2.3

A TECHNIQUE FOR VECTOR CORRELATION AND ITS APPLICATION TO MARINE SURFACE WINDS

Breaker, L. C. and W. H. Gemmill
NOAA/NWS/NMC
Washington, D. C. 20233

D. S. Crosby
NOAA/NESDIS and American University
Washington, D. C. 20233

1. INTRODUCTION

The problem of correlating vector quantities has been of interest to meteorologists for at least the past 75 years (e.g., Sverdrup, 1917; Durst 1957; Court, 1958; Breckling, 1989). However, it appears that a completely satisfactory definition for vector correlation has yet to emerge. Crosby et al. (1992) proposed a definition for vector correlation which arose outside the meteorological community, originating with Hooper (1959) and later expanded upon by Jupp and Mardia (1980). We apply the results of Crosby et al. to the problem of comparing marine surface winds for two different situations. In the first situation, the above definition for vector correlation is applied to marine surface winds from buoys at two locations in the NW Atlantic approximately 700km apart; in the second, we compare marine surface winds derived from NMC's Global Data Assimilation System with those acquired from NDBC buoys located primarily in U.S. open coastal waters and in the Gulf of Alaska. The data selected in the first case are time series, and as such, allow us to examine the time variation in vector correlation over the length of record. In the second case, the observed and analyzed data were simply grouped by month, permitting intermonthly and seasonal comparisons.

First, we present a brief review of the theory and a description of the properties associated with the definition of vector correlation originally given by Hooper. Then we apply the technique to marine surface winds in two different situations. We summarize our results and comment on the technique in the final section of the paper.

2. THEORETICAL BACKGROUND AND PROPERTIES

2.1 Theory

Given the two-component vectors $\vec{W}_1 = u_1 i + v_1 j$ and $\vec{W}_2 = u_2 i + v_2 j$ in Cartesian coordinates, we can express the covariance matrix for $\vec{W}_1$ and $\vec{W}_2$ as

$$\sum_{\vec{w}} = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix} \quad (1)$$

where the Σ_{ij} submatrices are given by

$$\sum_{11} = \begin{bmatrix} \sigma(u_1, u_1) & \sigma(u_1, v_1) \\ \sigma(v_1, u_1) & \sigma(v_1, v_1) \end{bmatrix} \quad (1a)$$

$$\sum_{12} = \begin{bmatrix} \sigma(u_1, u_2) & \sigma(u_1, v_2) \\ \sigma(v_1, u_2) & \sigma(v_1, v_2) \end{bmatrix} \quad (1b)$$

$$\sum_{21} = \begin{bmatrix} \sigma(u_2, u_1) & \sigma(u_2, v_1) \\ \sigma(v_2, u_1) & \sigma(v_2, v_1) \end{bmatrix} \quad (1c)$$

$$\sum_{22} = \begin{bmatrix} \sigma(u_2, u_2) & \sigma(u_2, v_2) \\ \sigma(v_2, u_2) & \sigma(v_2, v_2) \end{bmatrix} \quad (1d)$$

In Eq. (1), the σs are the variances or the covariances of the u and v components. The vector correlation between $\vec{W}_1$ and $\vec{W}_2$ is then defined as

$$\rho_v^2 = TR\left(\left(\sum_{11}\right)^{-1}\sum_{12}\left(\sum_{22}\right)^{-1}\sum_{21}\right) \quad (2)$$

where ρ_v^2 is the square of the vector correlation coefficient, ρ_v, and TR represents the trace of the products of the Σ_{ij} submatrices (Jupp and Mardia, 1980). Eq. (2) can be expanded in algebraic form to yield

$$\rho_v^2 = f/g, \quad (3)$$

where

$$f = \sigma(u_1,u_1)(\sigma(u_2,u_2)(\sigma(v_1,v_2))^2 + \sigma(v_2,v_2)(\sigma(v_1,u_2))^2) + \sigma(v_1,v_1)(\sigma(u_2,u_2)(\sigma(u_1,v_2))^2 + \sigma(v_2,v_2)(\sigma(u_1,u_2))^2) + 2(\sigma(u_1,v_1)\sigma(u_1,v_2)\sigma(v_1,u_2)\sigma(u_2,v_2)) + 2(\sigma(u_1,v_1)\sigma(u_1,u_2)\sigma(v_1,v_2)\sigma(u_2,v_2)) - 2(\sigma(u_1,u_1)\sigma(v_1,u_2)\sigma(v_1,v_2)\sigma(u_2,v_2)) - 2(\sigma(v_1,v_1)\sigma(u_1,u_2)\sigma(u_1,v_2)\sigma(u_2,v_2)) - 2(\sigma(u_2,u_2)\sigma(u_1,v_1)\sigma(u_1,v_2)\sigma(v_1,v_2)) - 2(\sigma(v_2,v_2)\sigma(u_1,v_1)\sigma(u_1,u_2)\sigma(v_1,u_2))$$

and

$$g = [\sigma(u_1,u_1)\sigma(v_1,v_1) - (\sigma(u_1,v_1))^2][\sigma(u_2,u_2)\sigma(v_2,v_2) - (\sigma(u_2,v_2))^2].$$

This version of Eq. (2) may be more convenient for computational purposes. Because W is two-dimensional, Eq. (2) yields values of ρ_v^2 which vary between 0.0 (no correlation) and 2.0 (perfect correlation).[1] For convenience, we have calculated, and quote, values of ρ_v^2 (vice $\sqrt{\rho_v^2}$) throughout the study.

2.2 Properties

The definition for vector correlation given by Eq. (2) has the following properties, where we have replaced the population parameter, ρ_v^2, by the corresponding sample parameter, r_v^2.

(i) r_v^2 is symmetric in the sense that

$$r^2_{\vec{w}_1|\vec{w}_2} = r^2_{\vec{w}_2|\vec{w}_1}$$

(ii) r_v^2 is independent of coordinate transformations.

(iii) r_v^2 is equal to 2.0, for two-dimensional vectors, if $\vec{W}_1$ and $\vec{W}_2$ are completely dependent.

(iv) If $\vec{W}_1$ and $\vec{W}_2$ are independent, then r_v^2 will approach 0.0 as the sample size increases without limit. For $\vec{W}_1$ and $\vec{W}_2$ independent, nr_v^2 is asymptotically distributed as chi-square, where n is the sample size for which r_v^2 is calculated. For the two-dimensional case, the chi-square distribution has four degrees-of-freedom.

(v) For the one-dimensional (scalar) case, r_v^2 simplifies to the square of the Pearson product-moment correlation coefficient.

(vi) The vector correlation (squared), r_v^2, is equal to the sum of the squared first (r_1^2) and second (r_2^2) canonical correlations (Crosby et al., 1992), $r_v^2 = r_1^2 + r_2^2$.

3. APPLICATIONS

3.1 Marine Surface Winds at Two Locations in the NW Atlantic

In the first situation, we calculate vector correlations between surface winds at two locations in the NW Atlantic. The wind observations were acquired by NDBC environmental data buoys located at 40.5°N, 69.5°W (buoy number 44008) and at 34.9°N, 72.9°W (buoy number 41001). These buoys, whose locations are shown in Figure 1, are approximately 700km apart, close enough so that synoptic-scale disturbances that typically pass through the region will, in most cases, influence the winds at both sites. An expected winter storm track for this region has been included (Klein, 1957). As winter low-pressure systems leave the east coast of the U.S., they often deepen over the Gulf Stream and expand as they propagate to the NE. Thus, the winds at both buoys are expected to be strongly influenced by the passage of these low pressure systems which pass through the area during the winter months. The observations, taken hourly, extend from 1 December 1987 to 4 February 1988, a period of 65 days. The stick diagram shown in the upper two panels of Figure 2 depict the time series of wind vectors at each location.

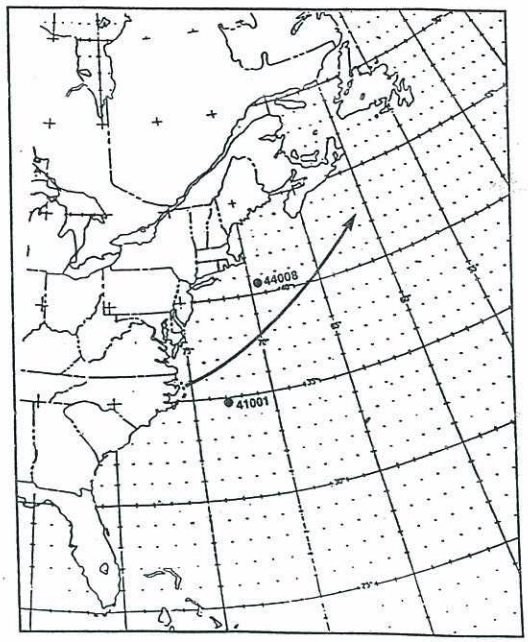

FIGURE 1. Locations of the two NDBC environmental data buoys from which time-series surface winds were extracted. Period covers 1 December, 1987 to 4 February, 1988. A typical winter storm track has been included (Klein, 1957).

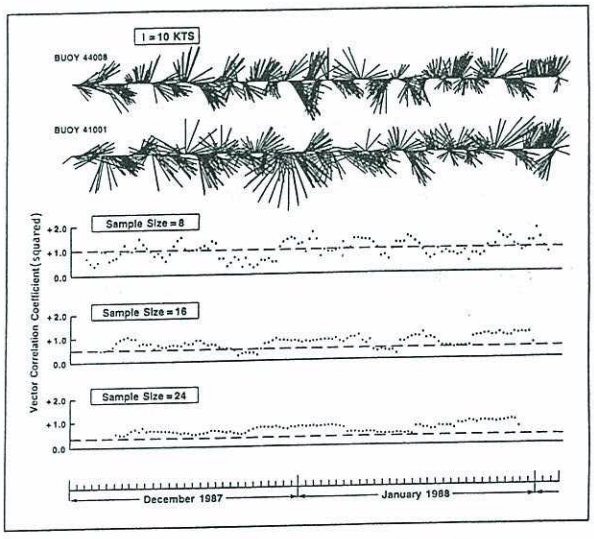

Figure 2. Wind vector sequences for NDBC buoys 44008 (top panel) and 41001 (next-to-top-panel), and the corresponding vector correlations for sample sizes of 8, 16, and 24 (lower panels). The horizontal dashed lines in the lower three panels indicate the 5% level of significance.

[1]This definition, of course, can be scaled to vary between 0.0 and 1.0, if so desired.

Autocorrelation analyses were initially conducted to estimate the time scales of persistence. Autocorrelation analysis of the u and v wind components indicated correlation time scales that were rather consistently of the order of half a day; consequently the original data have been subsampled every 12th point to produce series with observations which are approximately independent.

Vector correlations have been calculated for four sample sizes, 8, 16, 24 and the entire series (i.e., 130) corresponding to periods of 96, 192, 288 and 1560 hours, respectively. A sliding sample window was employed which was stepped forward one data interval at a time for each sample size. The results are shown in Figure 2 (lower three panels). The upper 95th percentile of the distribution has been included to determine whether or not the individual values of r_v^2 are statistically significant at the 5% level, assuming that the points within the series are independent (Crosby et al., 1992).

Our choices of sample size are based primarily on the synoptic time scales of variation in the surface wind fields. The winds shown in Fig. 2 indicate time scales of variation (i.e., "event" time scales) on the order of 2-4 days. Sample sizes of 8 (4 days), 16 (8 days) and 24 (12 days) clearly encompass these time scales. It is important to recognize that the sample size must be sufficient to include significant variation in the vector sequences being correlated. For sample sizes that are too small in this respect, spurious correlations may arise.

The results for a sample size of 8 indicate that significant variation in r_v^2 itself occurs over the length of the series. The sample parameter r_v^2 exceeds the 95th percentile slightly less than 50% of the time. Relatively high values (r_v^2 = 1.5 or greater) tend to occur where major changes in surface wind (particularly noticeable in wind direction) are similar at both locations. Relatively low values of r_v^2 (less than about 0.4) tend to occur throughout the record, but we find no obvious explanation for their occurrence.

As sample size increases from 8 to 16 and from 16 to 24, the correlations tend to be statistically significant in most cases but the changes in r_v^2 tend to reflect to a lesser extent the major 2-4 day event-scale changes in surface wind. It becomes increasingly difficult to relate the values of r_v^2 to individual events in the wind field. In the limit, when N equals 130, we obtain a single value for r_v^2 that represents the correlation between the surface wind fields at the two locations over the entire record. In this case r_v^2 is equal to 0.54, a value which is statistically significant at the 5% level.

3.2 Comparing Analyzed and Observed Marine Surface Winds

In the second situation, we compare analyzed and predicted marine surface winds with observed winds for various buoy locations around the coastal U.S. and in the Gulf of Alaska. For this comparison, winds were acquired at 20 locations. The analyzed and predicted winds were derived from NMC's Global Data Assimilation System (GDAS). The lowest level winds from GDAS are located at the mid-point (~45m) of the lowest layer in the model. These winds are then adjusted to a height of 10m by assuming a neutrally stable, constant flux, layer using the well-known log-profile relation (e.g., Monin and Obukhov, 1954). These winds are produced on a 2.5° (latitude) x 2.5° (longitude) grid for forecast periods of 00(analyzed values), 24, 48 and 72 hours. The period during which these comparisons were made runs from 12/89 through 12/90, a period of 13 months. As indicated earlier, these data have simply been grouped by month for each of the 13 months. First, the appropriate analyzed values (i.e., the u and v components) are obtained by bilinear interpolation to the various buoy locations. Then the u and v components from each buoy and the corresponding interpolated values from the analysis taken over all buoys enter into the calculation of a single vector correlation for the entire month. Since there are many reports from each buoy we note that the total number of observations that enter into the calculation for a given month greatly exceeds the total number of buoys (199-531 versus 20). To further interpret our results, we have also included conventional scalar correlations[2] for the wind speeds to help distinguish between the effects of speed and direction. Also, to help identify any possible seasonal trends in the 13-month sequence of vector correlations which we present, we have calculated confidence limits for these vector correlations using the bootstrap method, an empirical approach for estimating the mean square error for any statistic.

The results of these calculations are presented in Fig. 3. Vector and speed correlations are both shown with, and without, one particular buoy (buoy number 46003, located in the Gulf of Alaska). At first unknown to us, the wind directions from this buoy were erroneous from 2/21/90 to 4/13/90 (U.S. Dept. of Commerce, 1990). Our initial calculations indicated a major and unexpected decrease in vector correlation for 3/90 (and to lesser extents for 2/90 and 4/90). A detailed examination of the data during this period indicated that incorrect wind directions from this buoy had been included in the data set for that month, an observation that was later confirmed according to the above reference. As a result, we have recomputed the vector correlations for the entire 13-month period excluding the one offending buoy. The recomputed results reveal the sensitivity of this calculation to erroneous data. From the

[2]The one-dimensional Pearson product-moment correlation coefficient.

speed correlation during the same period, it is clear that the decrease in vector correlation is primarily due to problems in wind direction for this one buoy. Overall, this comparison, with and without buoy 46003, also shows that during the remainder of the period (excluding February and April 1990), the vector correlations are generally robust in the sense that when a few reports are removed from any of the remaining monthly groups, similar vector correlations are obtained.

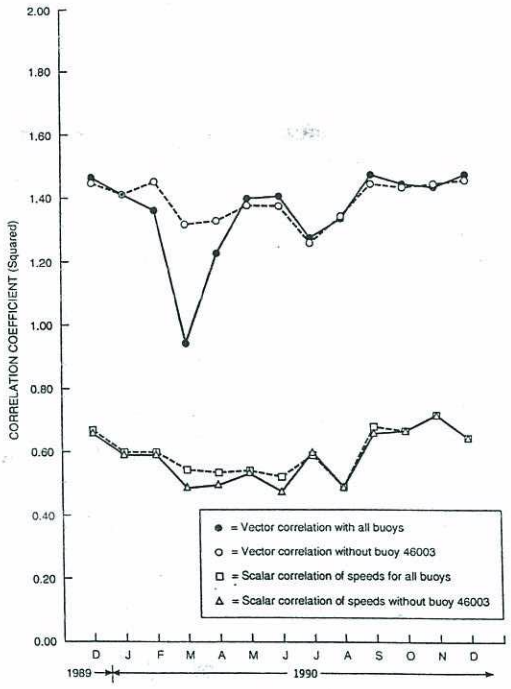

FIGURE 3. Vector and speed correlations between analyzed and observed winds by month for all buoys (dashed lines) and without buoy number 46003 (solid lines).

Fig. 4 shows the monthly vector correlations (without 46003) with the addition of confidence limits. Confidence limits were included to determine whether or not the variations in vector correlation from one month to the next and on a seasonal basis were significant. Since no theoretical basis exists for calculating these confidence limits, we used an empirical approach, referred to as the bootstrap technique, to estimate these limits (e.g., Yang and Robinson, 1986). In particular, to employ the bootstrap technique, we take the u and v components from the analysis and from the buoys for a given month and resample each of the component series to produce new series where each value in the new series has an equal probability of being selected from the original series. We perform this procedure repeatedly, in our case 200 times, to generate a distribution of simulated vector correlations from which we obtain the 95% confidence interval by selecting the 2.5 and the 97.5 percentiles.

This technique assumes that the original observations are independent, which in our case is not strictly true, since the observations from closely spaced buoys may be spatially correlated, and there is most likely correlation over time for observations from the same buoy. Serial correlation in the data not withstanding, we have estimated the upper and lower confidence limits for the sequence of monthly vector correlations. The confidence limits associated with these vector correlations indicate ranges of uncertainty which overlap significantly from one month to the next and on seasonal time scales as well, implying that major seasonal variations or trends in vector correlation do not exist in this particular sequence (for example, a value of vector correlation equal to 1.4 falls within the ranges of uncertainty for all 13 months).

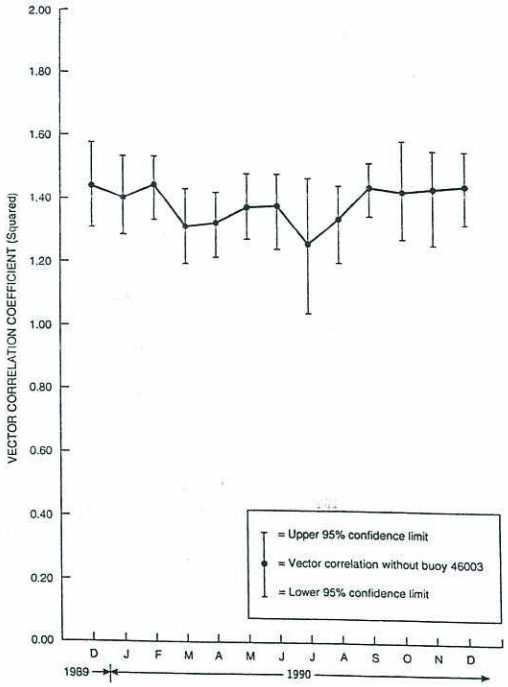

FIGURE 4. Vector correlations by month for all buoys except buoy number 46003. Upper and lower 95% confidence limits have been included.

In Fig. 5 we examine the relationship between vector correlation and forecast period. All of the monthly vector correlations are plotted versus forecast period from 00 hours out to 72 hours. Straight-line segments connect the mean vector correlations for each period and reveal that the correlations decrease with increasing forecast period, a result which was anticipated. Also, the rate of decrease in vector correlation increases beyond 24 hours. Finally, we note that the spread of vector correlations increases significantly as the forecast period increases.

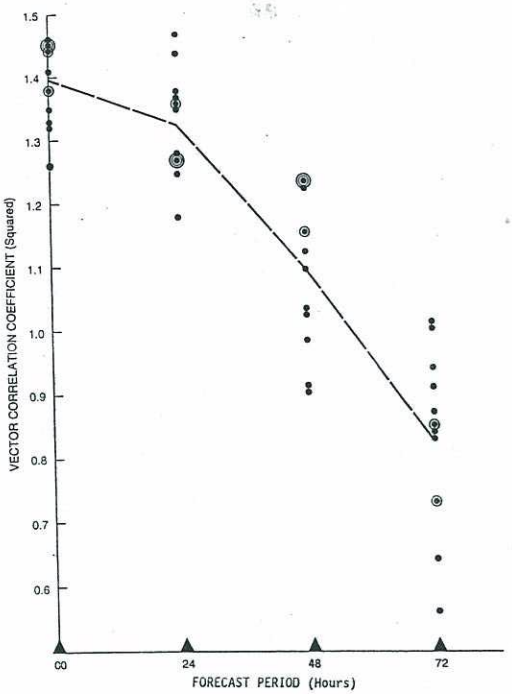

Figure 5. Vector correlations for all months (without buoy number 46003) versus forecast period. A dashed line connects the mean values of the vector correlations for each forecast period. Circles indicate two (or more) data points at the same location.

4. SUMMARY AND CONCLUSIONS

In the first case, with respect to the surface wind data at two locations in the NW Atlantic, it is clear that care must be exercised in selecting the "proper" sample size for calculating vector correlations for time series data. At one extreme, choosing a sample size which is too small may lead to spurious correlations that will not be amendable to interpretation. At the other extreme, when vector correlations for the entire series are calculated, a single value is obtained which will be meaningful, but the opportunity to examine time variations within the series will be lost. In cases where the sample sizes are small enough to reveal correlations related to individual events within the series, it may be possible to interpret ρ_v^2 in terms of these events. We have not attempted to do so here because these vector correlations may well depend on information that we did not have access to.

In the second case, we compared analyzed and observed marine surface winds using the present definition of vector correlation to improve quality control procedures. The definition of vector correlation used here has provided a sensitive indicator of the relationship between analyzed and observed winds. This correlation coefficient was also useful in detecting erroneous data. To determine whether or not intermonthly variations in vector correlation existed, we adopted an empirical statistical technique called bootstrapping. Using this technique, we estimated confidence limits for each of the monthly vector correlations, which in turn allowed us to determine whether or not the monthly and seasonal changes in vector correlation were significant. These results indicated that the intermonthly changes in vector correlation were most likely not significant.

In meteorology, vectors are often compared by correlating the orthogonal scalar components separately. Thus, the need for a correlation technique that compares the vectors per se can be questioned. We note, however, that correlating the scalar components separately produces values which are not unique since the results depend upon the coordinate system one chooses to adopt for the scalar decomposition. For example, if one correlates the scalar components using a spherical, earth-oriented coordinate system, one will generally obtain different results than if one uses a natural coordinate system. The method presented here is independent of the choice of coordinate system that is used to define the vectors. However, separate one-dimensional correlations of the scalar components may be helpful in interpreting the results, as was done for wind speed in the second case here.

Our primary purpose has been to present the definition of vector correlation originally proposed by Hooper with application to comparing and evaluating marine surface winds. There are still many questions about its application to practical problems. For example, the distribution of this statistic is known for large samples when the correlation is zero and the sample points are independent. However, little is known about its distribution when the sample points are not independent, a situation often encountered in time series data. Consequently, considerably more effort should be devoted to the application of this technique to the practical problems that frequently arise in comparing vector quantities.

5. REFERENCES

Breckling, J., 1989: The Analysis of Directional Time Series: Applications to Wind Speed and Direction, Springer-Verlag, Berlin.

Court, A., 1958: Wind Correlation and Regression, AFCRC TN-58-230, Contract AF19 (604)-2060, Cooperative Research Foundation, San Francisco, California, 19 February 1958, (AD 152460), 16pp.

Crosby, D.S., L.C. Breaker and W.H. Gemmill, 1992: A New Definition for Vector Correlation in Geophysics: Theory and Application, Submitted to the Journal of Atmospheric and Oceanic Technology.

Durst, B.A., 1957: A Statistical Study of the Variation of Wind with Height, Prof. Notes No. 121, Meteorological Office, London, 10pp.

Hooper, J.W., 1959: Simultaneous Equations and Canonical Correlation Theory, Econometrica, 27, 245-256.

Jupp, P.E. and K.V. Mardia, 1980: A General Correlation Coefficient for Directional Data and Related Regression Problems, Biometrika, 67, 163-173.

Klein, W.H., 1957: Principle Tracks and Mean Frequencies of Cyclones and Anticyclones in the Northern Hemisphere, U.S. Weather Bureau Research Paper No. 40, 60pp.

Monin, A.S. and A.M. Obukhov, 1954: Basic Regularity in Turbulent Mixing in the Surface Layer of the Atmosphere, Trudy Geophys. Inst. ANSSSR, No. 24.

Sverdrup, H.U., 1917: Under die Korrelation zwischen Vektoren mit Anwendungen auf Meteorologische Aufgaben. Met. Zeit., 285-291.

U.S. Department of Commerce, 1990: NDBC Data Platform Status Report, April 26, 1990 - May 3, 1990, U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, Stennis Space Center, MS.

Yang, M.C.K. and D.H. Robinson, 1986: Understanding and Learning Statistics by Computer, Series in Computer Science, Vol. 4, World Scientific, Singapore.

(Ocean Products Center Contribution No. 58)