NOAA Technical Memorandum ERL PMEL-38

DATA INTERCOMPARISON THEORY

I. MINIMAL SPANNING TREE TESTS FOR LOCATION AND SCALE DIFFERENCES

Rudolph W. Preisendorfer
Curtis D. Mobley

Pacific Marine Environmental Laboratory
Seattle, Washington
December 1982

UNITED STATES
DEPARTMENT OF COMMERCE
Malcolm Baldrige,
Secretary

NATIONAL OCEANIC AND
ATMOSPHERIC ADMINISTRATION
John V. Byrne,
Administrator

Environmental Research
Laboratories
George H. Ludwig
Director

NOTICE

Mention of a commercial company or product does not constitute an endorsement by NOAA Environmental Research Laboratories. Use for publicity or advertising purposes of information from this publication concerning proprietary products or the tests of such products is not authorized.

Contribution No. 600 from NOAA's Pacific Marine Environmental Laboratory

TABLE OF CONTENTS

Abstract

When intercomparing two data sets, each of n samples of some field at p points in space, the question often arises about the relative sizes of their averages over time and about their relative variances. In this note we consider two geometric ways of answering this question. The basic geometric concept is that of a minimal spanning tree (MST) made from the union of the data sets when they are considered as n-point swarms in euclidean p-space E_p. The MST is the network of straight lines in E_p that connects the points of the pooled swarms with the least possible total length of its segments. The test of relative location of data sets based on the MST uses a generalized notion of *run* (which measures how much the points of the two sets intermingle in their MST) while the scale test for variance is based on the simple intuitive idea that the set with greater variance will have the branches of its part of the tree spread beyond those of the other. Power tests were run for the MST location and scale tests and it was found that the MST scale test is relatively powerful and useful. An application of the MST scale test was made to the problem of defining natural seasons over the U.S. mainland using a 46-year temperature record. The result is a novel partition of the 12 months of the year into new seasons based on months with comparable temperature variances.

Data Intercomparison Theory

I. Minimal Spanning Tree Tests for Location and Scale Differences

Rudolph W. Preisendorfer

Curtis D. Mobley

1. Introduction

A. The intercomparison of data sets in meteorology and oceanography in recent years has become an important activity of climate research on several different levels. For example, there may arise the question of how much a physical field (sea level air pressure, sea surface temperature) has changed from one epoch (month, year, decade) to another. The change could be in the sense of, say, average value, of variance, of spatial pattern, or of temporal evolution of the physical field. We shall call this the *data-data* intercomparison problem. Another example arises when a general circulation model (GCM) for atmosphere/ocean interactions is attempting to simulate an observed data field over some space-time domain. The goodness of fit of the model field to the data field is again measurable in various modes such as mean, variance, or space/time evolution. This we shall refer to as the *model-data* intercomparison problem. Finally, in the development of a GCM, it may be of interest to make internal-parameter changes. These changes, along with initial and boundary condition changes, give rise to model-produced sets which are to be intercompared for dissimilarities in the above-listed attributes, and we thereby have the third main problem--that of *model-model* intercomparisons.

B. When the data sets are spatially and temporally extensive, the intercomparisons must be done mechanically, and objectively, by using appropriately

chosen measures of dissimilarity (distance functions). Moreover, these distance functions are applied in a setting which is usually of a random nature, and so the question arises on the statistical significance of observed dissimilarities among the various attributes of the compared data sets. The resolution of the latter question requires the construction of an appropriate reference distribution for the particular statistic (the distance function) being applied to the data sets. *The purpose of the present series of notes is to define and study some data intercomparison procedures, based on new statistics and their associated reference distributions, which will be applicable to a wide variety of data-data, model-data, and model-model intercomparison problems.*

C. In the present note we shall examine the data intercomparison problem using the notion of a *minimal spanning tree* (MST), a graphical device that facilitates objective decisions as to the relative locations and relative sizes of swarms of points in p-dimensional space. Here p is the number of time series in a data set, and the number n of points in the swarm is the common size of the sample of the time series. Normally, the relative locations of two swarms are specified by their centroids. By constructing a certain network of lines connecting the points of the two swarms in an economical way (the task of the MST) it is possible to give a relatively novel quantitative measure of the intermingling of the two point sets (and hence obtain a measure of their relative locations). Further, instead of looking at the standard deviations of the swarms (the usual measure of radial size) we determine the relative sizes of the swarms by counting the outermost branches of the tree belonging to each swarm, and seeing which of the two sets of branches "sticks out" of the combined swarm the more.

D. We shall compare the MST-derived measures of location and scale with some other measures and obtain a preliminary impression of the relative power of the new methods. It will turn out that the MST procedure provides a relatively powerful scale test, while the location test is relatively weak (although it is competitive with the classical T^2 test for location). This good showing of the MST scale test encouraged us to apply it to an interesting practical problem, that of determining the months of the year that group together into natural seasons. Of course this grouping together depends somewhat on the choice of the physical field (temperature, precipitation, e.g.) and geographic location (North America, Equatorial Pacific, e.g.). We shall specifically consider the problem of natural seasons as defined by commonly-shared monthly temperature variances over the U.S. mainland as provided by temperature records at 32 U.S. cities over a 46 year period: 1931-1976. We shall determine the natural temperature-iso-variance seasons two ways: via the MST scale test and by variance calculations of the classical form; and the results will be compared.

E. Acknowledgments

Dr. Tim P. Barnett, of the Climate Research Group, Scripps Institution of Oceanography, La Jolla, California, provided the original interest in the general data intercomparison problem, and indications for the need of solutions to the problem in climate diagnosis and prediction studies. He also supplied both the inspiration and the U.S. temperature records for the MST scale test applications to the natural-season problem described in §5 of this study. Ryan Whitney of PMEL typed the manuscript and Gini May of PMEL drew the figures.

2. MST Test for Location

The test for location using the minimal spanning tree (MST) of a pair of data sets is readily understood by first studying its one-dimensional version. In Fig. 2.1(a), we have a linear array of circles and crosses denoting points of a model and data set (say). A natural measure of the degree of separation (or intermingling) of the sets is obtained by counting the number of runs of circles and crosses. A *run* of data points is an unbroken sequence belonging to a given data set. The runs can be counted by placing dotted lines (or *cuts*) between every two data points from *different* sets. The total number of runs will then be the number of cuts plus one. In Fig. 2.1(a) there are four cuts and hence five runs. It is intuitively clear that the more closely intermingled the two sets are, the greater will be the number of runs. For the six data and six model points we can have as many as 11 cuts and hence 12 runs. There can be as few as one cut and hence two runs when the sets are totally separated.

When we have data sets in 2-space, as sketched in Fig. 2.1(b), to determine the number of runs, we build a minimal spanning tree of the union of the two data sets. This will be the set of straight lines connecting the circles and crosses in such a way that the total lengths of these line segments is a minimum. (The construction details of an MST in a general p-space are given in Appendix A.) The generalization of the *run* idea to two dimensions is shown in Figs. 2.1(b), (c). A cut is made on a line segment whenever that segment joins points from the two different sets (the model or the data set). The number of runs is one more than the number of cuts. Notice how in Fig. 2.1(b), (c) the number of runs diminishes from 10 to 7 as one goes from a closely intermingled pair of sets to a more separated pair of sets.

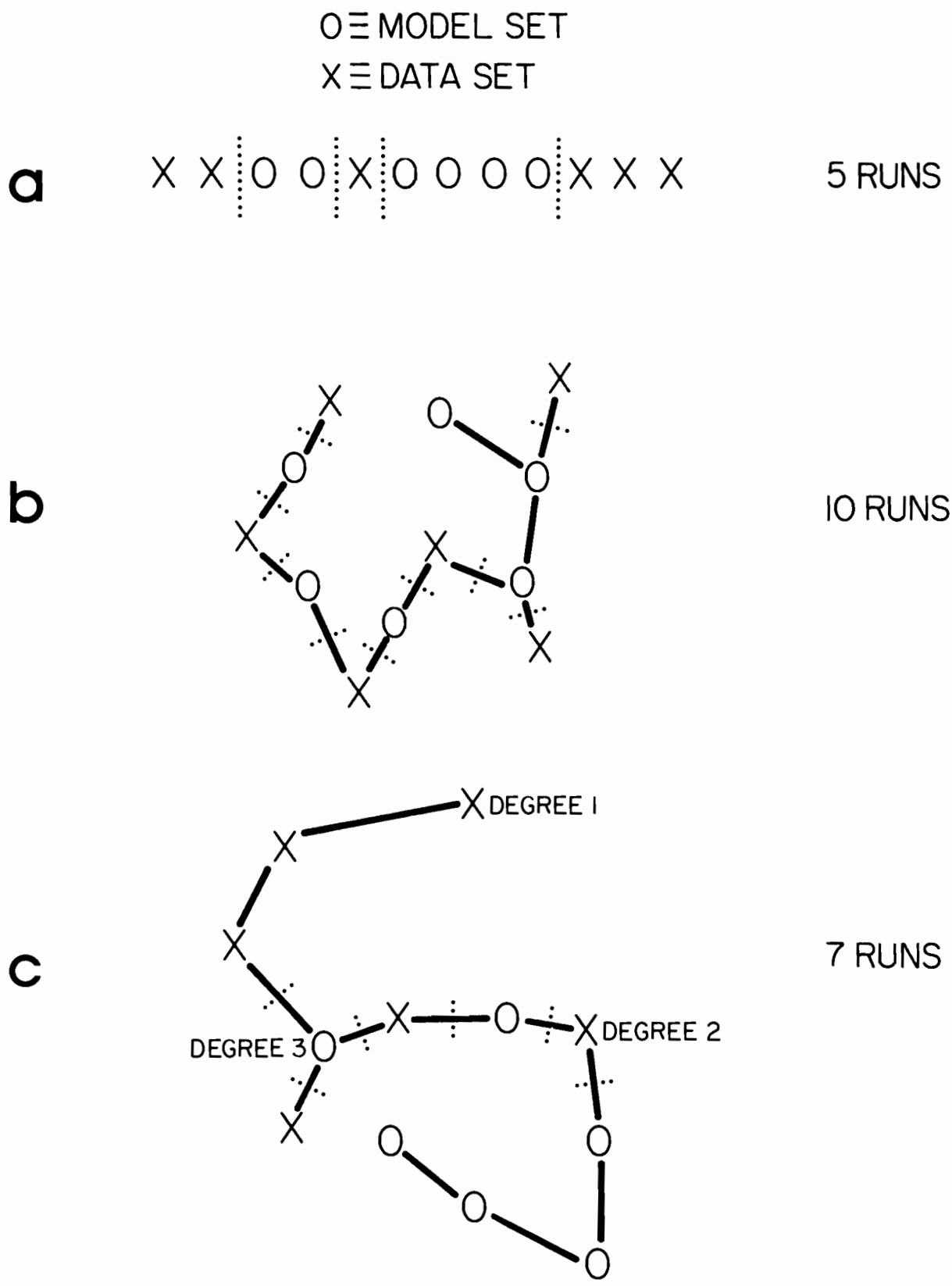

O ≡ MODEL SET
X ≡ DATA SET
a
X X O O X O O O O X X X
5 RUNS
b
10 RUNS
c
DEGREE 1
DEGREE 2
DEGREE 3
7 RUNS

Fig. 2.1

The preceding observations lead one to formulate the hypothesis H_o, namely

> H_o: The set $\underline{M}$ of model points and the set $\underline{D}$ of data points are randomly drawn from the same population.

Under this hypothesis we can build a useful statistical test of location using the *run statistic*. We would reject H_o if the number of runs for the MST of the union $\underline{D} \cup \underline{M}$ of $\underline{D}$ and $\underline{M}$ is significantly small. The decision of whether or not the number of runs is "significantly small" is based on the procedure in §2 of Appendix A. The procedure is based in turn on the observation that, under H_o, one can randomly interchange the labels of the $\underline{D}$ and $\underline{M}$ sets by means of arbitrary permutations. For each permutation we have a new $\underline{D}$ and $\underline{M}$ set and hence a new run number. By generating many such permutation-induced runs, we can build up a reference distribution for runs. If the runs-count for the originally given $\underline{M}$,$\underline{D}$ pair then falls in the left 5% tail (say) of this reference distribution we would reject H_o, and declare with confidence 95% that the $\underline{D}$ and $\underline{M}$ sets do have distinct locations.

3. MST Test for Scale

Testing for scale (i.e., spread or variance) using an MST of pooled model and data points (say) rests on the following observations. Let us define the *degree* of a point in an MST as the number of line segments in the MST leading to it from other points in the MST. Examples are shown in Fig. 2.1(c). Note that a data swarm $\underline{D}$ with relatively great variance compared to a model swarm $\underline{M}$ with relatively small variance will tend to have more degree-1 points than the *smaller-radius* set, providing they are relatively closely located or intermingled.

Under hypothesis H_o, it follows that the number of model points (or data points) of degree 1 obeys a hypergeometric distribution. Thus if n_D, n_M are

the number of points in $\underline{D}$ and $\underline{M}$, respectively, then the probability p(x) that x *model points* (say) will have degree 1 is

$$p(x) = \begin{bmatrix} n_M \\ x \end{bmatrix} \begin{bmatrix} n_D \\ n_1 - x \end{bmatrix} \begin{bmatrix} n \\ n_1 \end{bmatrix}^{-1}. \tag{3.1}$$

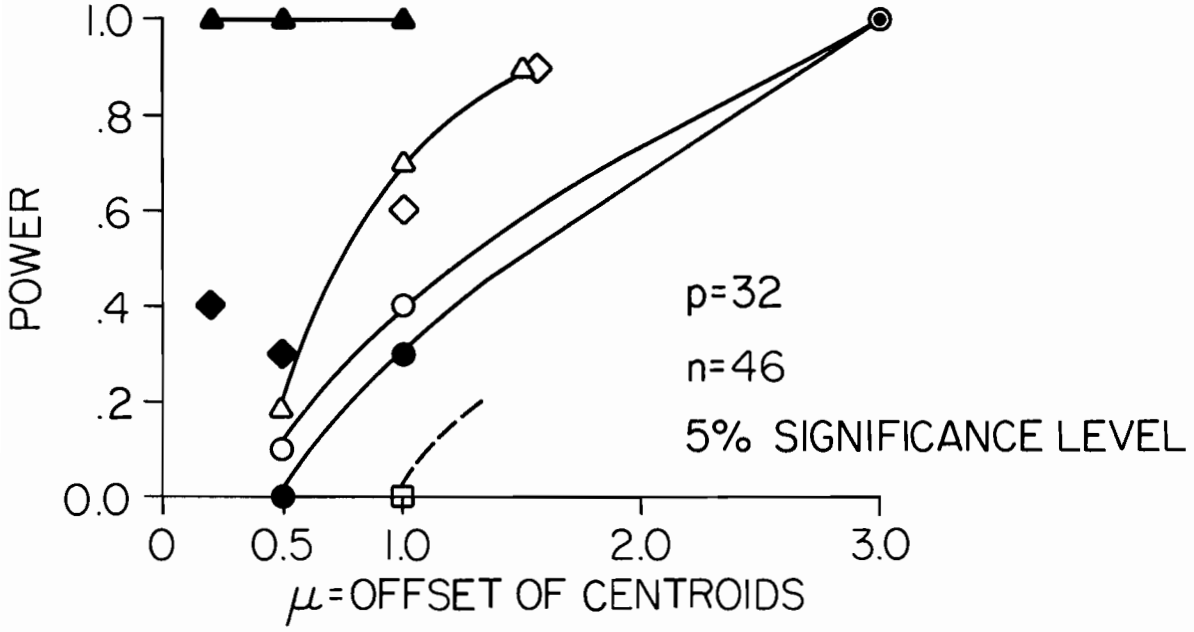
where $n = n_D + n_M$, and n_1 is the number of degree 1 points in the entire MST. We can, for given n_D, n_M and n_1, find the upper and lower critical values of x, and thereby decide on the relative scales of $\underline{D}$ and $\underline{M}$.

4. Power Curves for MST Location and Scale Tests

A. Power of the Location Test

The power of the MST location test was examined by means of suitably constructed gaussian populations in 32-dimensional space. The dimension of the space was suggested by the applications to be described in the following section. Thus we encounter p = 32 temperature time series of n = 46 samples each. To see the power of the MST location test under these conditions, we constructed two gaussian populations of the form $N_p(\underline{\mu}_j, \sigma_j^2 \underline{I}_p)$, j = 1,2, where p = 32 and where $\|\underline{\mu}_1\| = \underline{0}$, $\|\underline{\mu}_2\| = \mu$ and where μ was varied over the range $0 \leqq \mu \leqq 3.0$. Moreover, we set $\sigma_1 = \sigma_2 = 1.0$. Hence in effect we were sampling from two spherical gaussian swarms of unit variance in E_{32}, and whose centroids were an adjustable μ units apart. As we sampled from these various populations we determined runs-distributions under H_o for each μ, and noted when the runs were declared significantly small (a rejection of H_o). Whenever they were, we marked up a success for the MST location test. If the test is a good one, it would relatively early reject H_o, i.e., wake up to the fact that the swarm centers are a certain distance μ apart. The smaller the μ when this happens, the better the test. In Fig. 4.1 we have the results of the present MST

POWER OF MST LOCATION TEST AND OTHER TESTS



T^2 TEST ($\sigma_1=\sigma_2=1.0$)

MINIMAL SPANNING TREE ($\sigma_1=\sigma_2=1.0$)

MINIMAL SPANNING TREE ($\sigma_1=1.0, \sigma_2=0.2$)

DISTANCE PERMUTATION TEST ($\sigma_1=\sigma_2=1.0$)

DISTANCE PERMUTATION TEST ($\sigma_1=1.0, \sigma_2=0.2$)

CENTROID SEPARATION TEST ($\sigma_1=\sigma_2=1.0$)

CENTROID SEPARATION TEST ($\sigma_1=1.0, \sigma_2=0.2$)

Fig. 4.1

location test shown by means of the three solid circular dots. The abscissa measures separation of centroids in multiples of σ, which is 1.0 in all cases. Each solid circular dot represents the results of 10 experiments. Thus for $\mu = 1.0$, 3 out of 10 experiments resulted in a rejection of H_o. When the swarm centers were moved 3 units apart, then H_o was rejected in each of the 10 experiments. It was only after we had moved the swarm centers 0.5 units apart that the MST location test first declared the swarms differently located. Hence the MST power curve is zero in the range $0 \leqq \mu \leqq 0.5$. The form of the power curve for the present MST location test thus takes the shape indicated in Fig. 4.1.

To see this result in perspective, we performed a classical T^2 test for location under precisely the same sampling conditions as the MST location test. (For a brief review of the T^2 test, see Appendix B). The T^2 test results are shown by the open circles. We see that the MST test is closely comparable in power to the classical T^2 test. Later studies of the T^2 test (not recorded here) show that it is overly sensitive to the relative sizes of σ_1, σ_2, thereby causing the T^2 best to have false high power for μ values near 0. (Thus the T^2 test would, under $\sigma_1 \neq \sigma_2$ conditions, reject H_o, when H_o is nearly true, i.e., when $0 < \mu << 1$.) The MST location test on the other hand was well-behaved in this case, as indicated by the open square in Fig. 4.1, but, as its curve would rise to 1 only after this point, it is of relatively low power.

An interesting and potentially useful alternate location test is that summarized by the triangles. This is the result of the *distance permutation test* described in Appendix C. Two sets of triangles are shown: for the case $\sigma_1 = \sigma_2 = 1.0$ and for the case $\sigma_1 = 1.0$, $\sigma_2 = 0.2$. A *centroid test* (Appendix D) was also investigated. This is a naive variant of the T^2 test. It shows

somewhat higher power than the T^2 test when $\sigma_1 = \sigma_2 = 1.0$, but unfortunately shows the same erratic behavior as the T^2 test when $\sigma_1 = 1.0$, $\sigma_2 = 0.2$, and thus is not to be trusted.

In sum, the MST location test is comparable in power to the classical T^2 test when $\sigma_1 = \sigma_2$; but it drops in power when $\sigma_1 \neq \sigma_2$. The MST location test does not, however, exhibit the anomalous behavior of the classical T^2 under ($\sigma_1 \neq \sigma_2$) conditions. A potentially useful alternate test for location can be given by means of the distance permutation test of Appendix C.

B. Power of the Scale Test

The power of the MST scale test was examined by means of suitably constructed gaussian populations in 32-dimensional space. As in par A, we drew random samples from $N_p(\underline{\mu}_j, \sigma_j^2 \underline{I}_p)$ with p = 32. Recall that p is the number of time series of each data set. We set $\mu_1 = \mu_2 = 0$ (since data sets can always be given a common centroid prior to a variance analysis); we also set $\sigma_1 = 1$, and let σ_2 be of a variable magnitude σ in the interval $0 < \sigma < 1$. Fig. 4.2 summarizes the power curve of the scale test. Each dot is the result of ten experiments, one experiment consisting of the following steps. First, we made n = 46 random draws from $N_{32}(\underline{0}, \underline{I}_{32})$ to produce $\underline{D}$ and similarly 46 draws from $N_{32}(\underline{0}, \sigma\underline{I}_{32})$ to produce $\underline{M}$. Then, an MST was constructed on $\underline{D} \cup \underline{M}$; next, the number n_1 of degree 1 points of the MST were tallied; then the lower 5% critical size for x the number of degree 1 points of $\underline{M}$ was determined via (3.1); and finally a decision was made concerning the acceptance or rejection of the hypothesis H_o. The resultant curve (sketched on the basis of 3 choices of σ = 0.9, 0.8, and 0.3) shows satisfyingly great power, indicating for example that when $\sigma = 0.9$ the test detected 7 times out of 10 that the sampled swarm from $N_{32}(0, \underline{I}_{32})$ had a different scale than that of $N_{32}(\underline{0}, 0.9\ \underline{I}_{32})$.

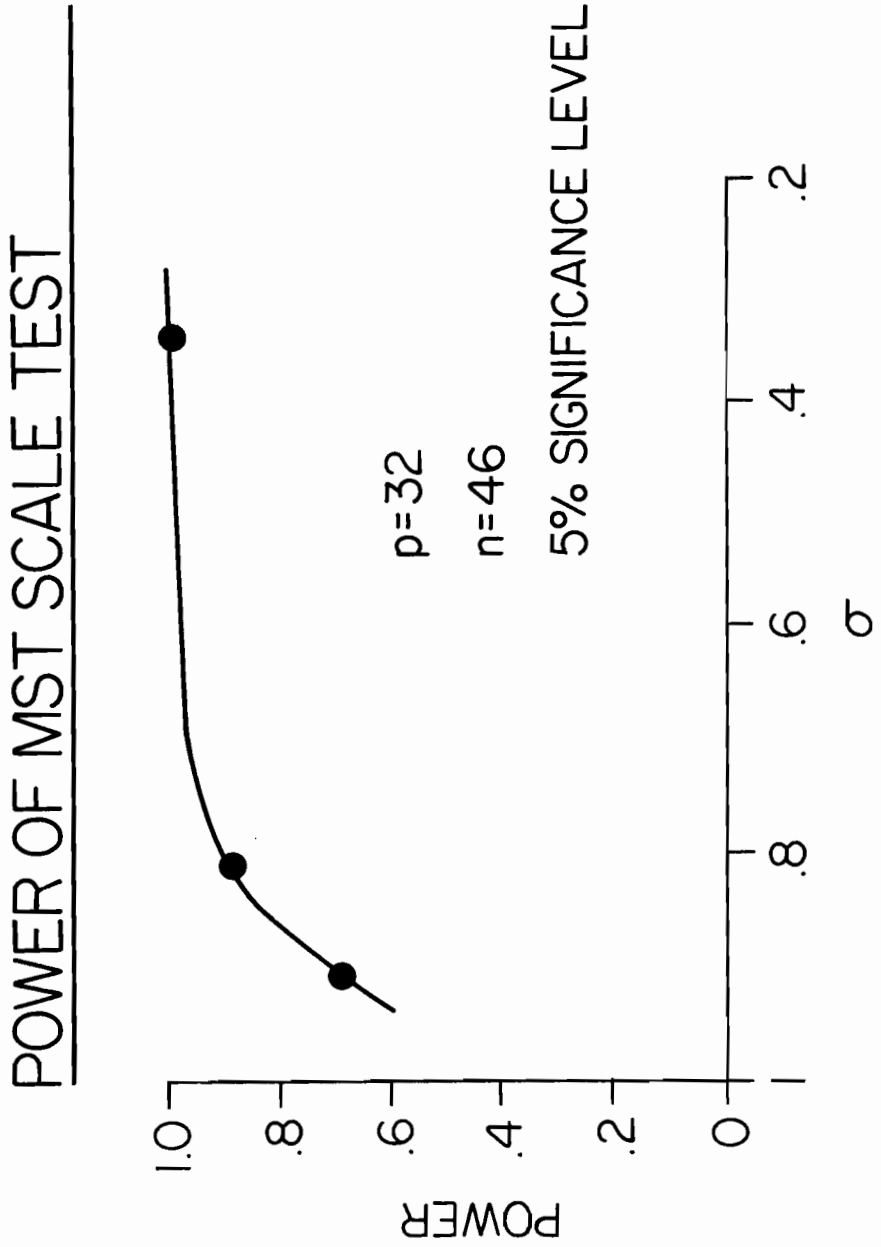

POWER OF MST SCALE TEST
POWER
1.0
.8
.6
.4
.2
0
.2
.4
.6
.8
1
σ
p=32
n=46
5% SIGNIFICANCE LEVEL

Fig. 4.2

In sum, the data-centered MST scale test has a workably high absolute power, and may be used with confidence to detect scale differences between centered data sets in high-dimensional euclidean spaces (i.e. with high-p parameters).

5. Natural Seasons: An Application of the MST Scale Test

A. We consider now a matter that is central to the problem of short-term climate predictability: how to group the months of a year into natural seasons of equal temperature variability. We envision taking each month as a base and then moving futureward of the month and finding those months whose temperature variabilities closely match that of the base month. Similarly, we can move backward in time from the base month to link up with those of its predecessors sharing equal temperature variability. In this way, having linked up each base month with its iso-variance partners, we can imagine a *running season* of temperature co-variable months, starting with January and its family of months, moving through February and March and their family of months, finally on to December with its family of months. It turns out that when we link up months this way, each base month is embedded in a "natural season," i.e., its family of temperature co-variable months. Predictions of monthly temperatures within a natural season should then be of nearly uniform skill--all other conditions affecting predictions being held the same. That is, as far as *variability* of temperature affects a prediction of monthly average temperature, the variability effect should be sensibly uniform within each natural season.

B. The data set we used in the natural-season study was supplied by Dr. T. P. Barnett of Scripps Institution of Oceanography and consisted of monthly averages of temperature (in °F) over the 46 year period from 1931 to 1976

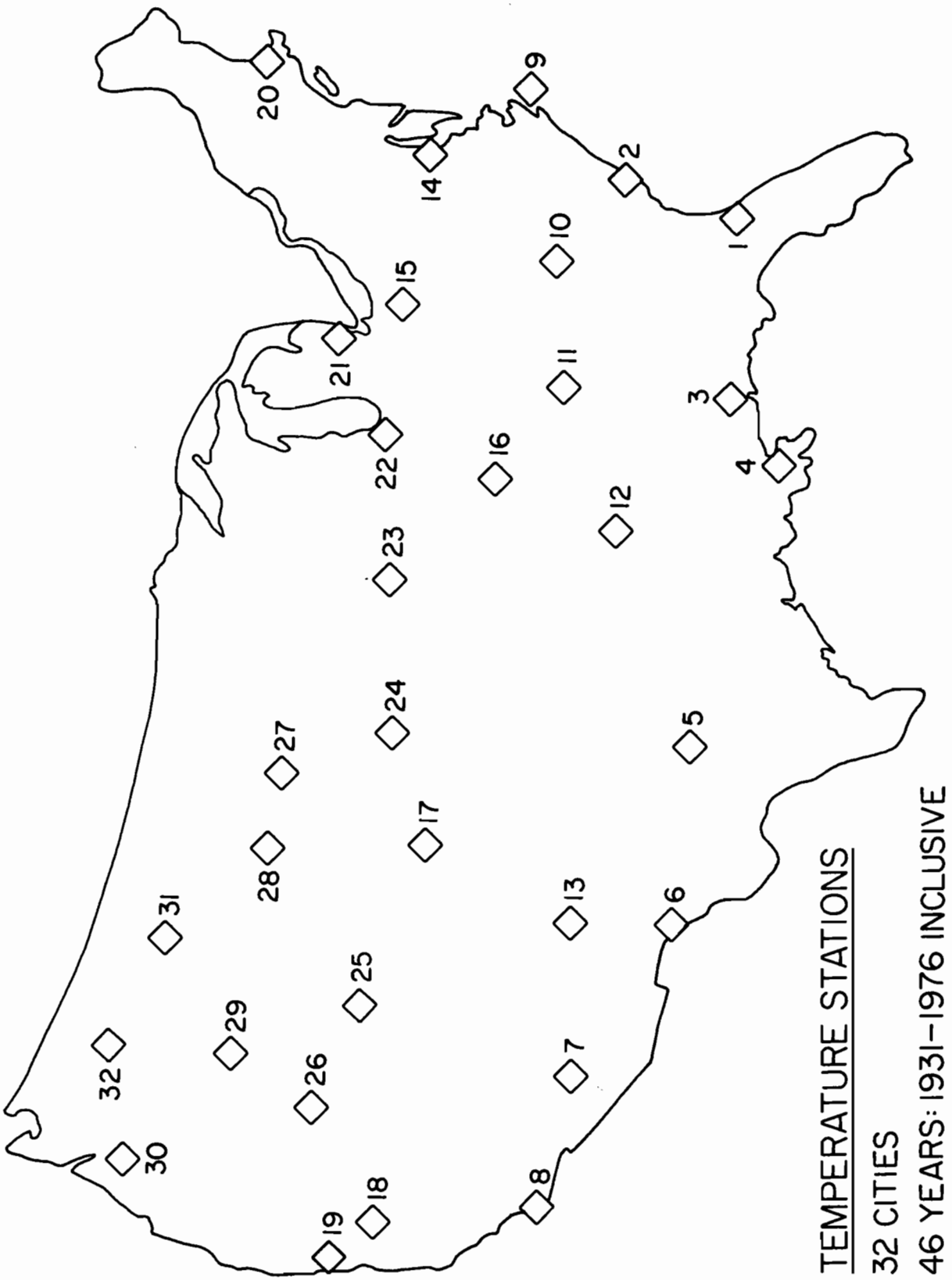

TEMPERATURE STATIONS
32 CITIES
46 YEARS: 1931–1976 INCLUSIVE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

Fig. 5.1

Table 5.1. 32 cities used. Numbers identify the cities on the accompanying U.S. map in Fig. 5.1.

1 is Jacksonville U/A to Waycross
2 is Charleston
3 is Mobile
4 is New Orleans
5 is Abilene
6 is El Paso
7 is Phoenix
8 is San Diego
9 is Cape Hatteras
10 is Asheville
11 is Nashville
12 is Little Rock
13 is Albuquerque
14 is Washington National
15 is Columbus
16 is St. Louis
17 is Denver
18 is Sacramento
19 is San Francisco
20 is Blue Hill Observatory
21 is Chicago
22 is Detroit
23 is Des Moines
24 is North Platte
25 is Salt Lake City
26 is Winnemucca
27 is Rapid City
28 is Sheridan
29 is Boise
30 is Portland
31 is Helena
32 is Spokane

collected at each of 32 cities of the U.S. mainland. Figure 5.1 shows the location of the cities and Table 5.1 lists their names. If we collected together in vector form the temperature anomalies (reckoned on the 46 year average) at each of the 32 cities for a given month, say January, then the representation of these January temperature anomalies would be as a point in a euclidean space of 32 dimensions, i.e., E_{32}. If then we project all 46 January points down on the plane spanned by the axes associated with any two cities, say Charleston and Jacksonville, we would obtain a sprinkling of crosses such as that shown in Fig. 5.2. Likewise the projection of the 46 points of E_{32} associated with August is given by the set of circles in Fig. 5.2. Because these points represent temperature *anomalies*, they are centered on the origin (0,0) of this plane. On inspecting the two sets of points, it is visually obvious that the temperature anomalies for January (as seen in the present two cities) have a greater variance than those for August. The relative variability of the temperature anomalies in these two cities for January and February are shown in Fig. 5.3. Now it is not so clear whether or not these variabilities are significantly different. When it comes to judging relative variability of the 46-point swarms in the full space E_{32}, we gladly turn this task over to the MST scale test.

C. We began the application of the MST scale test to the set of 46 January points in E_{32} by pooling them with the 46-point February set. We then found the number of degree-1 points in their MST, and determined with confidence 95% whether the February swarm had a significantly larger or smaller number of degree-1 points, relative to the January swarm, than expected. It turned out that it didn't, and so February was linked to January in the sense of having the same variability. Thus "F" was placed to the right of "J" in the bottom

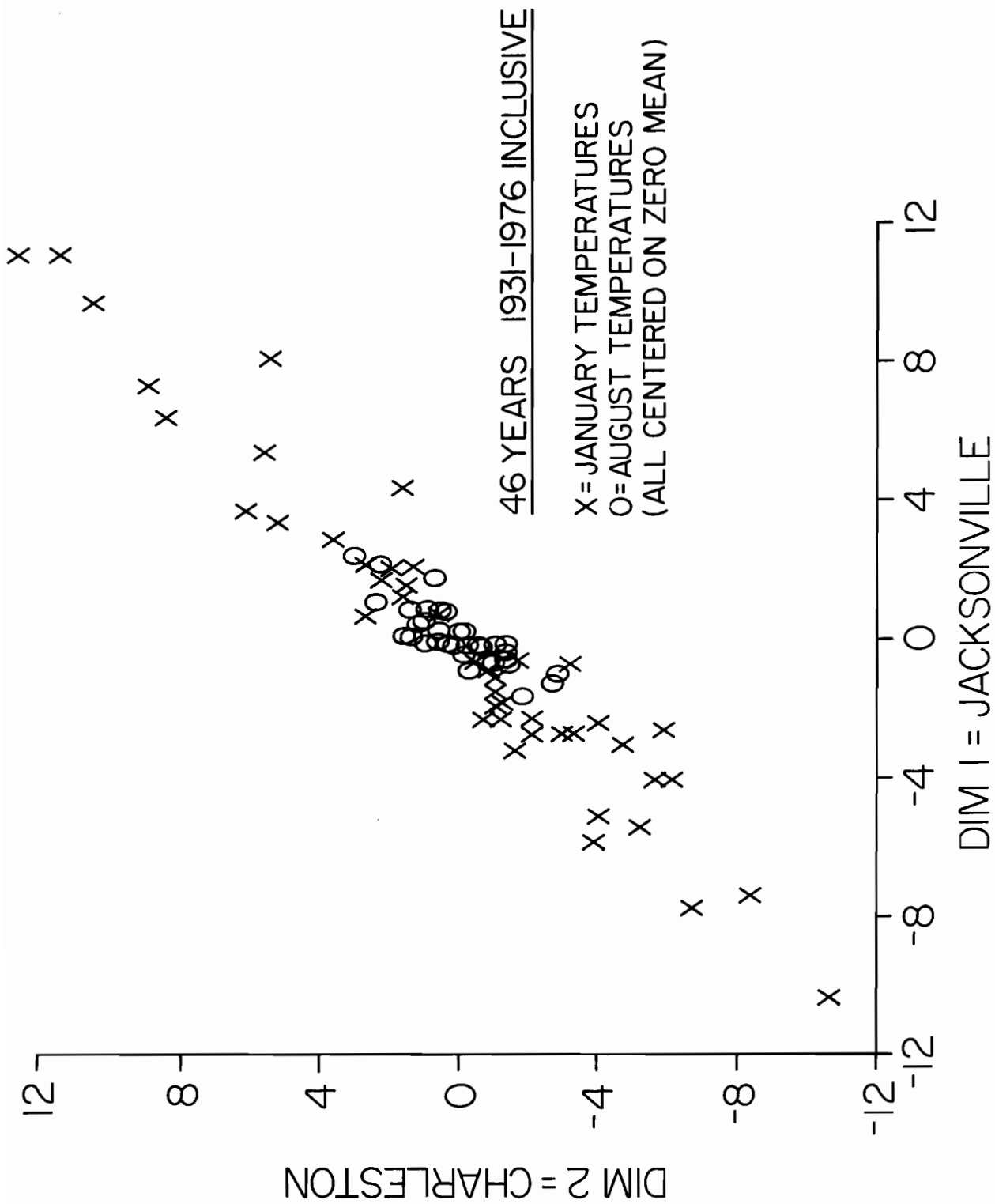

46 YEARS 1931-1976 INCLUSIVE
X=JANUARY TEMPERATURES
O=AUGUST TEMPERATURES
(ALL CENTERED ON ZERO MEAN)
DIM 1 = JACKSONVILLE
DIM 2 = CHARLESTON
12
8
4
0
-4
-8
-12

Fig. 5.2

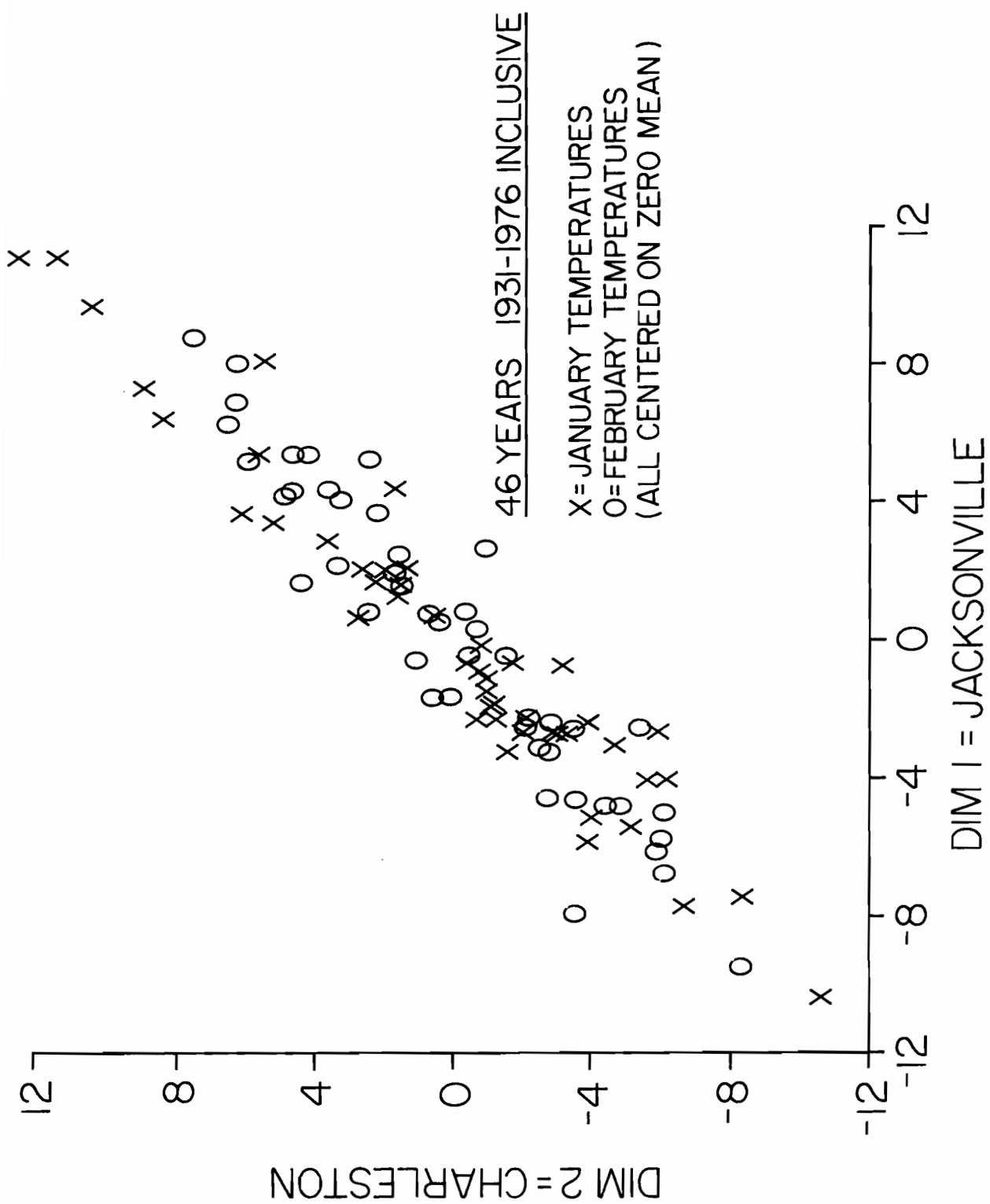

46 YEARS 1931-1976 INCLUSIVE
X=JANUARY TEMPERATURES
O=FEBRUARY TEMPERATURES
(ALL CENTERED ON ZERO MEAN)
DIM 1 = JACKSONVILLE
DIM 2 = CHARLESTON
12
8
4
0
-4
-8
-12

Fig. 5.3

rung of the ladder in Fig. 5.4. In like manner an MST for the January and March point swarms was constructed and it was found again that H_o could not be rejected on the 5% level. Thus "M" took its place to the right of "F" on the J-rung of the co-variability ladder. The process was repeated for all the swarms of the other months of the year, each swarm being pooled with January's and we found H_o rejected on the 5% level in each case. For this reason no other months are linked to the base month January. The natural season based on January therefore consists of the months January, February and March. This entire process, now based on February, was repeated for all other eleven months. The linkages are richer in this case, and the natural season based on February consists of December, January, February and March, as shown on the second rung of the co-variability ladder in Fig. 5.4. The remaining ten base months, embedded in their associated natural seasons, are shown in the remaining rungs of the co-variability ladder of Fig. 5.4.

During the calculations just outlined for the co-variability ladder, we also linked up those months with 90% confidence. This had the effect of discarding some of the months that had been linked up with 95% confidence. The results are shown in Fig. 5.5. They are replotted in a different way in Fig. 5.6(a). It is clear that, to some extent, a particular month's natural season membership depends on the significance level of the winnowing process: setting the significance level too low (~1%, say), would tend to link up more months to a base month. Setting it too high (~30%) would clip off more months from a base month than may be physically reasonable. Accordingly, to see if the couplings in Fig. 5.5 are reasonable, we undertook two alternate, somewhat different approaches to the natural-season problem, and these are discussed in the next two paragraphs.

MST-BASED SCALE TEST OF TEMPERATURE ISO-VARIANCE MONTHS

(HYPOTHESIZED CONNECTIONS REJECTED AT THE 5% LEVEL)

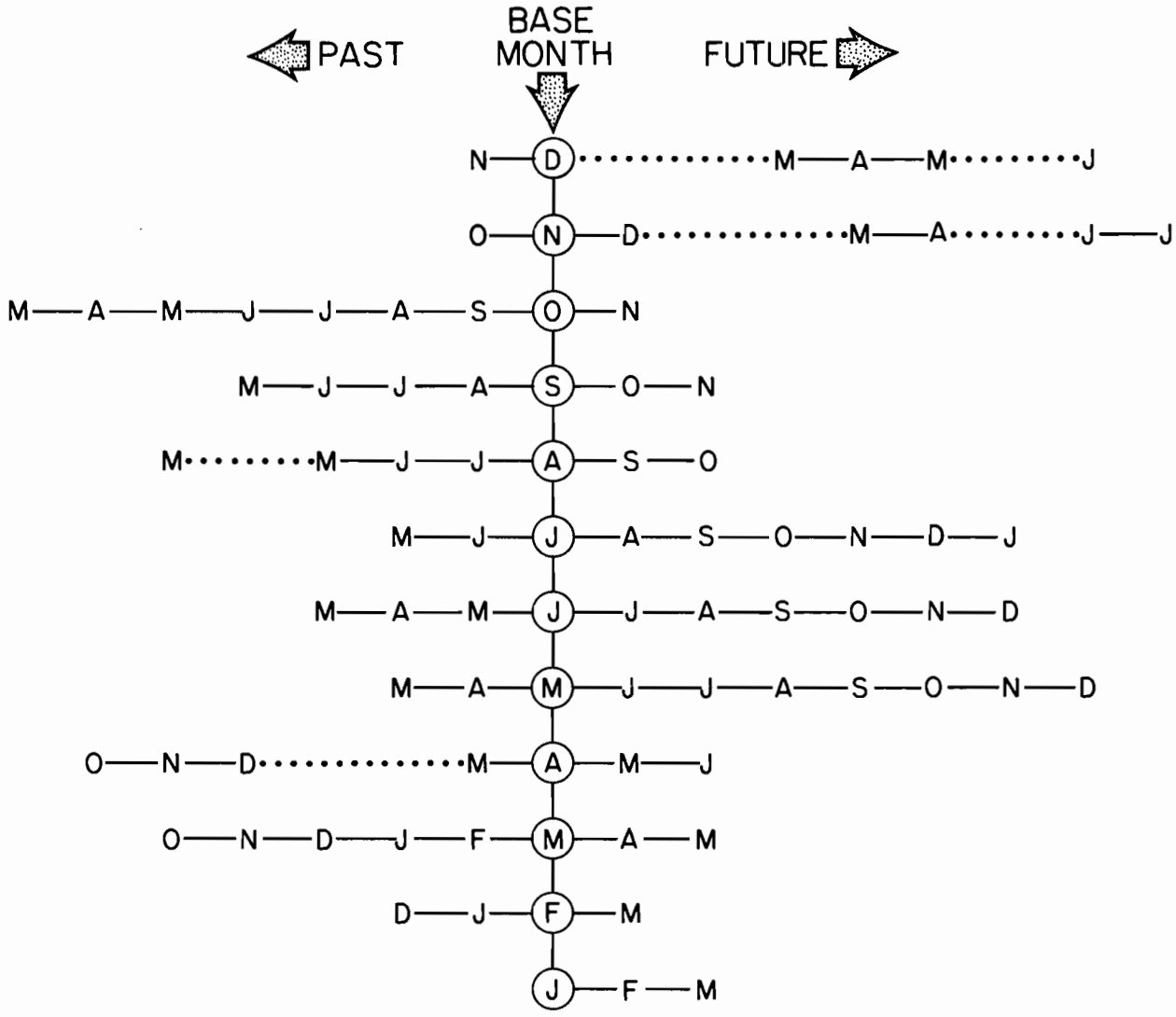


Fig. 5.4

MST-BASED SCALE TEST OF TEMPERATURE ISO-VARIANCE MONTHS

(HYPOTHESIZED CONNECTIONS REJECTED AT THE 10% LEVEL)

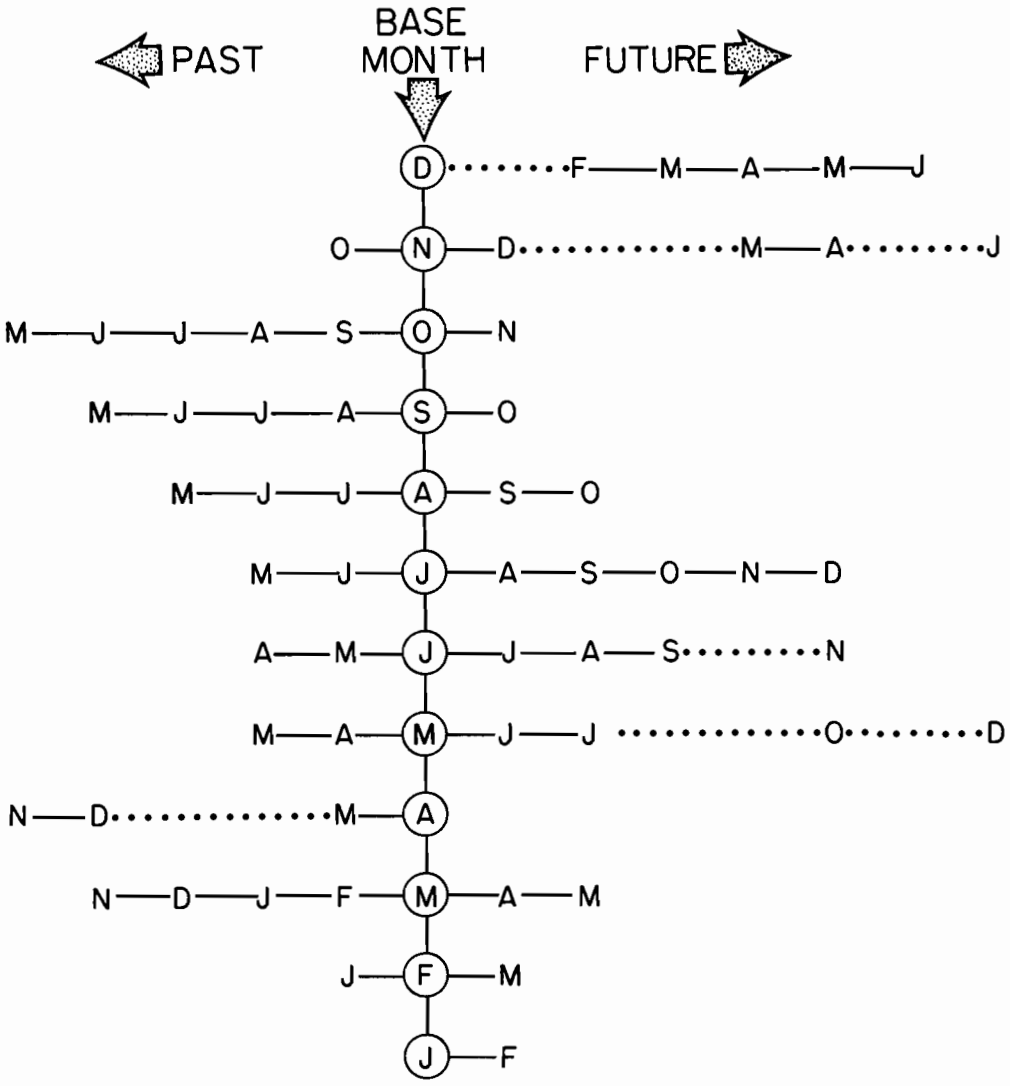


Fig. 5.5

Table 5.2. Values of x_{ij} and ξ_i derived from degree-1 points of the MST analysis of Scripps temperature data for 32 cities over 46 years (1931-1976).

index i		index j 1	2	3	4	5	6	7	8	9	10	11	12	x_i	$\xi_i = x_i/x$
JAN	1		21	21	24	28	26	22	22	28	26	28	25	271	.106
FEB	2	15		23	26	28	26	29	28	28	28	28	22	281	.110
MAR	3	13	17		17	20	23	25	23	23	21	21	20	223	.087
APR	4	13	14	12		22	25	29	30	28	25	19	16	233	.091
MAY	5	10	11	14	14		21	24	23	23	17	17	16	190	.074
JUN	6	11	12	14	17	16		20	24	24	26	16	14	194	.076
JUL	7	13	10	13	14	18	21		19	20	23	19	18	188	.073
AUG	8	11	10	14	13	15	19	21		23	15	14	12	167	.065
SEP	9	8	12	11	16	15	17	20	20		20	19	12	170	.066
OCT	10	11	13	14	18	21	19	22	18	17		20	9	182	.071
NOV	11	14	13	15	18	21	20	26	25	28	26		18	224	.088
DEC	12	11	14	21	18	22	23	24	25	24	29	26		237	.093

$x = 2560$

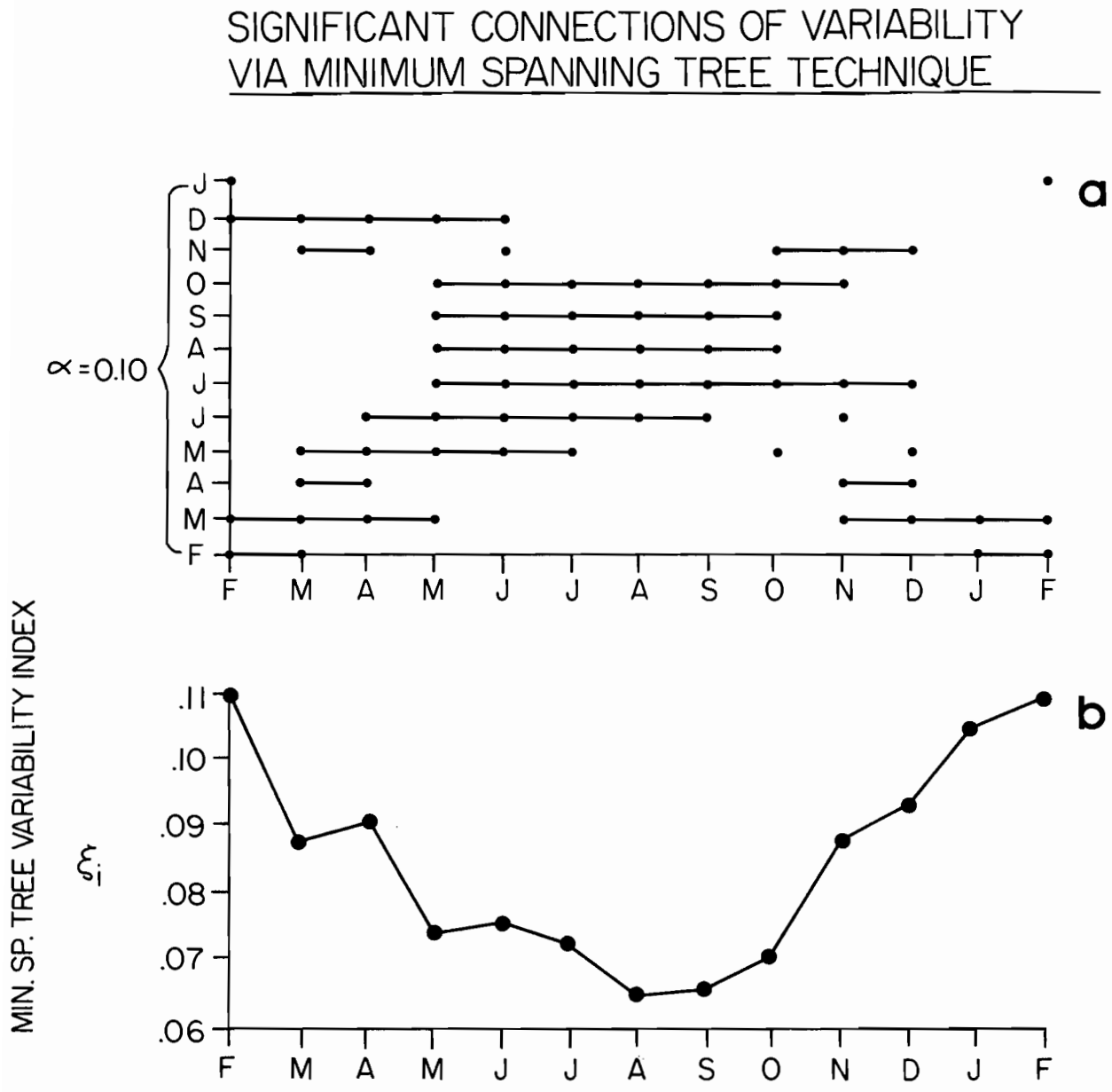

SIGNIFICANT CONNECTIONS OF VARIABILITY
VIA MINIMUM SPANNING TREE TECHNIQUE
a
α = 0.10
J
D
N
O
S
A
J
J
M
A
M
F
F M A M J J A S O N D J F
b
MIN. SP. TREE VARIABILITY INDEX
$\bar{\varsigma}$
.11
.10
.09
.08
.07
.06
F M A M J J A S O N D J F

Fig. 5.6

D. Returning to the setting defined in par C, let us denote by "x_{ij}" the number of the MST's degree-1 points in the swarm of the ith month when the ith month is pooled with the jth month (i, j = 1,...,12; 1 = Jan, 2 = Feb, etc.). An *absolute* measure of the spread (scale) of January temperatures is then defined as $x_1 \equiv x_{12} + \ldots + x_{1,12}$. In general, the absolute spread of the ith month's temperatures is $x_i = x_{i,1} + \ldots + x_{i,12}$ (omitting x_{ii}). Let $x = x_1 + \ldots + x_{12}$.

Table 5.2 lists the x_{ij}, the x_i, and the fractions $\xi_i = x_i/x$. A plot of ξ_i vs. i is shown in Fig. 5.6(b). By construction, ξ_i is a measure of the relative temperature variability of the ith month as sampled in the present U.S. mainland data set. It is immediately clear from Fig. 5.6(b) that the temperature variability is a minimum in August and a maximum in February, as seen using the MST degree-1 points. It can also be seen that the ξ_i curve is roughly sinusoidal, reflecting the expected annual swing of variance from maximum to minimum and back to maximum. If the variation of ξ_i with i = 1, ...,12, were exactly sinusoidal, then it would appear as shown in Fig. 5.7(b). Observe that, in such an idealized case, the variance of January matches that of March, while February's variance matches only that of itself, and moreover, the variance of April matches that of December, and so on. Under the relatively random conditions of the real world, we could imagine that the variance of January would match not only that of March, but also to some extent that of its neighbor February. Moreover, still being guided by the hypothesized exact sinusoid changes in ξ_i shown in Fig. 5.7(b), a month such as August would probably share similar variances with the neighbors, say June, July, September and October. These iso-variance connections are drawn systematically from the ξ_i-sinusoid and are depicted in Fig. 5.7(a). We used an approximately .01 ξ_i-interval above and below each level in Fig. 5.7(b) to define the iso-variance

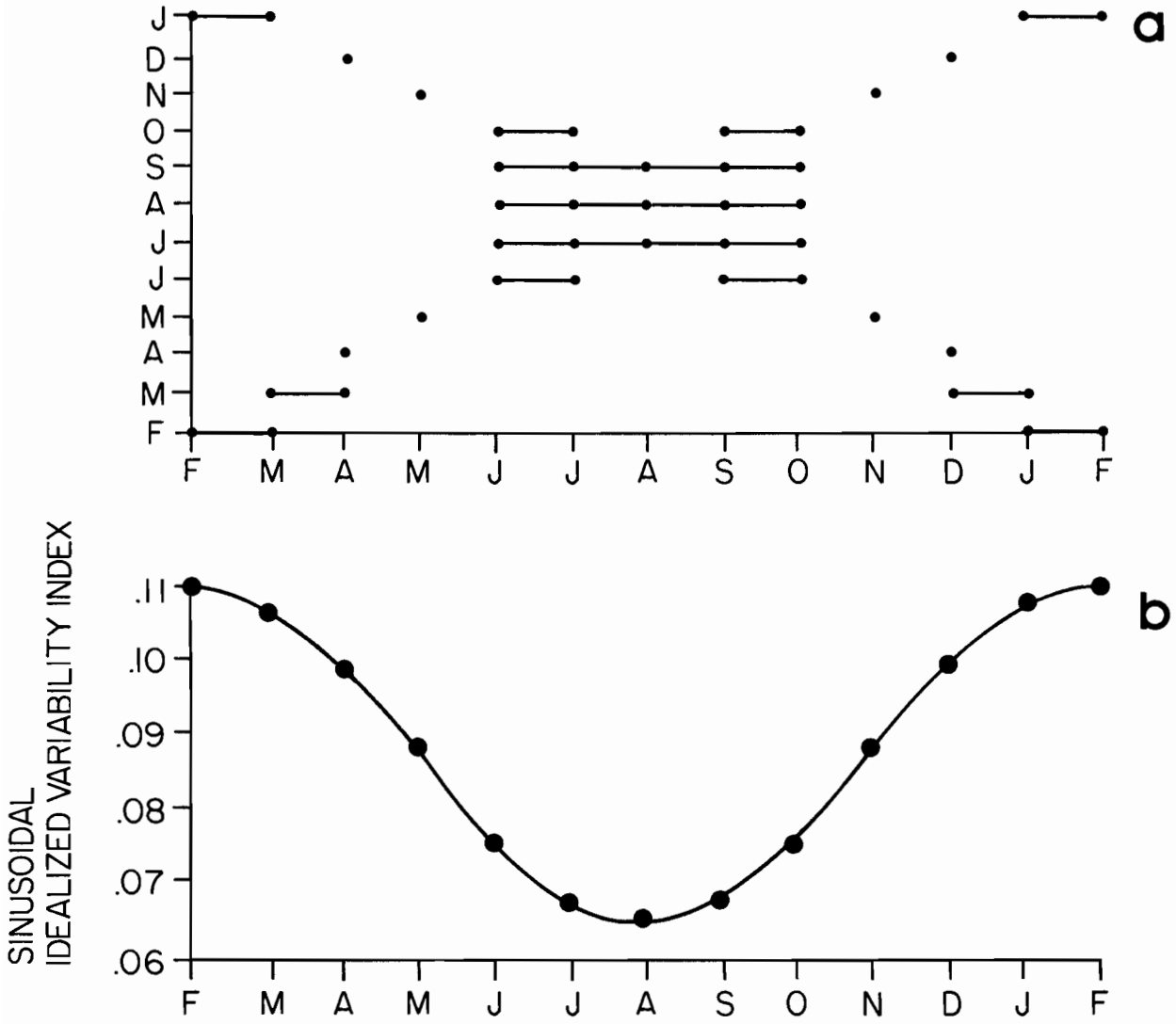

a
J
D
N
O
S
A
J
J
M
A
M
F
F M A M J J A S O N D J F
b
SINUSOIDAL
IDEALIZED VARIABILITY INDEX
.11
.10
.09
.08
.07
.06
F M A M J J A S O N D J F

Fig. 5.7

IDEAL CONNECTIONS OF TEMPERATURE ISO-VARIANCE MONTHS

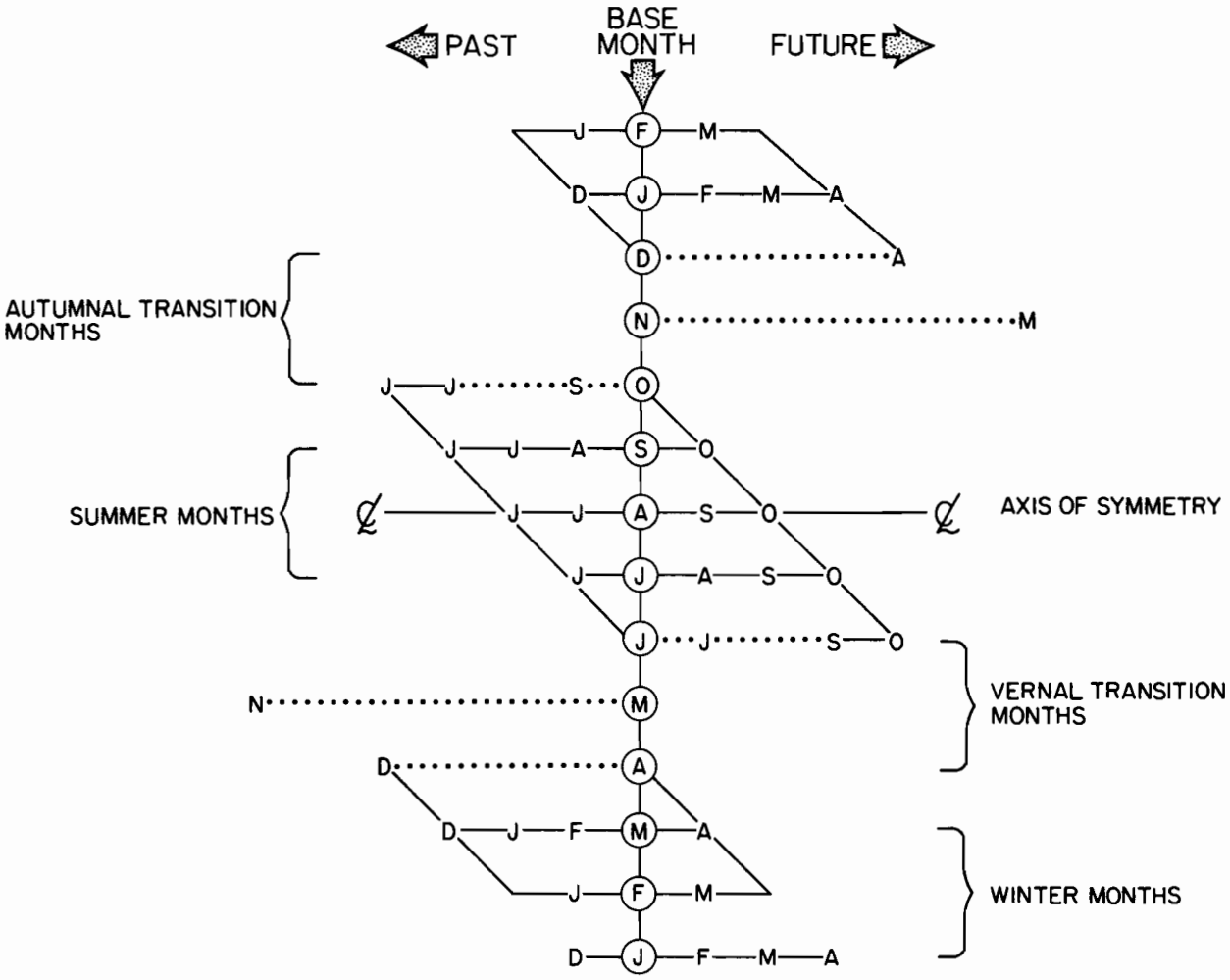


Fig. 5.8

neighbors of each month, as just sketched above. The pattern of connections in Fig. 5.7(a) is reminiscent of that in Fig. 5.6(a) (which, recall, is Fig. 5.5 redrawn). In this way we see the potential connection between the ξ_i curve of Fig. 5.6(b), and the natural seasons depicted in Fig. 5.6(a), which are simply the natural seasons of Fig. 5.5 drawn in the style of Fig. 5.7(a). Conversely, if we draw the idealized natural seasons of Fig. 5.7(a) in the style of Fig. 5.5, we obtain Fig. 5.8. This mode of representation brings out the ideal forms of the two sets of vernal and autumnal transition months, and the two solstice seasons: summer and winter.

E. As a check on the ξ_i index and the natural seasons defined in Fig. 5.5 or Fig. 5.6(a), we computed from the same data set the standard deviation of the temperature of the jth month, according to the formula

$$s_j^2 = \frac{1}{32} \cdot \frac{1}{46-1} \cdot \sum_{t=1}^{46} \sum_{x=1}^{32} [T_j(t,x) - \bar{T}_j(x)]^2 \qquad (5.1)$$

$$\bar{T}_j(x) = \frac{1}{46} \sum_{t=1}^{46} T_j(t,x)$$

$$j = 1,\ldots,12.$$

A plot of s_j vs. $j = 1,\ldots,12$ is shown in Fig. 5.9. A plot of ξ_i, rescaled by $452(\xi_i) - 25.5$, is shown plotted on the same diagram. The resemblance between the two curves is close, considering their diverse numerical bases. The ξ_i minimum in August is approximated by the s_i minimum in July. The s_i curve is somewhat the smoother of the two, but it does not suggest the ideal sinusoidal swing that ξ_i does around February. The ideal curve that seems to fit the s_i data is a parabola, as shown in Fig. 5.10(b). The "X" pattern of natural seasons belonging to this curve is shown in Fig. 5.10(a) in two forms: one

SUMMARY OF TWO VARIABILITY MEASURES FOR MONTHLY TEMPERATURES
32 CITIES, 46 YEARS, U.S. MAINLAND

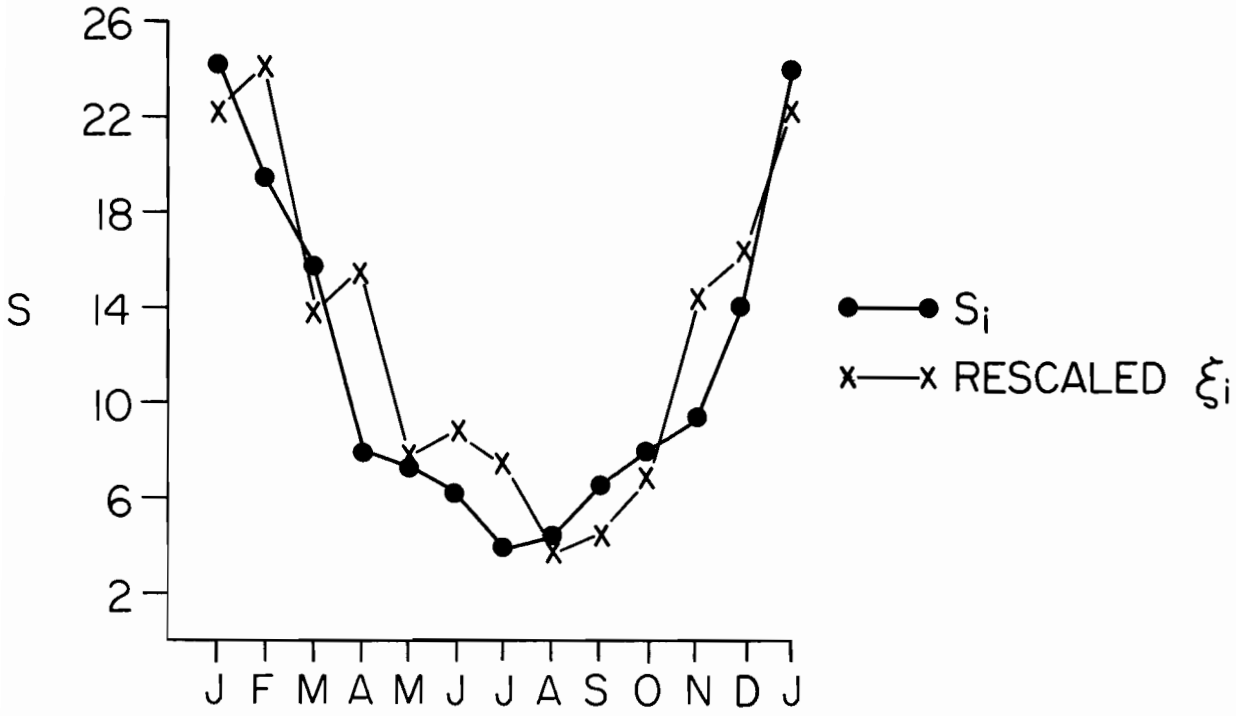


Fig. 5.9

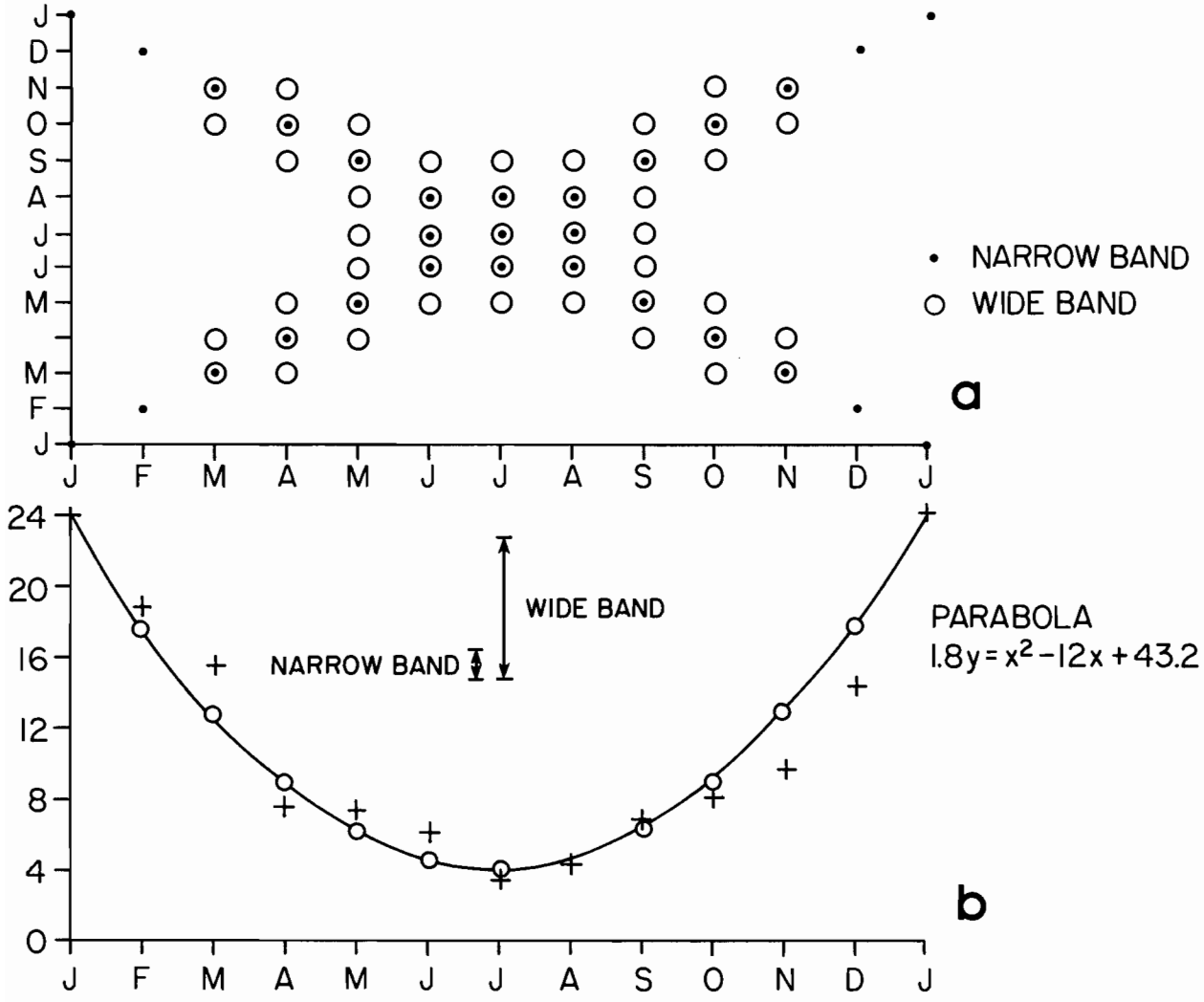

J
D
N
O
S
A
J
J
M
M
F
J
J F M A M J J A S O N D J
• NARROW BAND
○ WIDE BAND
a
24
20
16
12
8
4
0
WIDE BAND
NARROW BAND
PARABOLA
1.8y = x² − 12x + 43.2
b
J F M A M J J A S O N D J

Fig. 5.10

produced by a narrow-band neighbor snatcher (shown by dots), and one produced by a wide-band (shown by circles). These may be compared to the idealized natural seasons defined in Fig. 5.7(a). A natural season diagram may now be made from the ξ_i curve in Fig. 5.6(b), in the same manner. The connections will on the whole resemble those in Fig. 5.6(a). However, we prefer to remain with the connections of Fig. 5.6(a) (or Fig. 5.5), as they have been established using a rigorous method of statistically significant linkages.

In summary, we have discerned a set of natural seasons of months based at each month of the year, the common tie between the months being that of temperature variability. These natural seasons can be defined either via the MST method of determining the spread of a data set, or by the usual arithmetic statistical definition of standard deviation. The agreement between these two methods of definition appears to be close, but not exact. The MST method is based on a novel geometric method of defining variance similarities and establishing their statistical significance.

6. Bibliographic Notes

The main inspiration for the statistical method of this study rests in the paper of Friedman and Rafsky (1979), which applies the notion of a minimal spanning tree (MST) to the problem of the relative location and scale of a pair of data sets. Their paper in turn rests on two diverse areas of research. On the one hand there is the now classical runs-test of Wald and Wolfowitz (1940) which Friedman and Rafsky generalized to p-dimensional data space. On the other hand, there is the rapidly developing research area of graph theory (see, e.g., Harary, 1969) on which the theory of the MST is based. We have written our own program for MST constructions based on the simple intuitive

idea of a minimal spanning tree. As shown in Appendix A, it is a relatively trivial matter to grow the tree from a seed. However, it is possible to formalize the procedure, as has been done in particular by Prim (1957), and to program it efficiently (Whitney, 1972). Moreover, it is possible to grow trees with maximal speed in very high dimensional spaces, following the procedures of Bentley and Friedman (1975) and Rohlf (1977).

7. References

Works in this series on Data Intercomparison Theory (NOAA Technical Memorandums, ERL-PMEL):

DIT(I): Minimal Spanning Tree Tests for Location and Scale Differences.

DIT(II): Trinity Statistics for Location, Spread and Pattern Differences.

DIT(III): S-Phase and T-Phase Tests for Spatial Pattern and Temporal Evolution Differences.

DIT(IV): Tercile Tests for Location, Spread and Pattern Differences.

DIT(V): Case Study: Effects of Objective Analysis on a Tropical Pacific Sea Surface Temperature Set.

Anderson, T. W., (1958) An Introduction to Multivariate Statistical Analysis, Wiley and Sons, N.Y.

Bentley, J. L., Friedman, J. H. (1975) Fast algorithms for constructing minimal spanning trees in coordinate spaces. Stanford Linear Accelerator Report (SLAC) PUB-1665.

Friedman, J. H., Rafsky, L. C. (1979) "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests." Ann. of Statist. 7, 697.

Harary, F. (1969) Graph Theory, Addison-Wesley, Reading, Mass.

Mielke, P. W., Berry, K. J., Johnson, E. S., (1976) "Multi-response Permutation Procedures for A Priori Classifications." Comm. Statist. - Theor. Meth. A5, 1409.

Mielke, P. W., K. J. Berry, G. W. Brier (1981) "Application of Multi-response Permutation Procedures for Examining Seasonal Changes in Monthly Mean Sea-level Pressure Patterns," Mon. Wea. Rev. 109, 120.

Prim, R. C. (1957) "Shortest connection networks and some generalizations." Bell System Tech. J. 36, 1389.

Rohlf, F. J. (1977) A probabilistic minimal spanning tree algorithm. IBM Research Report C6502.

Wald, A., Wolfowitz, J. (1940) "On a test whether two samples are from the same population." Ann. Math. Statist. 11, 147.

Whitney, V. K. M. (1972) "Algorithm 422, minimal spanning tree." Comm ACM 15, 273.

Appendix A

Determining the Minimal Spanning Tree and Related Constructs

1. Constructing the Tree

A. We assume given two data sets $\underline{D}$ and $\underline{M}$ such that $\underline{D} = \{\underline{d}(1),\ldots,\underline{d}(n_D)\}$ and $\underline{M} = \{\underline{m}(1),\ldots,\underline{m}(n_M)\}$, where $\underline{d}(t) = [d(t,1),\ldots,d(t,p)]^T$, and $\underline{m}(t) = [m(t,1),\ldots,m(t,p)]^T$, i.e., we visualize $\underline{M}$ and $\underline{D}$ as two point swarms in euclidean p-space, respectively of n_D and n_M members. For the purpose of constructing the minimal spanning trees of $\underline{D}$ and $\underline{M}$, we consider their set union $\underline{X} = \underline{D} \cup \underline{M}$, where $\underline{X} = \{\underline{x}(1),\ldots,\underline{x}(n_D), \underline{x}(n_D + 1),\ldots,\underline{x}(n_D + n_M)\}$. Here we have identified the first n_D $\underline{x}$'s with the $\underline{d}(t)$'s: $\underline{x}(t) \equiv \underline{d}(t)$, $t = 1,\ldots,n_D$; and also we have $\underline{x}(t + n_D) \equiv \underline{m}(t)$, $t = 1,\ldots,n_M$. We assign the integers 1 through $n \equiv n_D + n_M$ to the points of this union, and retain these assigned numbers, throughout the constructions below, as permanent markers for the points of the $\underline{D}$ and $\underline{M}$ sets.

B. The jth point of $\underline{X}$ is $\underline{x}(j) = [x(j,1),\ldots,x(j,p)]^T$, $j = 1,\ldots,n$. To every pair $\underline{x}(i)$, $\underline{x}(j)$ of points of $\underline{X}$ we assign the distance

$$\mathrm{DIST}(i,j) = \left[\sum_{\ell=1}^{p} (x(i,\ell) - x(j,\ell))^2\right]^{\frac{1}{2}} \qquad \text{(A1.1)}$$

$i,j = 1,\ldots,n.$

These distances can be stored for use in the subsequent constructions. Since DIST(i,j) = DIST(j,i), we need only store DIST(i,j) for pairs i,j such that $i \leq j$.

C. The tree is constructed by first choosing a point of $\underline{X}$ and calling it the *seed*. With this choice, say point 1 of $\underline{X}$, there is created one point in the tree and n-1 outside the tree. Using the procedure below, we add one point to the tree during each stage. Hence at the end of the kth stage there are k points in the tree and n-k outside. To keep track of the points inside and outside of the tree, we define two arrays such that

$$\begin{aligned} &\text{INTREE}(i), \quad i = 1,\ldots,k \\ &\text{EXTREE}(j), \quad j = 1,\ldots,n-k \end{aligned} \tag{A1.2}$$

at stage k, k = 1,...,n. Thus "INTREE(i)" is the label of one of the k points in $\underline{X}$ which at this stage are in the tree. The arguments in the array have no *permanent* relation to the actual labels of the points of $\underline{X}$ assigned in par A.

We initialize these arrays by writing:

$$\left.\begin{aligned} &\text{INTREE}(1) = 1 \\ &\text{EXTREE}(1) = 2 \\ &\text{EXTREE}(2) = 3 \\ &\quad\vdots \\ &\text{EXTREE}(n-1) = n \end{aligned}\right\} \quad \begin{aligned} &\text{NIT} = 1 \\ &\\ &\\ &\text{NXT} = n-1 \\ &\\ &\\ &\end{aligned} \tag{A1.3}$$

Here NIT is the number of points in the tree, and NXT is the number outside the tree. These quantities, along with the INTREE and EXTREE arguments (defined below), will be updated during the constructions. This initialization in (A1.3) is stage 1.

D. Suppose we are at stage k, k = 1,...,n-1, with points in and out of the tree as given by (A1.2). We are ready to add a new point to the tree. Find

$$\text{DIST}(\text{INTREE}(i), \text{EXTREE}(j)) \tag{A1.4}$$

for i = 1,...,k and j = 1,...,n-k, using the pre-computed distances of pairs of points (recall (A1.1)). A subroutine finds the minimum of this set; or a simple running search of minimum distances can be incorporated as a few lines in the main program. In any case, when the k(n-k) distances have been examined, one should have in hand a particular pair of points of minimal distance apart, one in the tree, and the other its nearest neighbor outside of the tree, namely the pair:

$$\begin{aligned} &\text{MIN} = \text{INTREE}(\text{MN}) \\ &\text{MEX} = \text{EXTREE}(\text{MX}) \end{aligned} \qquad \text{(A1.5)}$$

Here MN is the INTREE index of the point MIN in the tree nearest to point MEX outside the tree, with MX the momentary EXTREE index of the latter. Observe that "MIN" and "MEX" are alternate and momentary names for the *original* labels of the points in $\underline{X}$.

E. To keep track of the accumulating set of points in the tree, as they are linked up to the tree in each stage, we write, at the end of stage $k \geq 2$, (recalling (A1.5)),

$$\begin{aligned} &\text{LINK}(k,1) = \text{MIN} \\ &\text{LINK}(k,2) = \text{MEX} \end{aligned} \qquad \text{(A1.6)}$$

F. To update the INTREE array at the end of stage $k \geq 2$, write

$$\text{INTREE}(k) = \text{MEX}$$

and to update EXTREE, proceed as follows (using (A1.5) notation):

$$\begin{aligned} &\text{For } j = 1,\ldots,n-k, \\ &\text{If } j < \text{MX}, \text{ then } \text{EXTREE}(j) = \text{EXTREE}(j) \\ &\text{If } j \geq \text{MX}, \text{ then } \text{EXTREE}(j) = \text{EXTREE}(j+1) \end{aligned} \qquad \text{(A1.7)}$$

G. The number of elements in INTREE and EXTREE are then updated:

$$\begin{aligned} &NIT = NIT + 1 \qquad \text{(number of points in tree)} \\ &NXT = NXT - 1 \qquad \text{(number of points not in tree)} \end{aligned} \tag{A1.8}$$

If NXT = 0, return to main program (tree is constructed)

H. Return to Step D above to enter a new stage. Thus (if NXT, above is $\geqq 1$), update the stage index k by 1.

2. Counting Runs

As described in the main text above (§2) the p-dimensional analog of the one-dimensional run is obtained from each permutation of point labels by looking for *linked* points that belong to *different* $\underline{M}$-like and $\underline{D}$-like sets. We now describe how to determine the runs in $\underline{D} \cup \underline{M}$ as well as in permutations of $\underline{D} \cup \underline{M}$. Let ϕ be a permutation of the integers $1,\ldots,n$. Let ψ be a function on $\{1,\ldots,n\}$ that gives the set membership of a point as being in either $\underline{D}$ or $\underline{M}$. Thus if $\underline{x}(j)$ is in $\underline{D}$ we write "$\psi(j) = 1$." If $\underline{x}(j)$ is in $\underline{M}$, we write "$\psi(j) = 2$." Recall that these assignations were made in §1A of this appendix. We are now formalizing this agreement by means of ψ.

We may now define the number of runs in the permuted-labels of the tree by initially setting for k = 1,

$$NRUN = 1.$$

Then for $k \geqq 2$,

$$\text{If } \psi(\phi[LINK(k,1)]) \neq \psi(\phi[LINK(k,2)]) \tag{A2.1}$$

set

NRUN = NRUN + 1,

otherwise, go on to k + 1 until n is reached. The final value of NRUN will give the number of runs in the MST for the particular permutation ϕ of the labels of the points. To build up the reference distribution of NRUN a large number of (say 100) permutations will be needed. To find the number of runs in the original union $\underline{D} \cup \underline{M}$, we simply use the identity permutation, i.e., work with the original $\underline{D}$ and $\underline{M}$.

3. Counting Degree Points

To find the degree of a point (the number of points in the tree to which it is linked) we use the LINK array. Thus consider point $\underline{x}(j)$. Go through all n-1 values LINK(k,1), k = 2,...,n, and tally up the number of times LINK(k,1) = j. Similarly, tally up the number of times LINK(k,2) = j. Then NDEG(j), the degree of $\underline{x}(j)$, is the sum of these two tallies.

4. Example

Consider the sets $\underline{D}$ (of three points) and $\underline{M}$ (of two points) as depicted in Fig. A4.1(a). Hence $\underline{D} = \{\underline{d}(1), \underline{d}(2), \underline{d}(3)\}$, and $\underline{M} = \{\underline{m}(1), \underline{m}(2)\}$. The union is $\underline{X} = \{\underline{x}(1), \underline{x}(2), \underline{x}(3), \underline{x}(4), \underline{x}(5)\}$. (The italicized integers will be explained later.) There are 10 distinct distances between all pairs of these five points. Choose point 1 (= $\underline{x}(1)$) as a seed*. Initialize INTREE and

---

* There generally will be a different tree if it is grown from another seed. Hence the MST based on point $\underline{x}(1)$ is not unique to the point set $\underline{X}$. See the section on orthogonal trees (§5) below.

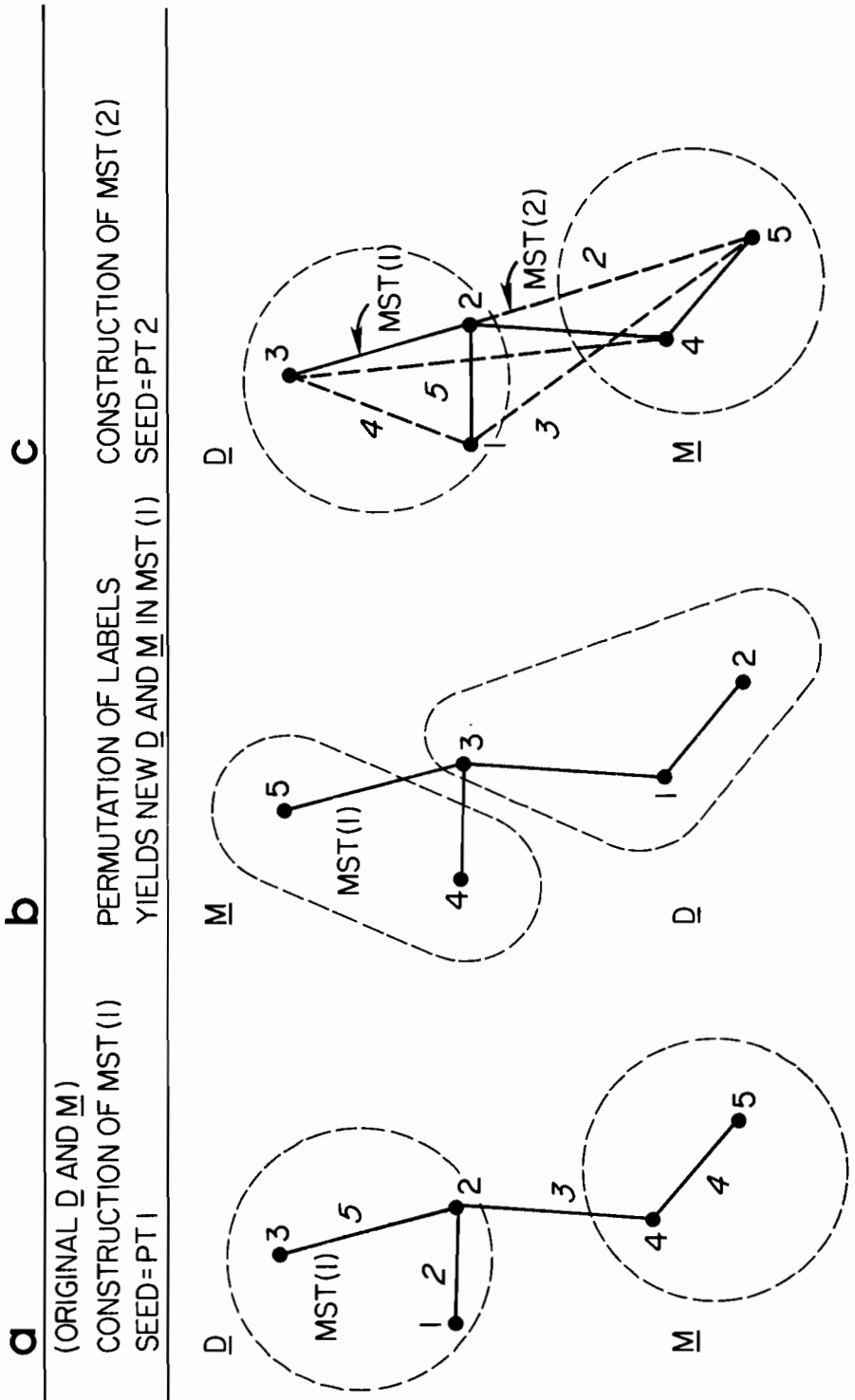

a
(ORIGINAL D AND M)
CONSTRUCTION OF MST (1)
SEED=PT 1
b
PERMUTATION OF LABELS
YIELDS NEW D AND M IN MST (1)
c
CONSTRUCTION OF MST (2)
SEED=PT 2
D
M
MST(1)
MST(2)

Fig. A4.1

EXTREE as in (A1.3). This ends stage 1. In stage 2, we find for (A1.5), i.e., for

MIN = INTREE(MN)

MEX = EXTREE(MX)

the values MIN=1, MEX=2 where MN=1, and MX=1.

Thus in stage 2, from (A1.6):

LINK(2,1) = 1
LINK(2,2) = 2 } end of stage 2

and further, for later stages

LINK(3,1) = 2
LINK(3,2) = 4 } end of stage 3 (A4.1)

LINK(4,1) = 4
LINK(4,2) = 5 } end of stage 4

LINK(5,1) = 2
LINK(5,2) = 3 } end of stage 5

The italic numbers in the diagram denote the stage in which the end points of the link are added to the tree. The reader can verify this sequence of inclusions in INTREE. Thus in stage 2, point 2 is added; in stage 3, point 4 is added. The diagram is drawn for the identity permutation. The number of runs in this tree is 2 (provided by link number 3). Here the set membership function ψ is given by $\psi(1) = \psi(2) = \psi(3) = 1$ and $\psi(4) = \psi(5) = 2$. The degree of point 1 is 1, and that of 2 is 3, as can be seen both from Fig. A4.1(a) and, by (A2.1), tallying up the number of times 1 or 2 (respectively appears on the right side of the LINK equations above.

When a permutation ϕ is applied to the point labels in the diagram we generate new $\underline{D}$ and $\underline{M}$ sets. Thus suppose ϕ is such that $\phi(1) = 4$, $\phi(2) = 3$, $\phi(3) = 5$, $\phi(4) = 1$, and $\phi(5) = 2$. Then the *same tree* now has $\underline{D}$ and $\underline{M}$ disposed

as sketched in Fig. A4.1(b). The seed of this tree is point 4 (i.e., point 1 relabeled).

This new pair of sets has three runs provided by links between points 3,4 and 3,5, in accordance with (A2.1). The degrees of points 4,5 and 2 are 1. Thus while the intrinsic degree value of a *point* in the tree is unchanged, the relabelling produces a new $\underline{D}$ or $\underline{M}$ set which may now have more or less points of a given degree than before, and together they may have a different number of runs. Thus set $\underline{D}$ originally had points 1,3 which were of degree 1. Now, the new $\underline{M}$ set has these same points (now labelled "4", "5") of degree 1. Before we had two runs, now we have three. By relabelling the same five points in the union of $\underline{D}$ and $\underline{M}$ in all the 120 various ways possible, and then counting the number of runs resulting from each new permutation, we can generate the cumulative reference distribution of this statistic for the purpose of conducting the location test described in the main text above. The reference distribution for the degree-1 points is, fortunately, known and of a simple type, and therefore does not require a permutation procedure (cf (3.1)).

5. Orthogonal Minimal Spanning Trees

The MST threads its sparse way through the set of n points of $\underline{X} = \underline{D} \cup \underline{M}$, missing many close neighbors whose inclusion in the graph could increase the power of the MST-based intercomparison tests. It is interesting to note that once an MST of points has been constructed, as shown in §4 of this appendix, we can start with a new seed, and build a completely different tree. For example, starting with point 2 of Fig. A4.1(a) and calling the tree there "MST(1)," the nearest neighbor is again point 1. But since the link between 1 and 2 is in MST(1) we look for the next nearest neighbor to point 2 with a

link not in MST(1). This is point 5, and we link up 2 and 5. Then from 2 and 5 we reach out to their nearest neighbors in $\underline{X} = \underline{D} \cup \underline{M}$ which have not been linked in MST(1). This is point 1, and we link up 5 and 1. Continuing this way we construct a new MST, namely MST(2) (see Fig. A4.1(c)) based on point 2 in $\underline{X} = \underline{D} \cup \underline{M}$. MST(2) in the sense of the construction just described, is *orthogonal* to MST(1).

The procedure for MST construction outlined in §1 may be followed in every detail in building MST(2). One only need add a new array CONN(i,j) defined as follows:

$$\text{CONN}(i,j) = \begin{cases} 0 \text{ if } i \text{ and } j \text{ are not linked by some MST} \\ 1 \text{ if } i \text{ and } j \text{ are linked by some MST} \end{cases} \qquad \text{(A5.1)}$$

This allows us to keep track of which point pairs of $\underline{X} = \underline{D} \cup \underline{M}$ have been used in some MST. See Fig. A4.1(c). When writing the program, for orthogonal tree constructions, we have found it helpful to label the various arrays by the index of the MST currently under construction. Thus in (A1.2) we now have

$$\begin{aligned} &\text{INTREE}(M,i), \quad i = 1,\ldots,k \\ &\text{EXTREE}(M,j), \quad j = 1,\ldots,n-k \end{aligned} \qquad \text{(A5.2)}$$

Here M is the MST index. In Fig. A4.1(c), as we begin the constructions, we set M = 2 as we use (A5.2). Starting with point 2 as seed, the order of occurrence of links is shown by the italic integers. Generally, M runs from 1 to n/2 or (n-1)/2, depending on the parity of n. In a similar fashion, the LINK array in (A1.6) can be expressed as

$$\begin{aligned} \text{LINK}(M,k,1) &= \text{MIN} \\ \text{LINK}(M,k,2) &= \text{MEX} \end{aligned} \qquad \text{(A5.3)}$$

during the construction of the Mth MST. As these successive MST's are being built, the array CONN in (A5.1) is continually updated.

When the number of orthogonal trees desired have been constructed (we used on the order of three or four) then the counting of runs and degree-1 points proceeds by going through each tree to find the runs and degree-1 points in that tree. The totals of the runs and degree-1 points are then found, and are used in the intercomparison tests just as in the case of single MST. In particular a fixed set of orthogonal MST's is now subject to the permutation procedure to find the reference distribution of runs associated with it.

APPENDIX B

The Classical T^2 Test for Location

Let $\underline{D} = \{\underline{d}(1),\ldots,\underline{d}(n)\}$ and $\underline{M} = \{\underline{m}(1),\ldots,\underline{m}(n)\}$ be two data sets of n points each in E_p. We shall need only this case where the number of points in each set is the same, namely n. Then define:

$$\underline{d} = n^{-1} \sum_{t=1}^{n} \underline{d}(t)$$

$$\underline{m} = n^{-1} \sum_{t=1}^{n} \underline{m}(t)$$

$$\underline{y}(t) = \underline{d}(t)-\underline{m}(t)$$

$$\underline{y} = \underline{d}-\underline{m} \quad .$$

Form the p×p covariance matrix

$$\underline{C} = (n-1)^{-1} \sum_{t=1}^{n} (\underline{y}(t)-\underline{y})(\underline{y}(t)-\underline{y})^T$$

and then the statistic

$$T^2 = n\, \underline{y}^T\, \underline{C}^{-1}\, \underline{y}.$$

If $\underline{D}$ and $\underline{M}$ are randomly drawn from $N(\underline{\mu}_D, \underline{\Sigma}_D)$ and $N(\underline{\mu}_M, \underline{\Sigma}_M)$ respectively, then the distribution of $[(n-p)/p(n-1)]T^2$ is noncentral F with p and n-p degrees of freedom and noncentrality parameter $n(\underline{\mu}_D-\underline{\mu}_M)^T\, \underline{\Sigma}^{-1}\, (\underline{\mu}_D-\underline{\mu}_M)$. If $\underline{\mu}_D = \underline{\mu}_M$, then the F distribution is central. In the present study we assume $\underline{\mu}_D = \underline{\mu}_M$ for the purpose of constructing the T^2 power curve. The theory of the T^2 test is given in Anderson (1958).

APPENDIX C

Distance Permutation Test for Location

Let $\underline{D}$ and $\underline{M}$ be as in Appendix B. Thus $\underline{D}$ and $\underline{M}$ may be visualized as n-point swarms in E_p. We define the ***separation index*** of any set $\underline{S} = \{\underline{s}(1), \ldots, \underline{s}(m)\}$ of m points in E_p as

$$SEP(\underline{S}) = \binom{m}{2}^{-1} \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \|\underline{s}(j) - \underline{s}(k)\|$$

where $\binom{m}{2} = \frac{1}{2}m(m-1)$

and $\|\underline{x} - \underline{y}\|$ is the euclidean distance between $\underline{x}$ and $\underline{y}$ in E_p. Thus $SEP(\underline{S})$ is the average distance between the set of points comprising $\underline{S}$.

Now let $\underline{X} \equiv \underline{D} \cup \underline{M}$, i.e., let $\underline{X}$ be the union of $\underline{D}$ and $\underline{M}$. We may write $\underline{X} \equiv \{\underline{x}(1),\ldots,\underline{x}(n), \underline{x}(n+1),\ldots,\underline{x}(2n)\}$, where the first n elements are from $\underline{D}$ and the remainder from $\underline{M}$. If ϕ is a permutation of the set $\{1,\ldots,n, n+1,\ldots,2n\}$ of 2n integers, then consider $\underline{S}_1(\phi) = \{\underline{x}(\phi(1)),\ldots,\underline{x}(\phi(n))\}$, $\underline{S}_2(\phi) = \{\underline{x}(\phi(n+1)), \ldots,\underline{x}(\phi(2n))\}$, which define a partition of $\underline{X}$ into two subsets of n elements each. Compute

$$SEP(\phi(\underline{X})) \equiv \tfrac{1}{2}(SEP(\underline{S}_1) + SEP(\underline{S}_2))$$

which is a measure of the average separation of points in the partition subsets $\underline{S}_1$, $\underline{S}_2$ of $\underline{X}$.

If we produce many permutations $\phi_1(\underline{X}),\ldots,\phi_r(\underline{X})$ ($r \cong 100$) of $\underline{X}$ and each time partition $\phi_j(\underline{X})$, as shown above, into $\underline{S}_1(\phi_j)$, $\underline{S}_2(\phi_j)$, and find $SEP(\phi_j(\underline{X}))$, then we can arrange the r $SEP(\phi_j(\underline{X}))$ values in ascending order to form a cumulative distribution of the SEP statistic.

This will give a reference background for the value

$$SEP(\underline{X}) \equiv \tfrac{1}{2}(SEP(\underline{D}) + SEP(\underline{M})).$$

Now, if $\underline{D}$ and $\underline{M}$, considered as point swarms in E_p, are relatively distant (compared to their average radii, say) then $SEP(\underline{X})$ will be relatively small compared to the average value of the set of values $SEP(\phi_j(\underline{X}))$, $j = 1,\ldots,r$. Hence if $SEP(\underline{X})$ falls in the left 5% tail, say, of the reference distribution, we would decide $\underline{D}$ and $\underline{M}$ have different locations. This is the essence of the distance permutation test. Our brief examination of its power in §4 of the text shows that it has useful power and, along with the MST, is less sensitive to the radii of the swarms $\underline{M},\underline{D}$ than the classical T^2 test. A general theory of such permutation tests as the present one is given in Mielke, Berry and Johnson (1976), and an application to meteorology is given in Mielke, Berry, and Brier (1981).

APPENDIX D

Centroid Test For Location

Let $\underline{D}$ and $\underline{M}$ be as in Appendix B, and let $\underline{X} = \underline{D} \cup \underline{M}$ be the union of these sets, as point swarms in E_p. We form the statistic $\|\underline{d} - \underline{m}\|$ which is a measure of their relative location. Introducing the permutations ϕ of $\underline{X}$, as in Appendix C, we can produce a reference distribution for $\|\underline{d} - \underline{m}\|$, and decide with confidence 95%, by examining the 5% right tail of the distribution, if $\underline{D}$ and $\underline{M}$ have different locations. If $\|\underline{d} - \underline{m}\|$ falls in this tail, we would say that the sets are differently located.

NOAA ERL technical reports, technical memoranda, and data reports published by authors at Pacific Marine Environmental Laboratory in Seattle, Washington, are listed below. Microfiche copies are available from the USDOC, National Technical Information Service (NTIS), 5285 Port Royal Road, Springfield, Virginia 22161 (703-487-4650). Hard copies of some of these publications are available from the ERL Library in Boulder, Colorado (303-497-3271). Hard copies of some of the technical reports are sold by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (202-275-9251).

NOAA Technical Report Series

ERL82-POL1 Naugler, Frederic P. (1968)
Bathymetry of a region (PORL-421-2) North of the Hawaiian Ridge, pre-NTIS.

ERL93-POL2 Grim, Paul J. (1968)
Seamap deep-sea channel, Jan. 1969, 2 824 50 060, pre-NTIS.

ERL118-POL3 Le Mehaute, Bernard (1969)
An introduction to hydrodynamics and water waves, 2 vols. 725 pp.
NTIS: PB192 065, PB192 066.

ERL146-POL4 Rea, David K. (1970)
Bathymetry and magnetics of a region (POL-421-3) 29° to 35°N, 155° to 165°W.
NTIS: COM-71-00173.

ERL191-POL5 Reed, R.K. (1970)
Results from some parachute drogue measurements in the central North Pacific Ocean, 1961-1962, 9 pp.
NTIS: COM-71-50020.

ERL214-POL6 Lucas, William H. (1971)
Gravity anomalies and their relation to major tectonic features in the North Central Pacific, 19 pp.
NTIS: COM-71-50409.

ERL229-POL7 Halpern, David (1972)
Current meter observations in Massachusetts Bay, 36 pp.
NTIS: AD-745 465.

ERL230-POL8 Lucas, William H. (1972)
South Pacific RP-7-SU-71 Pago Pago to Callao to Seattle.
NTIS: COM-72-50454.

ERL231-POL9 Halpern, David (1972)
Description of an experimental investigation on the the response of the upper ocean to variable winds, 51 pp.
NTIS: COM-72-50452.

ERL232-POL10 Stevens, H. R., Jr. (1972)
RP-1-OC-71 Northeast Pacific geophysical survey, 91 pp.
NTIS: COM-72-50677.

ERL234-POL11 Lucas, William H. (1972)
Juan de Fuca Ridge and Sovanco fracture zone.
RP-5-OC-71, 39 pp.
NTIS: COM-72-50854.

ERL240-POL12 Halpern, David (1972)
Wind recorder, current meter and thermistor chain measurements in the northeast Pacific-August/September 1971, 37 pp.
NTIS: COM-73-50107.

ERL247-POL13 Cannon, G. A. and Norman P. Laird (1972)
Observations of currents and water properties in Puget Sound, 1972, 42 pp.
NTIS: COM-73-50402.

ERL252-POL14 Cannon, G. A., N. P. Laird, T. V. Ryan (1973)
Currents observed in Juan de Fuca submarine canyon and vicinity, 1971. 57 pp.
NTIS: COM-73-50401.

ERL258-POL15 Lucas, William H., and Richard R. Uhlhorn (1973)
Bathymetric and magnetic data from the northeast Pacific 40° to 58°N, 125° to 160°W. 9 pp.
NTIS: COM-73-50577.

ERL259-POL16 Ryan, T. V., N. P. Laird, G. A. Cannon (1973)
RP-6-OC-71 Data Report: Oceanographic conditions off the Washington coast, October-November 1971, 43 pp.
NTIS: COM-73-50922

ERL260-POL17 Cannon, Glenn A. (1973)
Observations of currents in Puget Sound, 1970, 77 pp.
NTIS: COM-73-50666/9.

ERL261-POL18 Stevens, H. R. Jr., (1973)
RP-1-OC-70 Southeast Pacific geophysical survey, 60 pp.
NTIS: not available.

ERL271-POL19 Reed, Ronald K., and David Halpern (1973)
STD observations in the northeast Pacific, September-October 1972, 58 pp.
NTIS: COM-73-50923/4.

ERL292-PMEL20 Reed, R. K. (1973)
Distribution and variation of physical properties along the SEAMAP standard section, 16 pp.
NTIS: COM-74-50334/3.

ERL323-PMEL21 Erickson, B. H. (1975)
Nazca plate program of the international decade of ocean exploration--OCEANOGRAPHER Cruise-RP 2-OC-73, 78 pp.
NTIS: COM-7540911/6.

ERL325-PMEL22 Halpern, D., J. M. Helseth, J. R. Holbrook, and R. M. Reynolds (1975)
Surface wave height measurements made near the Oregon coast during August 1972, and July and August 1973, 168 pp.
NTIS: COM-75-10900/9.

ERL327-PMEL23 Laird, N. P., and Jerry A. Galt (1975)
Observations of currents and water properties in Puget Sound, 1973, 141 pp.
NTIS: COM-73-50666/9.

ERL333-PMEL24 Schumacher, J. D., and R. M. Reynolds (1975)
STD, current meter, and drogue observations in Rosario Strait, January-March 1974, 212 pp.
NTIS: COM-75-11391/0.

ERL339-PMEL25 Galt, J. A. (1975)
Development of a simplified diagnostic model for interpretation of oceanographic data.
NTIS: PB-247 357/7.

ERL352-PMEL26 Reed, R. K., (1975)
An evaluation of formulas for estimating clear-sky insolation over the ocean, 25 pp.
NTIS: PB-253 055/8.

ERL384-PMEL27 Garwood, Roland (1977)
A general model of the ocean mixed layer using a two-component turbulent kinetic energy budget with mean turbulent field closure, 81 pp.
NTIS: PB-265 434/1.

ERL390-PMEL28 Hayes, S. P., and W. Zenk (1977)
Observations of the Antarctic Polar Front by a moored array during FDRAKE-76, 47 pp.
NTIS: PB-281 460/6.

ERL390-PMEL29 Hayes, S. P., and W. Zenk (1977)
Observations of the Antarctic Polar Front by a moored array during FDRAKE-76, 49 pp.
NTIS: PB-281 460/6.

ERL403-PMEL30 Chester, Alexander J. (1978)
Microzooplankton in the surface waters of the Strait of Juan de Fuca, 26 pp.
NTIS: PB 297233/AS.

ERL404-PMEL31 Schumacher, J. D., R. Sillcox, D. Dreves, and R. D. Muench (1978)
Winter circulation and hydrography over the continental shelf of the northwest Gulf of Alaska, 16 pp.
NTIS: PB 296 914/AS.

ERL407-PMEL32 Overland, J. E., M. H. Hitchman, and Y. J. Han (1979)
A regional surface wind model for mountainous coastal areas, 34 pp.
NTIS: PB 80 146 152.

ERL412-PMEL33 Holbrook, J. R., R. D. Muench, D. G. Kachel, and C. Wright (1980)
Circulation in the Strait of Juan de Fuca: Recent oceanographic observations in the Eastern Basin, 42 pp.
NTIS: PB 81-135352.

ERL415-PMEL34 Feely, R. A., and G. J. Massoth (1982)
Sources, composition, and transport of suspended particulate matter in lower Cook Inlet and northern Shelikof Strait, Alaska, 28 pp.
NTIS: PB 82-193263

ERL417-PMEL35 Baker, E. T. (1982)
Suspended particulate matter in Elliott Bay, 44 pp.
NTIS: PB 82-246943.

ERL419-PMEL36 Pease, C. J., S. A. Schoenberg, J. E. Overland (1982)
A climatology of the Bering Sea and its relation to sea ice extent, 29 pp.
NTIS: not yet available.

ERL422-PMEL37 Reed, R. K. (1982)
Energy fluxes over the eastern tropical Pacific Ocean, 1979-1982, 15 pp.
NTIS: PB 83 138305

NOAA Data Report Series

ERL PMEL-1 Mangum, L., N. N. Soreide, B. D. Davies, B. D. Spell, and S. P. Hayes (1980)
CTD/O_2 measurements during the equatorial Pacific Ocean climate study (EPOCS) in 1979, 643 pp.
NTIS: PB 81 211203.

ERL PMEL-2 Katz, C. N., and J. D. Cline (1980)
Low molecular weight hydrocarbon concentrations (C_1-C_4), Alaskan continental shelf, 1975-1979, 328 pp.
NTIS: PB 82 154211.

ERL PMEL-3 Taft, B. A., and P. Kovala (1981)
Vertical sections of temperature, salinity, thermosteric anomaly, and zonal geostrophic velocity from NORPAX shuttle experiment, part 1, 98 pp.
NTIS: PB 82 163106.

ERL PMEL-4 Pullen, P. E., and H. Michael Byrne (1982)
Hydrographic measurements during the 1978 cooperative Soviet-American tsunami expedition, 168 pp.
NTIS: not yet available.

ERL PMEL-5 Taft, B.A., P. Kovala, and A. Cantos-Figuerola (1982)
Vertical sections of temperature, salinity, thermosteric anomaly and zonal geostrophic velocity from NORPAX Shuttle Experiment--Part 2, 94 pp.
NTIS:

ERL PMEL-6 Katz, C.N., J.D. Cline, and K. Kelly-Hansen (1982)
Dissolved methane concentrations in the southeastern Bering Sea, 1980 and 1981, 194 pp.
NTIS:

NOAA Technical Memorandum Series

ERL PMEL-1 Sokolowski, T. J. and G. R. Miller (1968)
Deep sea release mechanism, Joint Tsunami Research Effort, pre-NTIS.

ERL PMEL-2 Halpern, David (1972)
STD observations in the northeast Pacific near 47°N, 128°W (August/September 1971), 28 pp.
NTIS: COM-72-10839.

ERL PMEL-3 Reynolds, R. Michael and Bernard Walter, Jr. (1975)
Current meter measurements in the Gulf of Alaska--Part I: Results from NEGOA moorings 60, 61, 62A, 28 pp.
NTIS: PB-247 922/8.

ERL PMEL-4 Tracy, Dan E. (1975)
STD and current meter observations in the north San Juan Islands, October 1973.
NTIS: PB-248 825/2.

ERL PMEL-5 Holbrook, James R. (1975)
STD measurements off Washington and Vancouver Island during September 1973.
NTIS: PB-249 918/4.

ERL PMEL-6 Charnell, R. L. and G. A. Krancus (1976)
A processing system for Aanderaa current meter data, 53 pp.
NTIS: PB-259 589/0.

ERL PMEL-7 Mofjeld, Harold O. and Dennis Mayer (1976)
Formulas used to analyze wind-driven currents as first-order autoregressive processes, 22 pp.
NTIS: PB-262 463/3.

ERL PMEL-8 Reed, R. K. (1976)
An evaluation of cloud factors for estimating insolation over the ocean, 23 pp.
NTIS: PB-264 174/4.

ERL PMEL-9 Nakamura, A. I. and R. R. Harvey (1977)
Versatile release timer for free vehicle instrumentation over the ocean, 21 pp.
NTIS: PB 270321/AS.

ERL PMEL-10 Holbrook, James R. and David Halpern (1977)
A compilation of wind, current, bottom pressure, and STD/CTD measurements in the northeast Gulf of Alaska, February-May 1975.
NTIS: PB 270285.

ERL PMEL-11 Nakamura, A. I. and R. R. Harvey (1978)
Conversion from film to magnetic cassette recording for the Geodyne 102 current meter, 17 pp.
NTIS: PB-283 349/9.

ERL PMEL-12 Hayes, S. P., J. Glenn, N. Soreide (1978)
A shallow water pressure-temperature gage (PTG): Design, calibration, and operation, 35 pp.
NTIS: PB 286 754/7.

ERL PMEL-13 Schumacher, J. D., R. K. Reed, M. Grigsby, D. Dreves (1979)
Circulation and hydrography near Kodiak Island, September to November 1977, 52 pp.
NTIS: PB 297421/AS.

ERL PMEL-14 Pashinski, D. J., and R. L. Charnell (1979)
Recovery record for surface drift cards released in the Puget Sound-Strait of Juan de Fuca system during calendar years 1976-1977, 32 pp.
NTIS: PB 299047/AS.

ERL PMEL-15 Han, Y.-J. and J. A. Galt (1979)
A numerical investigation of the Bering Sea circulation using a linear homogeneous model, 40 pp.
NTIS: PB 299884/AS.

ERL PMEL-16 Loomis, Harold G. (1979)
A primer on tsunamis written for boaters in Hawaii, 10 pp.
NTIS: PB80-161003.

ERL PMEL-17 Muench, R. D. and J. D. Schumacher (1980);
(Hayes, Charnell, Lagerloef, and Pearson, contributors)
Some observations of physical oceanographic conditions on the northeast Gulf of Alaska continental shelf, 90 pp.
NTIS: PB81-102584.

ERL PMEL-18 Gordon, Howard R., ed. (1980)
Ocean remote sensing using lasers, 205 pp.
NTIS: PB80-223282.

ERL PMEL-19 Cardone, V. J. (1980)
Case studies of four severe Gulf of Alaska storms, 58 pp.
NTIS: PB81-102519.

ERL PMEL-20 Overland, J. E., R. A. Brown, and C. D. Mobley (1980)
METLIB--A program library for calculating and plotting marine boundary layer wind fields, 82 pp.
NTIS: PB81-141038.

ERL PMEL-21 Salo, S. A., C. H. Pease, and R. W. Lindsay (1980)
Physical environment of the eastern Bering Sea, March 1979, 127 pp.
NTIS: PB81-148496.

ERL PMEL-22 Muench, R. D., and J. D. Schumacher (1980)
Physical oceanographic and meteorological conditions in the northwest Gulf of Alaska, 147 pp.
NTIS: PB81-199473.

ERL PMEL-23 Wright, Cathleen (1980)
Observations in the Alaskan Stream during 1980, 34 pp.
NTIS: PB81-207441.

ERL PMEL-24 McNutt, L. (1980)
Ice conditions in the eastern Bering Sea from NOAA and LANDSAT imagery: Winter conditions 1974, 1976, 1977, 1979, 179 pp.
NTIS: PB81-220188.

ERL PMEL-25 Wright, C., and R. K. Reed (1980)
Comparison of ocean and island rainfall in the tropical South Pacific, Atlantic, and Indian Oceans, 17 pp.
NTIS: PB81-225401.

ERL PMEL-26 Katz, C. N. and J. D. Cline (1980)
Processes affecting distribution of low-molecular-weight aliphatic hydrocarbons in Cook Inlet, Alaska, 84 pp.
NTIS: not yet available.

ERL PMEL-27 Feely, R. A., G. J. Massoth, A. J. Paulson (1981)
Distribution and elemental composition of suspended matter in Alaskan coastal waters, 119 pp.
NTIS: PB82-124538.

ERL PMEL-28 Muench, R. D., J. D. Schumacher, and C. A. Pearson (1980)
Circulation in the lower Cook Inlet, Alaska, 26 pp.
NTIS: PB82-126418.

ERL PMEL-29 Pearson, C. A. (1981)
Guide to R2D2--Rapid retrieval data display, 148 pp.
NTIS: PB82-150384.

ERL PMEL-30 Hamilton, S. E., and J. D. Cline (1981)
Hydrocarbons associated with suspended matter in the Green River, Washington, 116 pp.
NTIS: PB82-148677.

ERL PMEL-31 Reynolds, R. M., S. A. Macklin, and T. R. Heister (1981)
Observations of South Alaskan coastal winds, 49 pp.
NTIS: PB82-164823.

ERL PMEL-32 Pease, C. H., and S. A. Salo (1981)
Drift characteristics of northeastern Bering Sea ice during 1980, 79 pp.
NTIS: PB 83 112466

ERL PMEL-33 Ikeda, Motoyoshi (1982)
Eddies detached from a jet crossing over a submarine ridge: A study using a simple numerical model, 38 pp.
NTIS: PB82-217563.

ERL PMEL-34 Liu, Cho-Teng (1982)
Tropical Pacific sea surface temperature measured by SEASAT microwave radiometer and by ships, 160 pp.
NTIS: not yet available.

ERL PMEL-35 Lindsay, R.W., and A.L. Comiskey (1982):
Surface and upper-air observations in the eastern Bering Sea, 90 pp.
NTIS: not yet available.

ERL PMEL-36 Preisendorfer, R., and C. E. Mobley (1982)
Climate forecast verifications off the U. S. mainland, 1974-1982, 225 pp.
NTIS: not yet available.