of inaccurate measurements is the most common obstacle to the application of inverse methods. Usually, these inaccuracies can be treated as additive noise superimposed on the true data and, therefore, can be handled with statistical techniques. These additive errors have the effect of "blurring" or distorting our picture of the solution (model). Unfortunately, one cannot conclude that if the error noise is small that the model distortions also will be small. The reason for this is that most geophysical kernel functions act to smooth the model, thus changing the length scale of the response for both the forward and inverse problems. In other words, the solution obtained with inaccurate data using the inverse procedure may be very different from the model which actually generated the data. In addition, particular solutions to the model are not unique and a wide variety of solutions is equally possible.

In most oceanographic applications of inverse methods, we are primarily interested in finding a model which reproduces the observations. Here, the fundamental problem is the nonuniqueness of any inverse solution which is one of infinitely many functions that can reproduce a finite number of observations. This nonuniqueness becomes more severe when the data are inaccurate, as they must be in any practical oceanographic application. The key to the application of inverse methods in oceanography is to select the "correct" (by which we mean the most probable or the most reasonable) inverse model-solution.

Inverse construction in oceanography may take the form of parametric modeling. In this case, we write our model as $m = f(a_1, a_2, \ldots, a_N)$ and a numerical scheme is sought to find appropriate values of the parameters, a_i ($i = 1, \ldots, N$). Parameterization is justified when the physical system actually has this form and depends on a number of input parameters. The model is solved by collecting more than N data points and finding the parameters through a least-squares minimization of

$$\phi = \sum_{i=1}^{N} (e_i - e_i')^2 \tag{4.5.6}$$

where

$$e_i' = f(a_1, a_2, \ldots, a_N; \varepsilon_i) \tag{4.5.7}$$

In (4.5.7) ε_i is the ith kernel function.

4.5.2 Inverse theory and absolute currents

As reviewed by Bennett (1992), an important application of inverse theory to ocean processes was the computation of absolute currents for large-scale ocean circulation. In the 1970s, two different approaches to this problem were proposed. The first by Stommel and Schott (1977) was called the "beta spiral" technique, which demonstrated that the vertical structure of large-scale, open-ocean velocity fields could be explained using simple equations expressing geostrophy and continuity (conservation of mass). The second method, introduced by Wunsch (1977), showed that reference velocities could be estimated simultaneously around a closed path in the ocean. The resultant absolute velocities were consistent with geostrophy and the conservation of heat and salt at various levels. As a guide to oceanographic applications of inverse techniques, we provide succinct reviews of both applications.

4.5.2.1 The beta spiral method

Good reviews of the Stommel and Schott (1977) beta spiral method are provided by Olbers *et al.* (1985) and Bennett (1992). The basic equations for this application are the usual linearized beta (β)-plane equations for horizontal geostrophic flow (u, v) in a Boussinesq fluid

$$-\rho_o fv = -\partial p/\partial x \tag{4.5.8a}$$

$$\rho_o fu = -\partial p/\partial y \tag{4.5.8b}$$

the hydrostatic equation

$$0 = -\partial p/\partial z - \rho g \tag{4.5.9}$$

which relate pressure perturbations, $p(\mathbf{x}, t)$, to density fluctuations, $\rho(z, t)$, and the conservation of mass (or continuity) relation

$$\nabla\cdot \mathbf{u} + \partial w/\partial z = 0 \tag{4.5.10}$$

In these equations, f is the Coriolis parameter, u, v, and w are, respectively, the eastward (x), northward (y) and upward (z) components of current velocity, and $\rho = \rho(x, y, z)$ is the density perturbation about the mean density $\rho_o = \rho_o(z)$. Following Bennett (1992), we will reserve vector notation for horizontal fields and operators ($\mathbf{x} = (x, y)$, $\mathbf{u} = (u, v)$, etc.).

Using the above equations, we can derive the well-known "thermal wind" relation, whose vertically integrated velocity components are

$$u(\mathbf{x}, z) = u_o(\mathbf{x}) + (g/f\rho_o)\int_{z_o}^{z} \rho_y(x, \zeta)\, d\zeta \tag{4.5.11a}$$

$$v(\mathbf{x}, z) = v_o(\mathbf{x}) - (g/f\rho_o)\int_{z_o}^{z} \rho_x(x, \zeta)\, d\zeta \tag{4.5.11b}$$

where subscripts x, y refer to partial differentiation and $u_o(\mathbf{x})$, $v_o(\mathbf{x})$ are the velocity components at some reference depth. Equations (4.5.8–4.5.10) also give rise to the well-known Sverdrup interior vorticity balance

$$w_z = \beta v/f \tag{4.5.12}$$

where β is the northward (y) gradient of the Coriolis parameter, and $f = f(y) = f_o + \beta y$ in the beta-plane approximation.

These equations cannot be used alone to determine the full absolute velocity field ($\mathbf{u}$, w), even if the density field ρ were known. However, to resolve this indeterminacy, all we need is the velocity field at a particular depth where $\mathbf{u} = \mathbf{u}(\mathbf{x}, z_o)$ and $w = w(\mathbf{x}, z_o)$. Stommel and Schott (1977) demonstrated that these unknown reference values may be estimated by assuming the availability of measurements of some conservative tracer ϕ which satisfy the steady-state conservation law

$$\mathbf{u}\cdot\nabla\phi + w\phi_z = 0 \tag{4.5.13}$$

This tracer might be salinity (S) or potential temperature (θ), or some function of both

S and θ. Combining the vertical derivative of equation (4.5.13) with equations (4.5.11) and (4.5.12), yields

$$\left(\mathbf{u}\cdot\nabla + w\frac{\partial}{\partial z}\right)(f\phi_z) = (g/\rho_o)\mathcal{J} \tag{4.5.14}$$

where $\mathcal{J}$ is the Jacobian $\mathcal{J}(\rho, \phi) = \rho_x\phi_y - \rho_y\phi_x$. In equation (4.5.14), $f\phi_z$ represents the potential vorticity which would be conserved if density ρ were itself conserved. The tracer equation can be used again to eliminate the vertical velocity w

$$\mathbf{u}\cdot\mathbf{a} = (g/\rho_o)\mathcal{J}(\rho, \phi) \tag{4.5.15}$$

where the vector $\mathbf{a}$ is given by

$$\mathbf{a}(\mathbf{x}, z) = \nabla(f\phi_z) - \frac{\nabla\phi}{\phi_z} f\phi_{zz} \tag{4.5.16}$$

Using the integrated thermal wind equations (4.5.11) yields

$$\mathbf{u}_o\cdot\mathbf{a} = c \tag{4.5.17}$$

where $\mathbf{u}_o$ is the horizontal velocity at depth z_o and c is given by

$$c(\mathbf{x}, z) = -\mathbf{u}'\cdot\mathbf{a} + (g/\rho_o)\mathcal{J}(\rho, \phi) \tag{4.5.18}$$

In equation (4.5.18), the $\mathbf{u}'$ is that part of the horizontal velocity in the thermal wind relation that depends on the density field.

Since $\mathbf{a}$ and c depend on g, ρ, f, $\nabla\rho$, ∇f, ϕ_z and ϕ_{zz}, they can be determined using closely spaced hydrographic stations through measurements of $T(z)$ and $S(z)$. Thus, from (4.5.17), we can calculate $\mathbf{u}_o$ using the hydrographic data. Equation (4.5.17) holds at all levels so that two different levels can be used to specify u_o and v_o. We can then calculate the vertical velocity w from (4.5.14). The full velocity solution should be independent of the levels chosen for these computations. In reality, (4.5.17) is not an exact relation as it was derived from approximate dynamical laws and computed from data that contain measurement and sampling errors. As a consequence, our estimate of $\mathbf{u}_o$ from (4.5.17) should be done as a best fit to the data from the two levels chosen.

Suppose that N levels are chosen from the hydrographic data ($N \geq 2$). Let $c_n = x(\mathbf{x}, z_n)$ and $\mathbf{a}_n = \mathbf{a}(\mathbf{x}, z_n)$ for $1 \leq n \leq N$. The simple least-squares best fit minimizes

$$R^2 = \sum_{n=1}^{N} R_n^2 = \sum_{n=1}^{N} (c_n - \mathbf{u}_o\cdot\mathbf{a}_n)^2 \tag{4.5.19}$$

where R_n is the residual at level n and R is the root-mean-square (RMS) total error. R^2 is a minimum if $\mathbf{u}_o$ satisfies a simple linear system

$$\mathbf{M}\mathbf{u}_o = \mathbf{d} \tag{4.5.20}$$

where the 2 × 2 systematic, nonnegative matrix $\mathbf{M}$ depends on the components of $\mathbf{a}_n$, while $\mathbf{d}$ depends on $\mathbf{a}_n$ and c. If $\mathbf{a}$ or c varies with depth, equation (4.5.15) implies that the total velocity vector $\mathbf{u}$ must also depend on depth. For the β-spiral problem, we find that the large-scale ocean currents constitute a spiral with depth at each station. The β-spiral in Figure 4.5.2 is from the study by Stommel and Schott (1977) who used

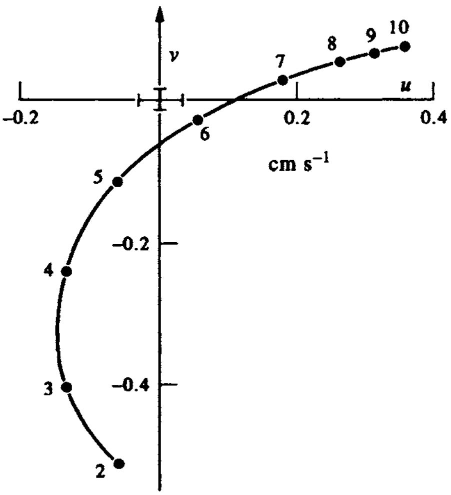

Figure 4.5.2. The β-spiral in horizontal velocity **u** *= (u(z), v(z)) at 28°N, 36°W, with depths in hundreds of meters. Error bars for the two components of velocity are given at the origin. (After Stommel and Schott, 1977.)*

hydrographic data from the North Atlantic to estimate $\mathbf{u}_o$ for a reference level of z_o = 1000 m depth. In this application they found, $u_o = 0.0034 \pm 0.00030$ m/s and $v_o = 0.0060 \pm 0.00013$ m/s at 28°N, 36°W.

The β-spiral problem includes two of the basic concepts common to inverse methods. First, we deal with an incomplete set of physical laws (4.5.8–4.5.10), or their rearrangement, as in the case of the thermal wind equations (4.5.11a, b) which includes the unknown reference velocity. Second, we often resort to the indirect measurement of an additional quantity which, in the case of the present example, is a conservative tracer. This application could have benefited from the inclusion of prior estimates of the errors in the dynamical equations and in the hydrographic data.

4.5.2.2 Wunsch's method

In a parallel development to the β-spiral technique, Wunsch (1977) used inverse methods to estimate reference velocities simultaneously around a closed path in the ocean (Bennett, 1992). As discussed by Davis (1978), both Wunsch's method and the β-spiral method are closely related. Both approaches assume the vertically integrated thermal wind equations (4.5.11) and both provide estimates for the reference velocity $\mathbf{u}_o$. In Wunsch's method, the thermal wind velocity, u$'$, is assumed to be zero at the reference level z_o, which in general may be a function of position $[z_o = z_o(x)]$. Wunsch chose the reference level to be the ocean bottom at $z_o(\mathbf{x}) = H(\mathbf{x})$, with $\mathbf{u}_o(\mathbf{x})$ defined to be the bottom velocity. He then divided the water column into a number of layers defined by temperature ranges. This is consistent with the general water mass structure of the North Atlantic as defined by Worthington (1976). These layers need not be uniform in depth at each hydrographic station. Together with the coastline of the U.S., the hydrographic stations formed a closed path in the western North Atlantic (Figure 4.5.3).

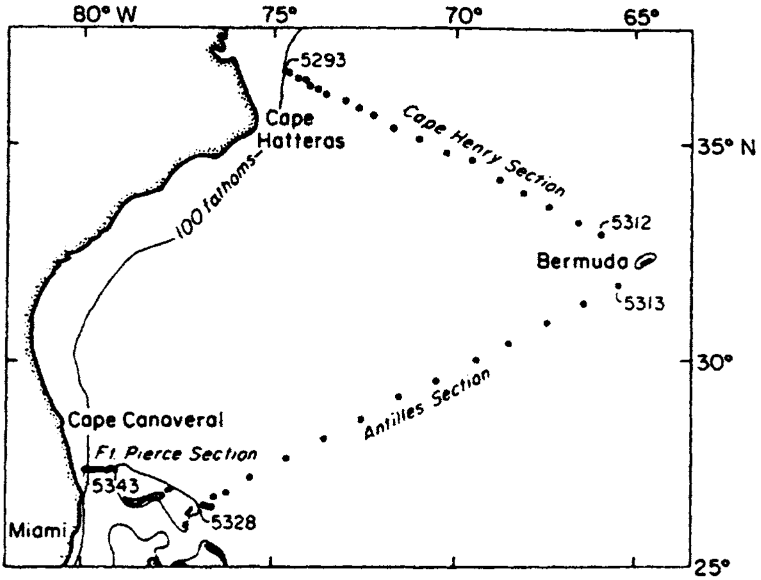

Figure 4.5.3. The locations of hydrographic stations in the North Atlantic used by Wunsch to obtain absolute current estimates using inverse theory. (After Wunsch, 1977.)

We now let v denote the outward component of velocity across the closed triangle formed by the lines of hydrographic stations in Figure 4.5.3. That is, $v = \mathbf{u} \cdot \mathbf{n}$ where $\mathbf{n}$ is the outward unit normal to the sections. We can further let $v' = \mathbf{u}' \cdot \mathbf{n}$ be the outward thermal wind velocity and $b = \mathbf{u}_o \cdot \mathbf{n}$ be the outward horizontal velocity at the seafloor. Let $v'_n(z)$ and b_n denote the thermal wind velocity estimate and unknown bottom velocity midway between the nth station pair, where $1 \leq n \leq N$, and let v'_{mn} denote the average value of v'_n in the mth layer of the water column, where $1 \leq m \leq M$. Wunsch chose the Mth layer to be the total water column, thus the Mth tracer is the total mass of the water column. The assumption of tracer conservation within each layer can be written as

$$\sum_{n=1}^{N} (v'_{mn} + b_n)\Delta z_{mn}\Delta x_n = 0, \; 1 \leq m \leq M \tag{4.5.21}$$

where Δz_{mn} is the thickness of the mth layer at the nth station pair, and Δz_{mn} is the separation distance between the nth station pair. This system of M equations for N unknowns b_n, $1 \leq n \leq N$, may be written in matrix notation as

$$\mathbf{Ab} = \mathbf{c} \tag{4.5.22}$$

where $\mathbf{A}$ is an $M \times N$ matrix and $\mathbf{c}$ is a column vector of length M with elements

$$A_{mn} = \Delta z_{mn}\Delta x_n \tag{4.5.23a}$$

$$c_m = -\sum_{n=1}^{N} \overline{v'_{mn}} A_{mn} \tag{4.5.23b}$$

Wunsch used $M = 5$ layers as defined by the ranges 12–17°C, 4–7°C, 2.5–4°C, and the entire water column (total mass). The hydrographic data were from $N = 43$ station pairs. For this problem, the matrix equation (4.5.23) represents five equations for 43 unknown velocities, so that the system is underdetermined and has many different solutions.

As reported by Bennett (1992), Wunsch (1977) somewhat arbitrarily selected the vector **b** with the shortest length. This was found by minimizing

$$t_1 = \mathbf{b}^T\mathbf{b} + 2\mathbf{I}^T(\mathbf{Ab} - \mathbf{c}) \tag{4.5.24}$$

where the superscript T denotes the transpose of the matrix and **I** is an unknown Lagrange multiplier consisting of a column vector of length M. It can be shown that t_1 is a minimum when

$$\mathbf{b} + \mathbf{A}^T\mathbf{I} = \mathbf{0} \tag{4.5.25}$$

which gives the minimum solution

$$\mathbf{b} = \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{c} \tag{4.5.26}$$

which satisfies (4.5.22). The symmetric matrix $\mathbf{AA}^T$ has dimensions $M \times M$ and is nonnegative (Bennett, 1992). However, $\mathbf{AA}^T$ may be singular. These singularities may be overcome by allowing errors in the hydrographic data and conservation laws; that is, by not seeking exact solutions of (4.5.22). We can instead write (4.5.22) in a quadratic form adding weights to each term. It can be shown that for positive weights, we are able to define an exact solution of the problem. This transfers the problem to the selection of these weights.

This cursory presentation of Wunsch's method for computing reference velocities demonstrates, once again, some of the basic elements of inverse methods: A system of incomplete physical laws and inexact measurements of related fields. It is necessary to admit errors into the equations and data values in order to stabilize the solution and to derive a unique solution. In his review, Davis (1978) concluded that both the underdetermined problem of Wunsch's method and the overdetermined problem of the β-spiral method are consequences of tacit assumptions made about noise levels and fundamental scales of motion. Davis suggested that a more orderly approach would be based on Gauss–Markov smoothing (Bennett, 1992) which should be an improvement, assuming explicit and quantitative estimates of the noise and its structure.

4.5.3 The IWEX internal wave problem

Another oceanographic example of the inverse method is found in Olbers *et al.* (1976) and Willebrand *et al.* (1977). Here, inverse theory is used to determine the three-dimensional internal wave spectrum from an array of moored current meters (Figure 4.5.4). In this example, the Fredholm equation (4.5.1) is written in matrix form and becomes

$$y_i = A_{ij}x_j; \quad 1 \le i \le N;\ 1 \le j \le K \tag{4.5.27}$$

where y_i are N observed velocity cross-spectra (the data), A_{ij} are the kernel functions (for matrix **A**) representing the physical relations from internal wave theory and x_j are the K internal wave parameters to be determined by the inverse method. The inverse

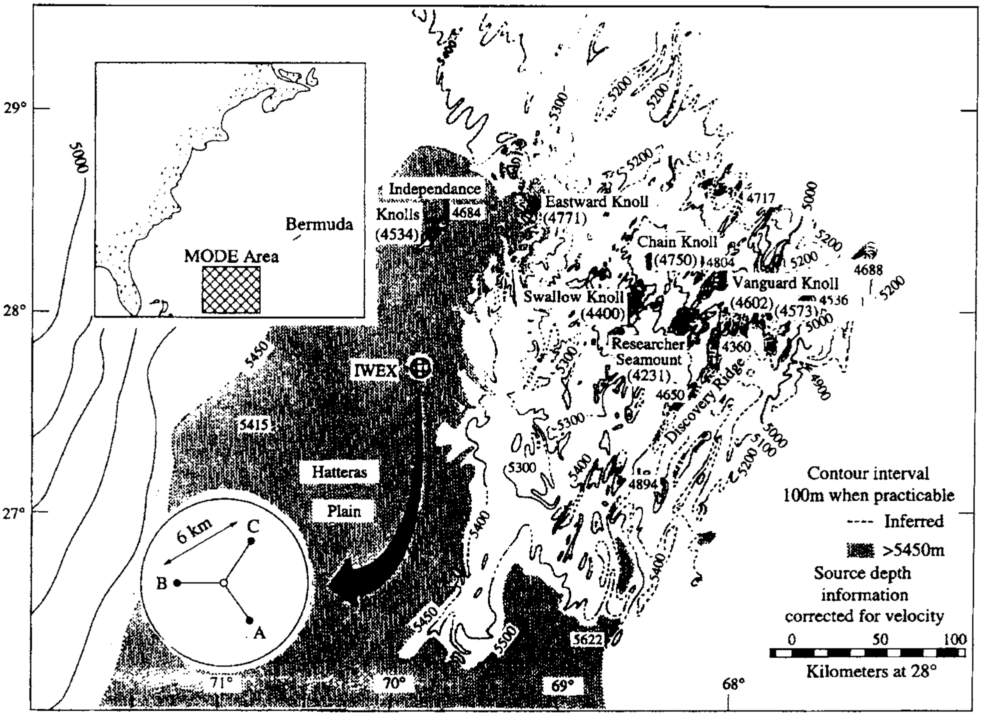

Figure 4.5.4. Location of the IWEX study area showing the positions of the three current meter moorings on the Hatteras Plain in the western North Atlantic. (From Briscoe, 1975.)

problem is to find the K parameters of the theoretical internal wave energy density cross-spectra using the N observed cross-spectra from the current meter array. We achieve this by using the least-squares method to minimize

$$\varepsilon^2(a) = [\hat{y} - y(a)]W[\hat{y} - y(a)]^* \tag{4.5.28}$$

where a represents a set of trial values used to find the minimum and the asterisk (*) denotes the complex conjugate. In equation (4.5.28), W is a weighting matrix used to scale the problem and to produce statistical independence (Jackson, 1972).

It is common to expand the kernel function matrix **A** into eigenvectors (Jackson, 1972). Thus, we write

$$\mathbf{A}V_j = \lambda_j u_j, \quad \mathbf{A}^T u_j = \lambda_i V_i \tag{4.5.29}$$

Following the singular value decomposition we conducted in the EOF analysis (Section 4.3.2), we can factor the matrix **A** as

$$\mathbf{A} = \mathbf{UBV}^T \tag{4.5.30}$$

where **U** is an $N \times P$ matrix whose columns are the eigenvectors u_i, $i = 1, \dots, P$; **V** is the $M \times P$ matrix whose columns are the eigenvectors v_i, $i = 1, \dots, P$, and **B** is the diagonal matrix of eigenvalues. After **U** and **V** are formed from the eigenvectors corresponding to the P nonzero eigenvalues of **A**, there remain $(N - P)$ eigenvectors U_j and $(K - P)$ eigenvectors V_j which correspond to zero eigenvalues. If we assemble these into columns of matrices, we have U_o (an $N \times (N - P)$ matrix) and V_o (a K

$\times(K - P)$ matrix). This is called *annihilator space* and reveals that our model is composed of both real model space (which corresponds to the data) and annihilator space which is linked to zeros in the data field. When we perform an inverse calculation, we usually recover a solution which lies in real model space. We must remember, however, that any function in space a can be added to the solution and still produce a solution that fits the data. With the kernel functions transformed into an orthogonal framework (expanded into eigenvectors) we construct the "smallest" or minimum energy model-solution.

When $P = N$, there is a solution to (4.5.28) and $P = M$ guarantees that a solution, if it exists, is unique. For $P < N$, the system is said to be overconstrained, while if $P < M$, the system is both overconstrained and underdetermined. In the latter case, an exact solution may not exist but there will be an infinite number of solutions satisfying the least squares criterion. This is the case for the present internal wave example, which is both overconstrained and underdetermined.

Returning to our internal wave problem, we find W in equation (4.5.28) using the least-squares method which produces the maximum likelihood estimator for a Gaussian distribution. This estimator is defined to be the inverse of the data covariance matrix. From the current meter array, 60 time series were divided into 25 overlapping segments. For each segment, cross-spectral estimates were computed for each of 600 equidistant frequencies. Averaging over segments and frequency bands to increase statistical significance, resulted in 3660 cross-spectra. The resultant 3660 × 3660 covariance matrix is difficult to invert. The diagonal of the weight matrix was selected to be

$$W = \text{diag}[1/\text{var}(y_i)] \tag{4.5.31}$$

which reproduces the main features of the maximum likelihood weight matrix (Olbers *et al.*, 1976). We note that, again for this problem, there are many more data points than parameters so that the system is overconstrained.

The least-squares solution procedure for this internal wave example is as follows:

(a) first find a parameter estimate $\hat{a}$ (the best guess);

(b) linearize at the value $a = \hat{a}$, such that

$$\hat{y}(a) = \hat{y}(\hat{a}) + \mathbf{D}(a - \hat{a}) + \ldots \tag{4.5.32}$$

where

$$\mathbf{D} = \left\{\delta\hat{y}_i/\delta a_j\right\}\Big|_{a=\hat{a}} \tag{4.5.33}$$

(c) improve the parameter estimate by using

$$a - \hat{a} = \mathbf{H}[\hat{y}(a) - \hat{y}(\hat{a})] \tag{4.5.34}$$

where the $N \times K$ matrix $\mathbf{H}$ is the generalized inverse of $\mathbf{D}$ derived from the linear terms of (4.5.32). If the matrix $\mathbf{D} < \mathbf{TWD}$ is nonsingular and well conditioned then

$$\mathbf{H} = (\mathbf{D}^T\mathbf{W}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{W} \tag{4.5.35}$$

and equation (4.5.11) becomes the least-squares solution of (4.5.29). Since $\mathbf{D}^T\mathbf{WD}$ is an $K \times K$ matrix, it can be easily inverted using standard diagonalization routines.

Having now arrived at a solution, $\mathbf{A} = \{A_{ij}\}$ of our problem in (4.5.27), we are left with two additional questions: (1) How well are the data reproduced by our solution? and (2) How accurately do we know our parameters $a_{\min}$? Since our data are subject to random errors, we can treat y as a statistical quantity and test the hypothesis that y and the model estimate $\hat{y}(a_{\min})$ are the same with a 95% probability (inverse estimate must be within the 95% confidence interval of our data point). Using the central limit theorem for our segment and frequency-averaged spectral values, we can approximate the 95% confidence interval on y as

$$\varepsilon^2_{95\%} = \overline{\delta y W \delta y}\left[1 + O(L^{-1})\right] = L \tag{4.5.36}$$

where $\delta y = y - \bar{y}$, and $O(\cdot)$ indicates the order of magnitude. Now if

$$\varepsilon^2(a_{\min}) \le \varepsilon^2_{95\%} \tag{4.5.37}$$

the model is a statistically consistent representation of the data. The consistency of the IWEX model is provided by the results in Figure 4.5.5, where we have plotted the measured, $\varepsilon^2(a)$, and expected, ε^2, values of the parameter ε^2. In this case, all values have been normalized so that magnitudes provide some indication of the percentage to which the observed and estimated (modeled) values of the data, y, coincide. For the most part, the measured values of ε^2 are scattered about the expected values of this parameter. Except at the M_2 tidal frequency and for frequencies greater than 1 cph, the hybrid IWEX model gives a consistent description of the IWEX data set to the 95% level.

Our second question regarding the accuracy of the parameter solution $a_{\min}$, can be answered by calculating the covariance matrix of the parameters. Using equation (4.5.30), we obtain the $K \times K$ covariance matrix of the parameters,

$$\overline{\delta a \delta a} = \mathbf{H}\overline{\delta y \delta y}\mathbf{H}^T \tag{4.5.38}$$

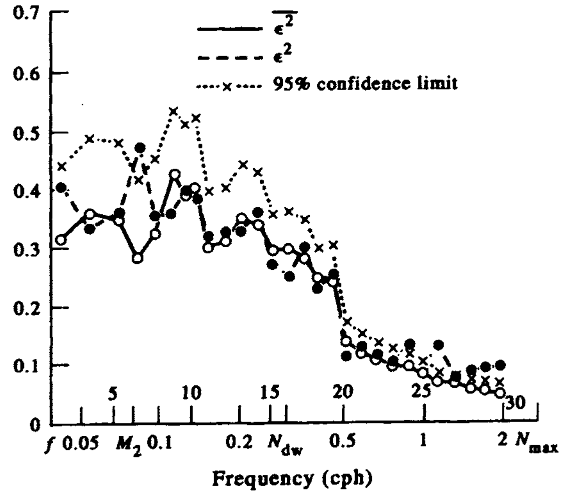

Figure 4.5.5. Consistency for the IWEX study. The error estimate ε^2 is the squared difference between the observed data and the modeled data obtained by inverse methods. Except for motions in the M_2 tidal band and at frequencies greater than about 1 cph, the results are within the 95% confidence level. N_{max} and N_{dw} are the maximum Nyquist frequency and the Nyquist frequency for the deep water, respectively (Briscoe, 1975).

from the data covariance matrix $\overline{\delta y \delta y}$. As usual, there is a reciprocal relation between the variance and the resolution of the parameters. Statistically uncorrelated parameters can be found by diagonalizing the matrix in (4.5.38).

4.4.4 Summary of inverse methods

In this section we have presented the basic concepts of the general inverse problem and have set up the solution system for two different applications in physical oceanography. Our treatment is by no means comprehensive and is intended to serve only as a guide to understanding the process of forming linear inverse solutions to fit observed oceanographic data.

The first example we treated is the computation of absolute geostrophic velocity by specifying an unknown reference velocity. Both the β-spiral (Stommel and Schott, 1977) and Wunsch's method are discussed. The dynamics are restricted to geostrophy and the conservation of mass. The second example was the specification of parameters in theoretical internal wave cross-spectra to reproduce the velocity cross-spectra of an array of moored current meters. The statistical nature of both the data and the model are considered and the accuracy of the results are expressed in probabilistic terms. Readers interested in further discussion of these and other related applications of inverse methods are referred to Bennett (1992). This book contains a complete review of inverse methods along with discussion of most of the popular applications of inverse techniques in physical oceanography. We also direct the interested reader to a recent paper by Egbert *et al.* (1994) in which a generalized inverse method is used to determine the four principal tidal constituents (M_2, S_2, K_1, O_1) for open ocean tides. The tides are constrained (in a least squares sense) by the hydrodynamic equations and by observational data. In the first example, solutions are obtained using inversion of the harmonic constants from a set of 80 open ocean tide gauges. The second example uses cross-over data from TOPEX/POSEIDON satellite altimetry. According to the authors, "The inverse solution yields tidal fields which are simultaneously smoother, and in better agreement with altimetric and ground truth data, than previously proposed tidal models."