Scale-Controlled Objective Analysis

Katsuyuki V. Ooyama
Hurricane Research Division, AOML, NOAA, Miami, FL 33149

(Manuscript received 23 December 1986, in final form 6 April 1987)

ABSTRACT

The major topic of this paper is the resolvable spatial scales that can be analyzed by statistical interpolation of an undersampled dataset. The inquiry was motivated by the need to design the most appropriate procedures for spatial analysis of the upper-air sounding data from the GARP Atlantic Tropical Experiment. A reliable representation of horizontal scales in the analyzed wind fields was a matter of utmost concern, since the derived fields of vorticity, divergence and vertical motion were also of vital interest. To achieve our goal, it was found that the traditional premise of statistical interpolation had to be reexamined. The main conclusions of this theoretical inquiry are (i) resolvable scales are determined by the geometrical distribution of observing stations; (ii) precise knowledge of the second-moment statistics improves the analysis by de-aliasing the amplitudes of resolvable scales, but has no effect on the definition of resolvable scales; (iii) residual effects of unresolvable signals in the data are removable by a spatial filter and must be so removed; and (iv) spatial phases of the de-aliased resolvable scales may still be in error.

On the basis of these findings, the objective analysis procedures we have developed are targeted on the best achievable analysis of resolvable scales. The procedures include the following: an adequate estimation of "true" statistical fields from the given ensemble of data, a search for the optimum spatial filter by monitoring the targeted error variance, and a rational method of desensitizing the analysis to statistically errant data. In order to reduce the spatial phase error of propagating disturbances, the procedures are extended to the analysis of the timewise Fourier-transformed dataset (actually in the frequency-band analog). Since the wind is a physical vector, the entire procedure for the wind analysis is given in the tensor-invariant form, which is decidedly advantageous for very practical reasons. For example, the tensor approach eliminates the notorious ambiguity in normalization that is encountered in the multivariate approach. The paper also describes, in the Appendix, a method of filtered mechanical interpolation, which is specifically designed, with a variety of optional boundary conditions, for application to analysis in a finite domain.

1. Introduction

Analysis, in this paper, is a process of estimating the continuous spatial field of a physical variable from a set of discrete observational data. (The product of such a process, if not ambiguous, is also called an analysis.) In the ideal case, in which the domain of interest is densely covered by data of reasonable accuracy, all that is required of analysis may be mechanical interpolation of the discrete data with some smoothing. For most meteorological observations, however, especially those obtained by direct probes, the data are collected by a less-than-ideal number of irregularly placed stations; that is, some small-scale variations of the field are not adequately sampled. The common problem in analysis of such undersampled data is known as aliasing, or misrepresentation of spatial scales. In the case of irregularly placed data, the amplitudes of aliased scales are not necessarily conserved and may result in further distortion by overshooting.

Spatial analysis of the undersampled data must perform two tasks, related but distinct. One is the mechanical task, as before, of producing a continuous field. The other is a judgmental task of managing the total information input to the mechanical task, for the purpose of ensuring the resulting field to be a reasonable approximation of the true field. The latter task obviously requires some information about the true field, real or presumed, in addition to the given data. The required information does not substitute for real data but, rather, determines the ground rules of judgment for reducing the multitude of possibilities in the undersampled data to fewer, more likely possibilities. To reduce merely the number of possibilities, we may simply limit the degrees of freedom in the representation of the analyzed field, e.g., by fitting a low-order polynomial to the data. To "de-alias" the analysis of undersampled data, however, we need more factual information concerning the likely spatial structure of the field to be analyzed. This is the reason why, in meteorological analysis, judgments by a knowledgeable and experienced human analyst are respected, if not unanimously accepted.

While flexible human judgment is hard to emulate, automated implementation by a computer of any codified rule has the advantage of being consistent under the rule. Furthermore, codified rules are more amenable to orderly refinements than are subjective judgments. Thus, even in judgmental analysis, it is desirable to have the analysis procedure objectively defined and implemented. Procedural objectivity is not sufficient, however. The quality of judgmental analysis depends critically on the informational content of the implemented rules. Objectivity must be extended to the derivation of those rules from well-defined sources of information. In this respect, many methods in the literature of objective analysis are *objective* only in the procedural sense.

For example, the successive correction method by Cressman (1959) attempts to de-alias the analysis by assigning a spatially extended structure to each discrete datum. The assigned local structures are usually synthesized from a set of discretionary weighting functions. Since Cressman's original application to the analysis of synoptic-scale isobaric height fields, the method has been applied to a wide variety of meteorological situations differing both in spatial scales and in types of physical variables. In each new situation, however, the analyst must experiment with the weighting functions until, in his judgment, satisfactory results are obtained. The method itself contains no intrinsic information.

In contrast, a far greater degree of objectivity is possibly achieved by an analysis method that utilizes statistical interpolation or, as it is often referred to, optimum interpolation (Gandin, 1963). The principle of statistical interpolation is to minimize the error variance between the true and the analyzed fields. The minimization is taken at each point in space, over an ensemble of statistically similar events at different observation times. De-aliasing is made possible by the knowledge of spatial covariance functions for the ensemble of true fields. In this role, the covariance functions are similar to the local structure functions in Cressman's method, but are no longer discretionary. The error minimization also leads to a mathematical optimization of the combined effects of several adjacent data when their influences are overlapping. There is a hitch, however. The true fields are not knowable as postulated. If they were, there would be no need for analysis. Thus, the method is built on the premise that an adequate estimate of the true statistics is obtainable not from the knowledge of individual true fields but from statistical analysis of observational data.

In early applications of statistical interpolation to the initial data analysis for numerical prediction models (e.g., Eddy, 1967; Rutherford, 1972; Schlatter, 1975), attempts were made to estimate the required statistics from observational data, although autocovariance functions, mathematically modeled with a few disposable parameters, have been often used in operational models (e.g., Lorenc, 1981; Baker and Rosmond, 1985). These applications are mainly for the analysis of scalar fields, such as isobaric heights or streamfunctions, in a multivariate framework in which wind data are incorporated through the assumption of nondivergence, although the improved initialization techniques of current operational models permit the calculation of the initial divergence field that is consistent with the model. It is not the purpose of this paper to review the analysis for global models, but the design proposed by Daley (1985) for the direct analysis of both rotational and divergent components of winds, and the objective determination by Hollingsworth and Lönnberg (1986) of complete wind covariance functions from the massive FGGE data, are among the recent contributions to the wind analysis.

In the application of statistical interpolation we discuss below, the estimation of true covariance functions of highly divergent winds was from the outset the central concern of our objective analysis. The information encapsulated in the covariances is necessary not only for interpolating discrete data but also for de-aliasing the analysis of undersampled data. The latter, judgmental aspect of statistical interpolation has not attracted wide attention in the literature. (See section 3c for further comments.)

In 1974, the GARP Atlantic Tropical Experiment (GATE) was conducted for the purpose of studying scale interaction between convective cloud systems and their environment in the tropics (e.g., see Houze and Betts, 1981), deploying an unprecedented concentration of observation platforms and instrumentation over a maritime tropical region. Among the vast amount of data collected in GATE, there was a set of nearly 2000 upper-air soundings of wind, temperature and humidity taken by the international fleet of 15 ships over an hexagonal area of about 800 km across during the three weeks of the Phase III observation period.

The analysis of these soundings deserved high priority, since a reliable time history of the basic meteorological fields would serve as the foundation for interaction studies, allowing other data by radar, aircraft and satellites to be integrated. On the other hand, the double-hexagonal array of GATE ships with separation distances of 170 km or greater was not designed to resolve convective systems in detail. Thus, great care was needed to de-alias the analysis of such undersampled data in this convectively active region. The task of analysis was further complicated by the international mix of instrumentations which substantially differed from each other in accuracy, response and calibration. Consequently, the majority of diagnostic studies utilizing the GATE upper-air data were directed only to the areal mean properties and budgets over the entire ship array in categorized time-phases of the synoptic-scale wave (e.g., Thompson et al., 1979).

During his tenure at the National Center for Atmospheric Research (NCAR), the author made a determined effort to analyze the GATE upper-air soundings, applying the principle of statistical interpolation. The first goal was to quantify the time-varying, three-dimensional fields of horizontal wind, vorticity, divergence and vertical motion in the highest resolution that the given data would objectively allow. (The second goal of similarly analyzing thermodynamic fields was carried out later by S. K. Esbensen of Oregon State

University.) In GATE wind analysis, neither geostrophy nor nondivergence could be assumed, since horizontal divergence and vertical motions were the most critical fields of interest. Moreover, the reported pressure-height data were useless, due to instrumental biases. Thus, wind analysis had to be designed for direct use of wind data only. Statistical interpolation for this task required judgmental information in the form of tensor covariance functions of vector winds. No simplifying assumption that might prejudice spatial derivatives was allowed. Heuristic, mathematical modeling was also out of the question, since little was known of the statistical tensor structure of highly divergent, subsynoptic-scale winds in GATE or elsewhere.

The only way of deriving the required information was by mechanical interpolation, in space, of the discrete, sample covariances. This procedure would not yield the true covariance functions but only their estimates. Thus, followed the question: Would such estimates be adequate to achieve our goal? The data for the statistical estimation were an ensemble of the same, undersampled data that originally raised the specter of spatial aliasing. Any uncertainty in small-scale detail of those estimates would be amplified by spatial differentiation in calculation of divergence. On the other hand, unnecessary filtering or smoothing was to be avoided. Therefore, the question of adequacy required a clear answer before the *quality* of analysis could be associated with *objectivity* of procedures.

To answer the question above is the main purpose of this paper. It will be necessary to reexamine the principle of statistical interpolation in terms of resolvable spatial scales. The term *resolvable* refers here to the valid informational content of analyzed scales, and not to the mechanical degrees of freedom for representing them. The concept of resolvable scales has been obscure in statistical interpolation, since the traditional definition postulates the error variance to be minimized *independently* at each point of the analysis domain. To understand the consequence of this point-by-point minimization in the perspective of field analysis, we consider, first, the Fourier transform of statistical interpolation in an idealized case of equally spaced, one-dimensional data. The theoretical result, then, is generalized to the more realistic case of irregularly distributed, two-dimensional data. Specifically, the paper proceeds as follows:

The basic terminology and notation are defined in section 2, and the traditional principle of statistical interpolation is reiterated in section 3. Spectral properties of mechanical interpolation in the idealized case, especially the problem of the aliased main band, are reviewed in section 4. The idealized spectral examination is extended to statistical interpolation in section 5. It is concluded that the statistical method attempts to dealias the spectral amplitudes of the main band and that it also generates meaningless side bands. In other words, the resolvable scales are those scales in the main band and are solely determined by the geometrical placement of observing stations. Even the knowledge of true covariances cannot extend the resolution limit, although it will improve the quality of analysis within the resolvable scales. It is also shown that the residual components in the side bands are not only meaningless but are also harmful to field analysis, even though they contribute to minimization of the error variance. Therefore, the side bands must be removed from the analysis. A spatial filter for doing just that, without affecting the resolvable scales, will be called optimum. In brief, the true field, which is the assumed target of the traditional error minimization, is not an achievable target in analysis of discrete data.

The strategy for analyzing irregularly distributed data is discussed in section 6. Although the exact language of the Fourier transform does not apply any more, the concept of an optimum filter can be extended to a generalized definition of resolvable scales. The optimum-filtered true field, containing only the resolvable scales, is introduced as the best achievable target. Then, the minimization principle of statistical interpolation is redefined in terms of the error variance of analysis against the best achievable target. The new definition yields a practical procedure of determining the optimum filter for any particular ensemble of discrete data. It also answers our posted question, at least theoretically, by defining the *adequate* estimation of true statistics that is needed for analysis to attain the achievable goal. Actual estimation procedures are discussed in section 7.

The principle of statistical interpolation applies only to analysis of deviations from a norm. The norm must be defined and analyzed by other means. Only by using a model-predicted field as the norm has statistical interpolation become operationally practical in global applications. We did not have the benefit of a prediction model in the GATE analysis, but our task was a post-analysis of time-sequenced datasets. The question of the norm is discussed in section 8, together with other beneficial aspects of the time-sequenced data. For clarity of presentation, the discussions, summarized above, are developed for the analysis of a scalar variable. Additional comments for analysis of a vector variable are given in section 9. Both in theory and in application, some means of mechanical interpolation is needed as a working tool. The tool that we have developed for use in a finite domain with general boundary conditions is described in the Appendix.

2. Basic definitions and notation

a. Matrix notation

The purpose of this section is to define the terminology and mathematical notation that are basic to the rest of the paper. Although the results of the paper are applicable to multivariate analysis of several variables, our main presentation will be limited, for clarity, to the field analysis of a single scalar variable, $f(\mathbf{x})$. Extension to the analysis of a vector variable, $\mathbf{u}(\mathbf{x})$, is discussed in section 9.

By field analysis, we mean the estimation of $f(\mathbf{x})$, as a function of $\mathbf{x}$ in a prescribed domain $\mathcal{D}$, from a finite number of given data, $\hat{f}_j$, which are observational estimates of $f(\mathbf{x})$ at discrete points, $\mathbf{x} = \hat{\mathbf{x}}_j$, for $j = 1, 2, \cdots, J$, respectively. The resultant field, which we may simply call the *analysis,* will be denoted by $\tilde{f}(\mathbf{x})$ to clearly distinguish it from the target of estimation, the *true* field, $f(\mathbf{x})$. The analysis domain of interest is a finite horizontal area; that is, $\mathbf{x} = (x, y) \in \mathcal{D}$. The extent of the domain is largely dictated by the distribution of data points; $\mathcal{D}$ is normally an envelope of all $\hat{\mathbf{x}}_j$. Analysis is performed in space at a specific time, with the assumption that the given data are concurrent in time. The indication of time, t, will normally be omitted.

The concurrent set of the given data shall be written as a column matrix of J elements. Namely,

$$\{\hat{f}\} \equiv (\hat{f}_j)_{J\times 1}, \qquad (2.1)$$

where the pair of braces on the left-hand side signifies the column matrix and the right-hand side is a concise definition of the J-by-one column matrix with a typical element in parentheses. Similarly, the set of the observation points may be written as

$$\{\hat{\mathbf{x}}\} \equiv (\hat{\mathbf{x}}_j)_{J\times 1}, \qquad (2.2)$$

where each element of the column matrix is the position vector of a station. In the two-dimensional domain, there is no logically unique way to order a set of points. Thus, both the observing stations and the corresponding data may be indexed in any convenient order. However, once it is chosen, the same order must be kept.

It is important to note that the term *vector* is used in this paper only for physical and position vectors whose scalar components are subject to a definite rule of coordinate transformation. These vectors are printed in boldface type. On the other hand, a merely ordered set of discrete elements, which is a vector in linear algebraic terminology, is denoted by a column matrix, as it is in (2.1), and is always referred to as such. The discrete elements themselves may be vectors, as in (2.2), or functions of $\mathbf{x}$, as are introduced below. Extending the notation, we shall denote a doubly ordered set of discrete elements by a square matrix, in which the elements may be scalars or, later, tensors. The notational distinction, separating the ordered sets from the physical vectors and tensors, greatly facilitates the later transition from scalar analysis to vector analysis.

For the reason above, the use of braces { } and brackets [ ] will be reserved, in this paper, exclusively for the purpose of indicating those column matrices and square matrices, respectively, that are defined above. We also reserve angle brackets ⟨ ⟩ for the ensemble average.

b. *Representation of the analyzed field*

Since the number of given data, J, is finite, the analyzed field, $\tilde{f}(\mathbf{x})$, will not contain an infinite amount of independent information; thus, it can be adequately represented by finite degrees of freedom. The most common practice is to represent $\tilde{f}(\mathbf{x})$ by its values at regularly spaced grid points. However, also available are the spectral and other similar means of continuous representation by a finite number of basis functions (see the Appendix).

The number of basis functions, M, like that of grid points, defines the representational resolution. In a properly designed objective analysis, there is no need to restrict M by the number of data, J. In fact, unless the data coverage is redundantly dense, M should be sufficiently greater than J, so that enough degrees of freedom are left for the representation to accommodate judgmental information in addition to the data. In other words, we may choose M as large as necessary to suit the intended goal of analysis.

This does not imply, however, that the quality of analysis would improve with the increasing value of M. As we shall see, an absolute limit to the amount of meaningful information in the analysis is set, not by the mode of representation, but by the number and spatial distribution of the sampled data. Otherwise, the exact mode of representation is immaterial to our discussion; we shall maintain the function notation for $\tilde{f}(\mathbf{x})$, as being defined at every $\mathbf{x}$ in the analysis domain.

c. *The assumption of linear dependence*

The elementary principle that lies at the base of practically all the objective analysis methods is the assumption that the analyzed field, $\tilde{f}(\mathbf{x})$, at any $\mathbf{x}$, be linearly dependent on each and every datum $\hat{f}_j$. Under the standard convention for matrix operations, the assumption is expressed as

$$\tilde{f}(\mathbf{x}) = \{\psi(\mathbf{x})\}^{\mathrm{T}}\{\hat{f}\}, \qquad (2.3)$$

where

$$\{\psi(\mathbf{x})\} \equiv (\psi_j(\mathbf{x}))_{J\times 1} \qquad (2.4)$$

is a set of J spatial functions expressed as a column matrix. In (2.3), it is transposed to a row matrix as indicated by superscript T. Each element, $\psi_j(\mathbf{x})$, called an influence function, is the coefficient of linearity defining the influence of a datum at $\hat{\mathbf{x}}_j$ on the field at $\mathbf{x}$.

Under assumption (2.3), obtaining the analysis, $\tilde{f}(\mathbf{x})$, is equivalent to determining the set of influence functions, $\{\psi(\mathbf{x})\}$. If the observing stations, $\{\hat{\mathbf{x}}\}$, cover the analysis domain in sufficient density, or if the analyst so decides, $\{\psi(\mathbf{x})\}$ may be determined by a geometrical consideration of $\{\hat{\mathbf{x}}\}$, coupled with a spatial filter to remove the possible observational errors in $\{\hat{f}\}$. This is the case for a *mechanical* interpolation, implying that no additional information beyond the observational data, $\{\hat{\mathbf{x}}\}$ and $\{\hat{f}\}$, is prerequisite to the analysis. A method for mechanically interpolating irregularly distributed data in a finite domain is described in the Appendix.

On the other hand, if it is suspected that the given data inadequately sample the domain, the determination of $\{\psi(\mathbf{x})\}$ requires some knowledge of the true field, $f(\mathbf{x})$, in order to judge whether or not $\{\hat{f}\}$ is adequate and, if not, to compensate for the defects in sampling. This is the case for *judgmental* analysis. As discussed in the Introduction, statistical interpolation is the best known objective method for judgmental analysis, and it is the major topic of this paper.

3. Statistical interpolation

a. The traditional definition

The purpose of this section is to reiterate the traditional development of statistical interpolation. The problems that have been encountered in practical applications are also reviewed later in this section.

As elucidated by Gandin (1963), the theoretical principle of statistical interpolation is to minimize the second statistical moment of differences between the *true* field and the analyzed. The minimization requires an ensemble of datasets, of which $\{\hat{f}\}$ is a member set, drawn from statistically similar events, and a corresponding ensemble of true fields, of which $f(\mathbf{x})$ is a member. The datasets are given by the observations, but the true fields must be viewed as a rhetorical device for mathematical development.

The first statistical moment, or the norm, of the ensemble is not the subject of statistical interpolation. Following the tradition, we assume that the norm has been already subtracted from the data. Thus, in this paper, all the symbols for data and field variables, without customary primes ('), stand for deviations from the norm. Specifically,

$$\langle\{\hat{f}\}\rangle = 0 \quad \text{and} \quad \langle f(\mathbf{x})\rangle = 0, \tag{3.1}$$

where the pair of angle brackets $\langle \quad \rangle$ denotes the ensemble average.

The error variance, $\mathcal{E}(\mathbf{x})$, of analysis (2.3) against the true field is defined for the ensemble by

$$\mathcal{E}(\mathbf{x}) \equiv \langle(\tilde{f}(\mathbf{x}) - f(\mathbf{x}))^2\rangle, \tag{3.2}$$

and this is to be minimized by choosing $\{\psi(\mathbf{x})\}$. Since the minimization is taken independently at every $\mathbf{x}$, it leads to a formal definition of the optimum $\{\psi(\mathbf{x})\}$ as the solution of a matrix equation,

$$[\hat{m}]\{\psi(\mathbf{x})\} = \{m(\mathbf{x})\}, \tag{3.3}$$

where $[\hat{m}]$ is the J-by-J square matrix of sample-covariances, and $\{m(\mathbf{x})\}$ the column matrix of J covariance functions, defined by

$$[\hat{m}] \equiv (\hat{m}_{jj'})_{J\times J} = \langle\{\hat{f}\}\{\hat{f}\}^{\mathrm{T}}\rangle, \tag{3.4a}$$

$$\{m(\mathbf{x})\} \equiv (m_j(\mathbf{x}))_{J\times 1} = \langle\{\hat{f}\}f(\mathbf{x})\rangle, \tag{3.5a}$$

respectively. The same, in terms of matrix elements, are

$$\hat{m}_{jj'} = \langle\hat{f}_j\hat{f}_{j'}\rangle, \tag{3.4b}$$

$$m_j(\mathbf{x}) = \langle\hat{f}_j f(\mathbf{x})\rangle = \langle f(\hat{\mathbf{x}}_j)f(\mathbf{x})\rangle. \tag{3.5b}$$

The rightmost equality in (3.5b) is based on the assumption that, over the ensemble, observational errors in $\{\hat{f}\}$ are not correlated with $f(\mathbf{x})$ at every $\hat{\mathbf{x}}_j$ of $\{\hat{\mathbf{x}}\}$. Therefore, $\{m(\mathbf{x})\}$ is considered to be a set of the *true* covariance functions.

By definition, $[\hat{m}]$ is a positive definite matrix; its inverse matrix exists. Thus, with the solution of (3.3), analysis (2.3) becomes statistically optimum in the form

$$\tilde{f}(\mathbf{x}) = \{m(\mathbf{x})\}^{\mathrm{T}}[\hat{m}]^{-1}\{\hat{f}\}. \tag{3.6}$$

The minimized error variance of analysis (3.6) is

$$\begin{aligned} E(\mathbf{x}) &\equiv \min_{\psi}\mathcal{E}(\mathbf{x}), \\ &= s^2(\mathbf{x}) - \{m(\mathbf{x})\}^{\mathrm{T}}[\hat{m}]^{-1}\{m(\mathbf{x})\}, \end{aligned} \tag{3.7}$$

with the variance of the true fields given by

$$s^2(\mathbf{x}) = \langle f(\mathbf{x})^2\rangle. \tag{3.8}$$

b. Problems in practice

In application of the above theory, there are a few practical problems that must be addressed; in particular, analysis of the norm, estimation of true covariance functions $\{m(\mathbf{x})\}$, and hypersensitivity of the analysis (3.6).

Since statistical interpolation applies only to the deviations from a norm, the norm must be defined, first as discrete values at $\{\hat{\mathbf{x}}\}$, and then as a continuous field in $\mathbf{x}$. The latter, i.e., the norm field, must be added back to the analyzed deviation field in order to recover the total field. Although a constant field and a climatological mean field are often mentioned as possible candidates for the norm, these are too crude or unreliable in real applications. Ideally, the norm should define a slowly varying field of large spatial scales, while more transient, smaller-scale disturbances are contained in the deviations. In reality, a clear separation of scales between the norm and deviations is not always possible with a given ensemble of data.

For example, a meaningful analysis of the global norm field, directly from operational synoptic data, is practically impossible due to a large variation in data coverage from one region to another; the model-predicted field from an earlier time is now commonly used as the norm in the analysis of initial fields for global prediction models. In our GATE analysis, the number of ship stations in the array was too small to separate the norm-defining scales by spatial considerations only. Therefore, the separation had to be made, at each station, in terms of temporal frequencies, and the analysis of the norm field required mechanical interpolation with some subjective constraints (see section 8).

It is clear in (3.6) that the ability of this formula to interpolate the discrete data is derived solely from the knowledge of $\{m(\mathbf{x})\}$ as continuous spatial functions. If we are also interested in spatial derivatives of the analyzed field, the derivatives of $\{m(\mathbf{x})\}$ must be reliably known. On the other hand, as the means of determining $\{m(\mathbf{x})\}$, the formal definition (3.5a, b) is useless in practice, since we never know the true field, $f(\mathbf{x})$, as a spatial function. As we have discussed in the Introduction, we cannot always assume that a heuristic modeling of $\{m(\mathbf{x})\}$ is possible and acceptable. An attempt can be made to estimate $\{m(\mathbf{x})\}$ objectively from the discrete sample covariances, $[\hat{m}]$, but there still remains the question of the adequacy of such estimates. This paper will try to answer this question by examining the realistic goal of analysis that is achievable by statistical interpolation.

In most applications, especially in operational ones, a new dataset $\{\hat{f}\}$ is introduced at every analysis time, and the ensemble of $\{\hat{f}\}$ becomes an open-ended, theoretical concept. In practice, therefore, not only the elusive $\{m(\mathbf{x})\}$, but also the theoretically definable $[\hat{m}]$ are synthesized by an assumed model of autocovariance, so that $[\hat{m}]$ is not assured to be positive definite. Even if $[\hat{m}]$ is somehow made to be positive definite and computationally invertible, it is a notorious fact that the analysis (3.6) may unexpectedly produce bewildering results. The cause of this difficulty is the extreme sensitivity of (3.6) to those datasets, $\{\hat{f}\}$, that do not fit quite well to the statistically probable range of patterns expected by $[\hat{m}]$. Paradoxically, the more trivial the statistical mismatch is, the greater the sensitivity becomes.

Beginning with Gandin (1963), the most common remedy for the hypersensitivity has been an extra "observational" error variance that is added to the diagonal elements of $[\hat{m}]$. By raising all the eigenvalues of the matrix, it desensitizes the contribution from statistically errant data, but also degrades the overall quality of the analysis. Furthermore, this remedy introduces its own paradox that, as Gandin observed, the better the data coverage is, the greater becomes the need for the upward adjustment of variance; that is, the required amount of adjustment has little relation to the actual quality of observation. A more rational remedy, originally proposed by Petersen (1973), will be discussed in section 7.

c. The missing consideration of spatial scales

In the spectral analysis of time series, Blackman and Tuckey (1959) state that the aliased data due to poor sampling should be discarded and new measurements be taken under a better experimental design. The possibility of aliasing is equally present in the field analysis of discrete spatial data. During GATE, the on-board observations of every ship clearly registered frequent passages of squall lines in the wind and humidity measurements. The time series of the upper-air data at each ship also showed fairly strong signals associated with those convective-scale disturbances, especially in the lower atmosphere below the 60 kPa level. Although the array of the GATE ships was able to capture the cloud clusters, i.e., the subsynoptic-scale aggregates of convective-scale disturbances, the array was too coarse to determine the spatial structure of the individual disturbances that generated the observed strong signals. Therefore, in the spatial analysis of the GATE ship data, there was a real danger of aliasing the undersampled convective-scale signals to the cloud-cluster scale.

In spite of Blackman and Tuckey's advice, it is obviously unrealistic to hope for another GATE with a greater number of ships. However, we may take heed to its implied contrapositive that it is not possible to de-alias the aliased data without additional information. Thus, the question shifts to the availability of the additional information. As was discussed in the Introduction, our decision is to use the spatial covariances, calculated from the entire dataset (the ensemble) of the GATE Phase III period, in the attempt to de-alias the spatial analyses at individual map times. The meaning of de-aliasing in this attempt, and the degree of success, are discussed theoretically in sections 4 and 5, and the practical procedures are developed in the subsequent sections. In so doing, we find that the traditionally defined error variance (3.2) is not necessarily the best criterion for de-aliased analysis. For example, the theory of optimum estimates by Thiebaux (1973) or Thiebaux and Passi (1976) did not address the question of analyzable spatial scales by a given, and possibly undersampled, set of data.

It is noted that the de-aliased analysis would require spatially filtered covariance functions, even if the true functions were known in every detail. This conclusion is not necessarily new, in the sense that all the operational applications of statistical interpolation are employing smooth covariance functions. For example, the scale length in the Gaussian model of the autocovariance by Lorenc (1981) is set to 500 km, presumably for the best performance of the global prediction model. In the GATE analysis, we are not testing a prediction model but are trying to extract from the data a maximum amount of information that would be compatible with the study of interaction between convective systems and their environment. Therefore, the question is not just about the need for a filter, but, rather, the determination of the filter that is optimum to our goal.

4. Spectral properties of mechanical interpolation

a. Matrices for one-dimensional Fourier transform

The central problem of our inquiry is concerned with the smallest scale or scales that can be meaningfully analyzed when the distances between the observing stations are not arbitrarily small but finite. In this section and the next, we consider an idealized model of the equally spaced stations, so that the problem can be reduced to its essentials. The purpose of this particular section is to introduce Fourier spectral representation in a band-grouped matrix notation.

The analysis domain is assumed to be a one-dimensional cyclic domain of width D, in which the observing stations, $\hat{x}_j$, are placed with a constant separation distance, Δx. Any integer is allowed for j, but $j = 1, 2, \cdots, J$ will be considered as the primary cycle. Thus,

$$\hat{x}_j \equiv j\Delta x, \qquad (4.1)$$

$$D = x_J - x_0 = J\Delta x. \qquad (4.2)$$

The cyclic conditions on the data and the true field are

$$\hat{f}_{j\pm J} = \hat{f}_j \quad \text{and} \quad f(x \pm D) = f(x). \qquad (4.3)$$

Since observational errors in the data only divert our attention from the present question, the data are assumed to be error-free, that is,

$$\hat{f}_j = f(\hat{x}_j). \qquad (4.4)$$

Finally, it is assumed that the ensemble statistics of the true fields are spatially homogeneous and precisely known. In particular, the covariance function (3.5b) is, for every j,

$$m_j(x) = \mu(x'), \quad \text{for} \quad x' = x - \hat{x}_j, \qquad (4.5)$$

where $\mu(x')$ is the autocovariance function and is assumed to be known everywhere.

Under the assumed conditions, any analyzed field can be compared with the corresponding true field in terms of their Fourier spectra in wavenumber space. Since the domain has a finite cycle width, the wavenumbers to be considered are discrete but may range between plus and minus infinity. These wavenumbers will be denoted by $k_{b,n}$, with the two integer indices, b to indicate a band of J wavenumbers and n to specify an individual within the band. In order to center the main band, $b = 0$, at zero wavenumber, we shall assume J to be an odd integer, i.e.,

$$J = 2N + 1. \qquad (4.6)$$

Thus,

$$\left.\begin{aligned} k_{b,n} &= (bJ + n)\Delta k, \\ b &= -\infty, \cdots, -1, 0, 1, \cdots, \infty \\ n &= -N, \cdots, -1, 0, 1, \cdots, N \end{aligned}\right\}, \qquad (4.7)$$

with

$$\Delta k = 1/D, \quad \text{or} \quad J\Delta k\Delta x = 1 \qquad (4.8)$$

The assumption of odd J by (4.6) is only to avoid awkward notation. It is not critical to the rest of the discussion.

For compactness of notation, we abbreviate the complex Fourier function as

$$Q_{b,n}(x) \equiv \exp(-i2\pi k_{b,n}x). \qquad (4.9\text{a})$$

For each band b, there are J independent functions of the above form, with n running from $-N$ to N. These are grouped as a set and denoted by a column matrix of J elements,

$$\{Q_b(x)\} \equiv (Q_{b,n}(x))_{J\times 1}. \qquad (4.9\text{b})$$

We may further define, for each band, a J-by-J matrix, $[\hat{Q}_b]$, composed of discrete values of (4.9b) evaluated at $x = \hat{x}_j$ for $j = 1, \cdots, J$, and also for $n = -N, \cdots, N$. However, because of (4.8), $[\hat{Q}_b]$ of different bands are identical to each other. Therefore, we drop the subscript b and define

$$[\hat{Q}] \equiv (Q_{b,n}(\hat{x}_j))_{J\times J}, \qquad (4.10)$$

for any b on the right-hand side. The rows of $[\hat{Q}]$ are indexed by n, from $-N$ to N (J rows in total), and the columns by j from 1 to J.

Orthogonality relations of the defined Fourier matrices are listed below for convenience:

$$\left.\begin{aligned} &\Delta k \int_0^D \{Q_b(x)\}\{Q_{b'}(x)\}^{\mathrm{H}} dx = \delta_{bb'}[1] \\ &\Delta k \sum_{b=-\infty}^{\infty} \{Q_b(x)\}^{\mathrm{H}}\{Q_b(x')\} = \delta(x - x') \\ &[\hat{Q}][\hat{Q}]^{\mathrm{H}} = [\hat{Q}]^{\mathrm{H}}[\hat{Q}] = J[1] \end{aligned}\right\}, \qquad (4.11)$$

where [1] denotes the identity matrix, $\delta_{bb'}$ is Kronecker's delta, and $\delta(x - x')$ Dirac's delta function. Also in the above, superscript H indicates a transposed, complex conjugate matrix (Hermitian transpose).

b. Fourier representation of the true field

Being defined at every x, the true field can be expressed by an infinite Fourier series. In our band-grouped notation, it is

$$f(x) = \Delta k \sum_{b=-\infty}^{\infty} \{Q_b(x)\}^{\mathrm{H}}\{F_b\}, \qquad (4.12)$$

with

$$\{F_b\} = \int_0^D \{Q_b(x)\} f(x) dx, \qquad (4.13)$$

where $\{F_b\}$ is the column matrix of $F_{b,n}$, the Fourier coefficient for $k_{b,n}$, of band b.

The Fourier spectrum of $f(x)$, that is, $\{F_b\}$ of all the bands, is defined for each member of the ensemble. On the other hand, the statistical properties of the ensemble are related to the power spectral density, or power spectrum for short. For example, the ensemble-averaged variance (3.8), which is, here, spatially constant due to the homogeneity assumption, is given by

$$s^2(x) = s^2 = \Delta k \sum_{b=-\infty}^{\infty} \sum_{n=-N}^{N} P_{b,n}, \qquad (4.14)$$

where

$$P_{b,n} \equiv \Delta k \langle F_{b,n}^{*} F_{b,n} \rangle \qquad (4.15\text{a})$$

is the component of the power spectrum at wavenumber $k_{b,n}$. In the above, superscript $*$ denotes a complex conjugate. We may group J components of (4.15a), from $n = -N$ to N, and represent them for each band by a *diagonal* matrix,

$$[P_b] = (P_{b,n}\delta_{nn'})_{J\times J}. \tag{4.15b}$$

As is well known, the power spectrum is also the Fourier coefficients of the autocovariance function $\mu(x)$. By shifting the origin of $\mu(x)$ to individual stations $\hat{x}_j$, as shown in (4.5), we have the covariance functions expressed in terms of the power spectrum. Namely,

$$\{m(x)\} = \Delta k[\hat{Q}]^{\mathrm{H}} \sum_{b=-\infty}^{\infty} [P_b]\{Q_b(x)\}. \tag{4.16}$$

c. *Aliased analysis*

The true field, expressed by the infinite series (4.12), is only the target of discrete-data analysis. The best analysis obtainable by mechanical interpolation is a finite Fourier series, in the present case of the equally spaced data, $\{\hat{f}\}$, in a cyclic domain. The result is a continuous function of x, but does not contain any more information than exists in the data. Therefore, this analysis result is denoted here by $\hat{f}(x)$ and is given by

$$\hat{f}(x) = \Delta k\{Q_0(x)\}^{\mathrm{H}}\{\hat{F}\}, \tag{4.17}$$

with

$$\{\hat{F}\} = \Delta x[\hat{Q}]\{\hat{f}\}. \tag{4.18}$$

As explained for (4.10), $[\hat{Q}]$ has no band designation; neither does $\{\hat{F}\}$, the column matrix of Fourier coefficients. However, $\{\hat{F}\}$ is normally considered to belong to the main band, $b = 0$. This assumption is already reflected in (4.17) by the choice of $\{Q_0(x)\}$ rather than $\{Q_b(x)\}$ of a side band. (Exceptions may occur in the analysis of certain other datasets, such as the Doppler-radar observation of wind speeds. The problem there is to relocate, or *unfold,* the observed spectral peak to a correct side band, depending on wind conditions.)

Besides the missing side bands, the mechanical analysis (4.17) does not agree with the true field (4.12) even in the main band. To see this, we only need to compare the two at the discrete data points $\{\hat{x}\}$, where, on account of (4.4), both must reproduce $\{\hat{f}\}$. In fact, (4.17) yields

$$\{\hat{f}\} = \Delta k[\hat{Q}]^{\mathrm{H}}\{\hat{F}\},$$

while (4.12) is reduced, because of (4.10), to

$$\{\hat{f}\} = \Delta k[\hat{Q}]^{\mathrm{H}} \sum_{b=-\infty}^{\infty} \{F_b\}.$$

Since $[\hat{Q}]$ is not a singular matrix, it follows that

$$\{\hat{F}\} = \sum_{b=-\infty}^{\infty} \{F_b\}, \tag{4.19}$$

which implies the well-known fact that all of the side bands of the true spectrum are folded onto the one and only main band of the spectrum that is obtained by the mechanical interpolation. In other words, higher wavenumbers in the true side bands are *aliased* to appear like wavenumbers within the main band. We may note that the amplitudes of the true side bands are unchanged, even though their wavenumbers are aliased in (4.19). If the data points are not equally spaced, there will be no such assurance of amplitude preservation. In the mechanical interpolation of irregularly spaced data, therefore, aliasing is practically synonymous with overshooting.

For the convenience of later reference, the analysis (4.17) may be given in the one-sided real form, i.e.,

$$\hat{f}(x) = \sum_{n=1}^{N} \hat{G}_n(x), \tag{4.20}$$

with

$$\hat{G}_n(x) = 2\Delta k(\hat{A}_n \cos 2\pi k_{0,n}x - \hat{B}_n \sin 2\pi k_{0,n}x), \tag{4.21}$$

where $\hat{A}_n$ and $\hat{B}_n$ are the real and imaginary parts of $\hat{F}_n$, respectively. The spatial phase of each wave component $\hat{G}_n(x)$ is determined by the ratio $\hat{A}_n{:}\hat{B}_n$ and is also affected by aliasing.

Now, the statistical properties of the ensemble of $\hat{f}(x)$ can be written in the spectral form. The power spectrum is unique only in one band, and the components for $n = -N$ to N are

$$\hat{P}_n = \Delta k\langle \hat{F}_n^* \hat{F}_n\rangle, \tag{4.22a}$$

which may be set in the diagonal matrix form,

$$[\hat{P}] = (\hat{P}_n\delta_{nn'})_{J\times J}. \tag{4.22b}$$

Then, the sample-covariance matrix (3.4a) becomes

$$[\hat{m}] = \Delta k[\hat{Q}]^{\mathrm{H}}[\hat{P}][\hat{Q}]. \tag{4.23}$$

Because of (4.4), $[\hat{m}]$ can also be formed by arranging, side by side, J columns of (4.16), each at $x = \hat{x}_{j'}$ for $j' = 1, \cdots, J$. The result,

$$[\hat{m}] = k[\hat{Q}]^{\mathrm{H}}\Big(\sum_{b=-\infty}^{\infty} [P_b]\Big)[\hat{Q}],$$

must agree with (4.23). Thus, it follows that

$$[\hat{P}] = \sum_{b=-\infty}^{\infty} [P_b], \tag{4.24}$$

which shows that the power spectrum of the ensemble of $\hat{f}(x)$ is also folded, or aliased.

The ensemble variance of $\hat{f}(x)$ is given by

$$\hat{s}^2(x) = \Delta k \sum_{n=-N}^{N} \hat{P}_n = s^2, \tag{4.25}$$

which is constant in space, and, because of (4.24), is identical to s^2 of the true variance (4.14).

5. Spectral properties of statistical interpolation

a. In toto

In the idealized setting of the preceding section, we have seen that the mechanical interpolation of the discrete data produces no information in higher wavenumbers of the side bands, that the analysis in the main band may be incorrect due to aliasing, and, worst of all, that the method does not tell how severe or light the extent of aliasing might be. In the analysis of real observational data, aliased spatial scales are likely to be accompanied by overshooting amplitudes. A human analyst often has background knowledge of the observed physical event, so that he may recognize unrealistic distortions in the analysis. Removing the suspected aliasing from the analysis is another matter and requires some additional information about the true field.

In the objective analysis by statistical interpolation, the knowledge of covariance functions, or, equivalently, that of the true power spectrum of the ensemble, constitutes the supplied judgmental information. We are now ready to see, again in the idealized setting, how this extra information is used in an attempt to *de-alias* the main band. We shall also see that even a perfect knowledge of the statistics is no help for recovering the undersampled small-scale components in the side bands.

In the one-dimensional spectral form, the statistically optimized analysis (3.6) is represented by an infinite Fourier series,

$$\tilde{f}(x) = \Delta k \sum_{b=-\infty}^{\infty} \{Q_b(x)\}^{\mathrm{H}}\{\tilde{F}_b\}, \qquad (5.1)$$

where $\{\tilde{F}_b\}$ is the set of J Fourier coefficients of $\hat{f}(x)$ in band b, and is given by the Fourier transform of the right-hand side of (3.6),

$$\{\tilde{F}_b\} = \int_0^D \{Q_b(x)\}\{m(x)\}^{\mathrm{T}}[\hat{m}]^{-1}\{\tilde{f}\}dx.$$

Making use of (4.16), (4.18) and (4.23), as well as (4.8) and the orthogonality relations listed in (4.11), we can reduce the above to

$$\{\tilde{F}_b\} = [P_b][\hat{P}]^{-1}\{\hat{F}\}. \qquad (5.2)$$

Also for the present one-dimensional case, the minimized error variance (3.7) can be written as

$$E(x) = s^2 - \tilde{s}^2(x), \qquad (5.3)$$

where the constant s^2 is the ensemble variance of the true fields, (4.14), and $\tilde{s}^2(x)$ is the ensemble variance of the optimum analysis (5.1); specifically,

$$\tilde{s}^2(x) = \Delta k \sum_{b=-\infty}^{\infty} \sum_{b'=-\infty}^{\infty} \{Q_b(x)\}^{\mathrm{H}}[P_b][\hat{P}]^{-1}[P_{b'}]\{Q_{b'}(x)\}. \qquad (5.4)$$

The traditional minimization, as defined in section 3, works on the premise that the analysis would be better if the error at any x were made smaller. As the result, (5.3) is a varying function of x in the domain and normally becomes a minimum at every data point. In the present case of error-free data, we can in fact show, for every j, that

$$\min_x E(x) = E(\hat{x}_j) = 0. \qquad (5.5)$$

Although the above premise for point-by-point minimization seems to be commonly accepted without qualification, the obvious imprints of observing stations in the spatial variation of (5.3) and (5.4) should be viewed as a warning sign of trouble. For a better understanding of the problem, we divide (5.1) into two parts and shall examine them separately. Namely,

$$\tilde{f}(x) = \tilde{f}_0(x) + \tilde{f}'(x), \qquad (5.6)$$

where

$$\tilde{f}_0(x) = \Delta k \{Q_0(x)\}^{\mathrm{H}}\{\tilde{F}_0\}, \qquad (5.7)$$

$$\tilde{f}'(x) = \Delta k \sum_{b=1}^{\infty} (\{Q_b(x)\}^{\mathrm{H}}\{\tilde{F}_b\} + \{Q_{-b}(x)\}^{\mathrm{H}}\{\tilde{F}_{-b}\}). \qquad (5.8)$$

In the above, $\tilde{f}_0(x)$ is the primary part containing the main-band components only, and $\tilde{f}'(x)$ is the remainder that includes all of the side-band contributions.

Technical note. Because of the assumption (3.1), $\hat{P}_n$ vanishes at $n = 0$; in strict terms, $[\hat{P}]$ cannot be inverted as we have written in the above. However, $P_{0,0}$ and $\hat{F}_0$ also vanish for the same reason. Thus, the singularity at $n = 0$ is removable and will not appear in the explicit notation below.

b. The de-aliased main band

The meaning of the main-band analysis (5.7) becomes clear if it is compared with the aliased analysis (4.17) by mechanical interpolation. Since both $[P_0]$ and $[\hat{P}]$ are diagonal matrices, the individual Fourier coefficients in $\{\tilde{F}_0\}$, except for $n = 0$, are

$$\tilde{F}_{0,n} = R_n \hat{F}_n, \qquad (5.9)$$

where

$$R_n = P_{0,n}/\hat{P}_n \leq 1, \qquad (5.10)$$

in which the inequality follows from (4.24). Thus, (5.9) is a statistical attempt to *de-alias* the aliased amplitude $\hat{F}_n$, by the ratio, R_n, of the true power to the aliased power at each main-band wavenumber. We may write (5.7) in the one-sided real form as

$$\tilde{f}_0(x) = \sum_{n=1}^{N} R_n \hat{G}_n(x), \qquad (5.11)$$

which is to be compared with the similar form (4.20) for $\hat{f}(x)$. Since the de-aliasing factor, R_n, is a real number, the correction applies only to the amplitude of the real wave $\hat{G}_n(x)$; the spatial phase error that may already exist in $\hat{G}_n(x)$ stays uncorrected and will be carried over to $\tilde{f}_0(x)$.

The power spectrum for the ensemble of $\tilde{f}_0(x)$ is given by

$$\left.\begin{aligned}\tilde{P}_{0,n} &\equiv \Delta k\langle \tilde{F}^*_{0,n}\tilde{F}_{0,n}\rangle \\ &= R_n{}^2\hat{P}_n = R_n P_{0,n}\end{aligned}\right\}, \tag{5.12}$$

which implies a reduction of the aliased power by the square of factor R_n and also a reduction of the true power in the main band by R_n. Therefore, it appears that statistical interpolation de-aliases the power spectrum too much. Although unfortunate, this is the consequence of the attempt to minimize the analysis error of individual fields by the correction factors that are defined for the ensemble. To adjust the correction factors to individual fields, a far greater amount of judgmental information is required than is assumed by statistical interpolation.

The ensemble variance of (5.7), which represents the spatial distribution of the power, is given by

$$\left.\begin{aligned}\tilde{s}_0{}^2(x) &\equiv \langle \tilde{f}_0(x)^2\rangle \\ &= \tilde{s}_0{}^2 = 2\Delta k\sum_{n=1}^{N}\tilde{P}_{0,n}\end{aligned}\right\} \tag{5.13}$$

and is constant in space. In comparison with the full variance (5.4), the interesting aspect of (5.13) is not that it is exactly constant, but that it does not show imprints of the observing stations.

c. *Utility of the side bands*

We have seen that the analysis $\tilde{f}_0(x)$, containing only the main band, is an improvement over the mechanical analysis $\hat{f}(x)$. Although the wave phases are not corrected, the spectral amplitudes are statistically de-aliased. However, the error variance of $\tilde{f}_0(x)$ must be greater than the theoretical minimum (5.3), since $\tilde{f}_0(x)$ alone is not the optimum analysis. The minimum can be attained only by adding back the side-band contribution $\tilde{f}'(x)$ to the main band. Does this mean that $\tilde{f}_0(x)$, as a *field* analysis, could be further improved by the addition of side bands? The answer is that the side bands are not only useless but potentially harmful to the field analysis. It is not suggested, however, that mechanical interpolation (4.17) which produced no side band would be preferable. Although it sounds paradoxical, the utility of the side bands produced by statistical interpolation resides in their nature of being filterable.

To explain the answer unequivocally, we need an explicitly defined distribution of the true power spectrum in the side bands. It is general enough for the purpose to assume that the power spectrum varies linearly within each side band, or that, for $N \geqslant n \geqslant -N$ and $b \geqslant 1$,

$$P_{b,n} = P_{b,0} + n\Delta P_b, \tag{5.14}$$

where $P_{b,0}$ is the power at the central wavenumber, $k_{b,0}$, of side band b, and ΔP_b is a constant increment of power between two adjacent wavenumbers within the band. For $b \leqslant -1$, it holds that $P_{b,n} = P_{-b,-n}$, due to the inherent symmetry of the power spectrum.

The side-band contribution (5.8) can now be written as

$$\tilde{f}'(x) = 2\sum_{b=1}^{\infty}(C_b(x)\cos 2\pi k_{b,0}x + S_b(x)\sin 2\pi k_{b,0}x), \tag{5.15}$$

where

$$C_b(x) = \sum_{n=1}^{N}(P_{b,0}/\hat{P}_n)\hat{G}_n(x), \tag{5.16}$$

$$S_b(x) = \sum_{n=1}^{N}(n\Delta P_b/\hat{P}_n)\hat{G}_n\left(x+\frac{1}{4}k_{0,n}{}^{-1}\right). \tag{5.17}$$

It is immediately seen that $C_b(x)$, above, is identical to (5.11) except for the different statistical factors reducing the amplitudes of $\hat{G}_n(x)$. In $S_b(x)$, besides the altered amplitudes, the spatial phase of each $\hat{G}_n(x)$ is shifted by a quarter of its own wavelength. Therefore, both $C_b(x)$ and $S_b(x)$ contain oscillations only at the wavenumbers of the main band and represent essentially the same information as does $\tilde{f}_0(x)$. In (5.15), these low-wavenumber oscillations act as modulation factors on the rapidly oscillating sine and cosine at the central wavenumbers of the side bands.

The situation mathematically described in the above is analogous to a simultaneous radio transmission of the same, but muffled and slightly garbled, message over several carrier frequencies. Actually, the total effect of (5.15) is worse than this analogy implies. Due to the fact that

$$\cos 2\pi k_{b,0}\hat{x}_j = 1 \tag{5.18}$$

for all b at any $\hat{x}_j$, the summation of the first term in (5.15) leads to a sympathetic interference (coherent superposition), at and around the observing stations, of redundant messages $C_b(x)$ from all the side bands. The cumulative effect of interference will be more pronounced if $P_{b,0}$ decreases less rapidly with b or, equivalently, if the autocovariance function (4.5) has a narrower peak at zero lag. The effect of the second term in (5.15) is much less significant since the summation tends to cancel out incoherent waves.

It is now clear that the side-band components constituting $\tilde{f}'(x)$ have no useful information in them and that the interference between them creates the illusion of small-scale disturbances trapped near the observing stations. When spatial derivatives are calculated from $\tilde{f}(x)$ that includes $\tilde{f}'(x)$, the erroneous contributions of $\tilde{f}'(x)$ will be exaggerated due to their high wavenumbers. Therefore, there is no good reason to keep $\tilde{f}'(x)$ in the analysis. Since the side bands are distinct from the main band in the wavenumber spectrum, the removal of $\tilde{f}'(x)$ can be achieved by a proper spatial filter.

If the field analysis is desired in higher wavenumbers, the main band must be extended to cover those wavenumbers; it can be done only by making observations at shorter intervals.

We have seen in (5.5) that the total analysis $\tilde{f}(x)$ has no error variance at the observing stations, which implies that the analysis fits exactly to the given data at those points. For this reason, one might still argue for the retention of $\tilde{f}'(x)$ in the final analysis. In view of the fact that the fortuitous interference of the meaningless side bands is solely responsible for those dips of the error variance, the exact fit of an analysis to the data cannot be a virtue of overriding importance. If it were, even the mechanical interpolation $\hat{f}(x)$ would fill the bill.

d. *Summary*

The conclusions of the preceding discussions are now summarized in slightly generalized terms.

A discrete set of observational data normally contains signals of all scales; some scales are sampled sufficiently well to resolve a field, and others are not. The resolvable scales, i.e., the adequately analyzable spatial scales, are determined solely by the geometry of observing stations, $\{\hat{x}\}$. In the case of equally spaced stations, the resolvable scales are in the main band of wavenumbers bounded by the Nyquist frequency of spatial sampling, $(2\Delta x)^{-1}$. An analysis method cannot change the resolvable scales.

An intelligent analysis method analyzes only the resolvable signals, after separating them from the input signals in $\{\hat{f}\}$. The resolvable signals are those that are *judged* to belong to the resolvable scales. Unresolvable residual signals must be discarded. Mechanical interpolation, lacking the necessary intelligence, takes the input signals as if they all belong to the resolvable scales. There is no way to untangle the misinterpretation, once the interpolation has been made.

Statistical interpolation, aided by the statistical knowledge of de-aliasing factors, determines the most probable amplitudes of resolvable wave components and produces an analysis, $\tilde{f}_0(x)$, of the de-aliased resolvable signals. However, the method does not actually discard the unresolvable residual signals, but turns them into a spurious additional field, $\tilde{f}'(x)$, in the side bands. The spurious field is removable by a proper spatial filter and should be so removed. In next section, we modify the minimization principle of statistical interpolation such that it will only produce the analysis of resolvable signals.

Statistical interpolation does not correct the possible phase error even in the de-aliased wave components. In real applications, this may become a serious problem in locating the center of a disturbance. However, it is commonly known that, if the field analysis is being performed sequentially in time, the phase error of propagating disturbances can be alleviated by considering the time continuity of analysis. As we discuss in section 8, the time continuity can be taken into account objectively, by Fourier-transforming the time-sequenced station data into time-frequency bands.

6. Field analysis of resolvable scales

a. *The strategy*

The purpose of this section is to reformulate the principle of statistical interpolation in such a way that the conclusions of the last section can be extended to applications under normal conditions of real data. Specifically, we consider a finite number of observing stations irregularly distributed in a nonperiodic two-dimensional domain of limited extent. The obvious question here is how to define resolvable scales. The exact language of the Fourier transform no longer applies, and any answer must be somewhat empirical.

Our strategy is to determine the optimum level of filtering that would be just enough to remove the residue of unresolvable signals from the analysis, rather than to tackle directly a definition of the resolvable scales. To follow this strategy, we must modify the error variance (3.2) in such a way that the optimization procedure can be directly focused on the analysis of resolvable scales. Then, the optimum filter can be determined by monitoring the spatial distribution of the error variance.

We shall begin by introducing a spatial low-pass filter, $\mathscr{S}$, which operates on field variables in the analysis domain. The filter should have a disposable parameter specifying the limit of scales to be filtered. There are many ways to construct such a filter. In the case of a cyclic domain, a Fourier filter with sharp cutoff at a desired wavenumber may be used. This is the filter we assume when references are made, below, to the idealized case. For a finite domain with general boundary conditions, a filter is designable only with a tapered cutoff. For the matter of terminology, however, the cutoff point of the filter is defined as the wavenumber, or wavelength, of the half-response point that would result if the filter operated in a cyclic domain. In two-dimensional applications, the filter response can be made to vary in direction or with spatial coordinates. However, in the following discussion, we assume an isotropic and homogeneous filter, so that the filter may be referred to by a single cutoff wavenumber, k_c, or an equivalent wavelength, l_c. The filter we used in the GATE data analysis is described in the Appendix.

Using the low-pass filter $\mathscr{S}$ as a field operator, we may define a filtered *true* field, denoted by $\bar{f}(\mathbf{x})$, by

$$\bar{f}(\mathbf{x}) = \mathscr{S}(f(\mathbf{x}); l_c). \qquad (6.1)$$

It is noted that the filter makes no reference to observing stations or data, and also that the filter parameter l_c, at this point, is completely at our disposal. However, the purpose of defining the filtered true field is to allow the analysis of discrete data to aim at a potentially achievable target. As we have seen in the preceding section, the true field itself is not an achievable one.

b. Minimization of the targeted error variance

We shall now introduce a new form of error variance by including a specified target of analysis in its definition. Namely, for the ensemble of any analysis $\tilde{f}(\mathbf{x})$ of general form (2.3), we define

$$\mathscr{E}(\mathbf{x};l_c)=\langle(\tilde{f}(\mathbf{x})-\bar{f}(\mathbf{x}))^2\rangle+\langle(\bar{f}(\mathbf{x})-f(\mathbf{x}))^2\rangle, \quad (6.2)$$

where l_c on the left-hand side indicates the filter parameter that is implicit in $\bar{f}(\mathbf{x})$ on the right-hand side. The first squared term of (6.2) measures the error of the analysis $\tilde{f}(\mathbf{x})$ relative to the specified target $\bar{f}(\mathbf{x})$. However, this term alone will not properly account for the skill of the analysis, since the target itself can be made less demanding by increasing l_c. Thus, the second squared term, representing a penalty for lowering the target by filtering, is necessary to make (6.2) a balanced measure of the analysis error and skill. The new definition (6.2) is a generalization of the original definition (3.2) in the sense that the new one becomes identical to the old as l_c decreases to zero (no filter).

The next step is to minimize (6.2) by choosing the influence functions $\{\psi(\mathbf{x})\}$ while holding the filter parameter constant. Since the second squared term of (6.2) does not depend on $\{\psi(\mathbf{x})\}$, this step goes exactly the same as in section 3, except that $f(\mathbf{x})$ is replaced here by $\bar{f}(\mathbf{x})$. Thus, the new solution $\{\psi(\mathbf{x})\}$ that minimizes (6.2) is obtained by solving

$$[\hat{m}]\{\psi(\mathbf{x})\}=\{\bar{m}(\mathbf{x})\}, \quad (6.3)$$

where

$$\{\bar{m}(\mathbf{x})\}=(\bar{m}_j(\mathbf{x}))_{J\times 1}=\langle\{\hat{f}\}\bar{f}(\mathbf{x})\rangle, \quad (6.4)$$

and $[\hat{m}]$ is the same as (3.4). Since the filter does not operate on discrete data but only on the field, $\{\bar{m}(\mathbf{x})\}$ is obtainable by filtering (3.5), i.e.,

$$\{\bar{m}(\mathbf{x})\}=\mathscr{S}(\{m(\mathbf{x})\};l_c), \quad (6.5)$$

which implies that the elements of $\{\bar{m}(\mathbf{x})\}$ are not the covariance functions of the filtered true field but rather the filtered covariance functions of the true field.

With the solution of (6.3), the statistical analysis, optimized for the target (6.1), is given by

$$\tilde{f}(\mathbf{x})=\{\bar{m}(\mathbf{x})\}^{\mathrm{T}}[\hat{m}]^{-1}\{\hat{f}\}. \quad (6.6)$$

In the above, we could have denoted $\tilde{f}(\mathbf{x})$ with a bar added, since it is also obtainable by spatially filtering the original $\tilde{f}(\mathbf{x})$ defined by (3.6). We have not done so, in order to avoid an overcrowded symbol and also to hint at the fact that, in the actual process of analysis, we do not calculate (3.6) and then apply the filter.

If the cutoff taper of the filter is reasonably sharp, we may assume

$$\langle\bar{f}(\mathbf{x})f(\mathbf{x})\rangle=\langle\bar{f}(\mathbf{x})\bar{f}(\mathbf{x})\rangle. \quad (6.7)$$

Then, the minimum error variance which (6.2) attains with analysis (6.6) is

$$\begin{aligned}E(\mathbf{x};l_c)&=\min\mathscr{E}(\mathbf{x};l_c)\\&=s^2(\mathbf{x})-\{\bar{m}(\mathbf{x})\}^{\mathrm{T}}[\hat{m}]^{-1}\{\bar{m}(\mathbf{x})\},\end{aligned} \quad (6.8)$$

where $s^2(\mathbf{x})$ is the variance of the (unfiltered) true field, (3.8).

In the limiting case of $l_c = 0$ (no filter), (6.8) is identical to (3.7) and the analysis (6.6) will contain all the residual effects of unresolvable signals. In the opposite limiting case of infinite l_c, (6.8) takes the absolute maximum values equal to $s^2(\mathbf{x})$ and the analysis will be a total washout. Our next problem is to determine the optimum value of l_c, so that (6.6) will be targeted for the best analysis of resolvable signals.

c. Search for the optimum filter

For the moment, let us return to the idealized one-dimensional case of the preceding section. If the filter cutoff point is chosen exactly at the Nyquist wavenumber, or at the wavelength $2\Delta x$, the analysis (6.6) will be exactly equal to the main-band analysis $\tilde{f}_0(x)$ defined by (5.7). Therefore, $l_c = 2\Delta x$ is the optimum filter. Then, (6.8) is reduced to

$$\begin{aligned}E(x;l_c=2\Delta x)&=E_{\mathrm{opt}}\\&=s^2-\tilde{s}_0{}^2,\end{aligned} \quad (6.9)$$

which is constant throughout the domain, since both terms on the right-hand side, defined by (4.14) and (5.12), respectively, are independently constant.

If the cutoff wavelength is increased, the filter will begin to affect the main band, removing small-scale components within the band. Since the true variance s^2 is unchanged, the error variance will increase but still remain constant in the domain. Thus,

$$E(x;l_c>2\Delta x)=\text{const}>E_{\mathrm{opt}}.$$

On the other hand, if the cutoff wavelength is less than $2\Delta x$, the error variance will become smaller than (6.9), i.e.,

$$E(x;l_c<2\Delta x)<E_{\mathrm{opt}},$$

but it will not be a spatial constant. It is reduced more at the observing stations than between them, due to the phase interference of the side-band components, which are only partially filtered.

The above discussion clearly suggests an empirical procedure for determining the optimum filter that would remove just the undesirable residues of unresolvable scales and no more. In the idealized case, the optimum filter is the one that makes $E(x; l_c)$ a spatial constant of the least value. Since the procedure does not require an explicit knowledge of the resolvable scales, it can be extended to the general case of irregularly distributed stations in a finite domain. If the true variance itself is not constant over the domain, we may use the normalized error variance,

$$e(\mathbf{x}; l_c) = E(\mathbf{x}; l_c)/s^2(\mathbf{x}), \qquad (6.10)$$

provided that a reasonable estimate of $s^2(\mathbf{x})$ is available.

A more unsettling problem may arise in the general case, if the local density of observing stations varies within the analysis domain. Namely, the rate of approach of the error variance to a constant field may not be uniform. To avoid overfiltering in certain areas, other areas may have to be left underfiltered. This ambiguity is a reflection of the fundamental difficulty in the analysis of irregularly distributed data. The proposed procedure does not solve the problem, but allows us to make an informed compromise. If the disparity within the domain is too great, the use of a spatially variable filter may be considered.

In applying the above procedure to the GATE data analysis, we have found that the flatness of the error variance, either (6.8) or (6.10), is not a very sensitive measure to monitor. Therefore, we have also used another statistical measure, $\epsilon(\mathbf{x}; l_c)$, which is much more sensitive to the unfiltered residues of unresolvable signals. It is formally defined as follows.

Let us consider an imaginary ensemble of normalized pure random data at every station, or an ensemble of $\{\hat{\eta}\}$, such that

$$\langle \{\hat{\eta}\}\{\hat{\eta}\}^{\mathrm{T}} \rangle = [1], \qquad (6.11)$$

where [1] denotes the identity matrix. For each dataset $\{\hat{\eta}\}$, the analysis (6.6) would produce an obviously meaningless analysis,

$$\tilde{\eta}(\mathbf{x}) = \{\tilde{m}(\mathbf{x})\}^{\mathrm{T}}[\hat{m}]^{-1}\{\hat{\eta}\}. \qquad (6.12)$$

Then, $\epsilon(\mathbf{x}; l_c)$ is defined as the ensemble variance of $\tilde{\eta}(\mathbf{x})$. Namely,

$$\begin{aligned}\epsilon(\mathbf{x}; l_c) &= \langle \tilde{\eta}(\mathbf{x})\tilde{\eta}(\mathbf{x}) \rangle \\ &= \{\tilde{m}(\mathbf{x})\}^{\mathrm{T}}[\hat{m}]^{-1}[\hat{m}]^{-1}\{\tilde{m}(\mathbf{x})\}. \qquad (6.13)\end{aligned}$$

The behavior of $\epsilon(\mathbf{x}; l_c)$ with respect to the filter parameter is similar to that of $e(\mathbf{x}; l_c)$, except that when one increases, the other decreases and vice versa. In the idealized one-dimensional case, (6.13) for the optimum filter is given by

$$\epsilon(x; l_c = 2\Delta x) = 2J^{-1} \sum_{n=1}^{N} (P_{0,n}/\hat{P}_n)^2, \qquad (6.14)$$

which is constant throughout the domain. As l_c increases, (6.13) decreases but remains constant in the domain. If l_c decreases below $2\Delta x$, (6.13) generally increases but much more rapidly at the stations than between them. A similar pattern of the behavior with respect to l_c may be expected in the general two-dimensional case. Since unresolvable signals in the real data have little correlation between stations, they are hardly different from the random data that are assumed in the derivation of (6.13). Therefore, in terms of the response to unresolvable signals, the similarity between $e(\mathbf{x}; l_c)$ and $\epsilon(\mathbf{x}; l_c)$ is not surprising; the latter is only more sensitive than the former.

As we have noted earlier, the filter is not applied to the individual analysis $\tilde{f}(\mathbf{x})$. Instead, it is used only for filtering the true covariance functions, $\{m(\mathbf{x})\}$, as shown in (6.6). In reality, however, $\{m(\mathbf{x})\}$ is unknown and even its empirical estimate is difficult to obtain. What we may hope to do with an ensemble of real data is to estimate the filtered version, $\{\tilde{m}(\mathbf{x})\}$. In this sense, the search for an optimum filter becomes entwined with the empirical estimation of the covariance functions.

7. Estimation of the ensemble statistics

a. Variance

We have discussed, thus far, the analysis of resolvable signals as the goal of statistical interpolation, as well as the strategy for achieving it. We shall now turn to the actual problem of estimating the required covariance functions from a given ensemble of observational data. The degree of success in this task depends on the coverage and quality of observations and also on the complexity of the observed natural phenomena. In order to compensate for observational undersampling, additional assumptions such as spatial homogeneity and isotropy of the statistical fields are often used, but their justification depends on the given circumstances. Since the estimation of statistical fields is essentially an empirical procedure, it cannot be described without reference to the actual data; the following presentation reflects certain important decisions we made with GATE data.

As before, $\{\hat{f}\}$ represents a set of J concurrent data at J stations $\{\hat{\mathbf{x}}\}$, and it is assumed that the norm has been subtracted, so that $\langle\{\hat{f}\}\rangle = 0$ is still true. For the ensemble of $\{\hat{f}\}$, the sample-covariance matrix $[\hat{m}]$ can be immediately calculated, i.e., repeating (3.4) for convenience of reference,

$$\left.\begin{aligned}\hat{m}_{jj'} &= \langle \hat{f}_j \hat{f}_{j'} \rangle \\ [\hat{m}] &\equiv (\hat{m}_{jj'})_{J\times J}\end{aligned}\right\}. \qquad (7.1)$$

The diagonal terms, $j = j'$, of (7.1) are the sample variances; they are denoted, as elements and as a diagonal matrix, by

$$\left.\begin{aligned}\hat{s}_j^2 &= \hat{m}_{jj} \\ [\hat{s}^2] &\equiv (\hat{s}_j^2\delta_{jj'})_{J\times J}\end{aligned}\right\}. \qquad (7.2)$$

The positive square root of the variance matrix is also a diagonal matrix, defined by

$$\left.\begin{aligned}\hat{s}_j &= (\hat{s}_j^2)^{1/2} \\ [\hat{s}] &= (\hat{s}_j\delta_{jj'})_{J\times J} \\ [\hat{s}][\hat{s}] &= [\hat{s}^2]\end{aligned}\right\}. \qquad (7.3)$$

The inverse of $[\hat{s}]$, denoted by $[\hat{s}]^{-1}$, is also a diagonal matrix, such that

$$[\hat{s}]^{-1}[\hat{s}] = [\hat{s}][\hat{s}]^{-1} = [1]. \qquad (7.4)$$

The sample variance, its square root and inverse are all calculable directly from the ensemble of $\{\hat{f}\}$. The real challenge begins with the estimation of the true variance as a spatial function. Two problems are involved: how to estimate the station values of true variance and how to avoid, or reduce, aliasing in spatial interpolation of the discrete values. Of the first problem, the so-called random error in observations is least troublesome, if it is the only kind of observational error. In our GATE analysis, it was possible to remove practically all of the random error by preprocessing the data of each ship as a time series. On the other hand, the mixed instrumentation of these ships, differing in calibration and dynamic response, created a substantial but yet unknown magnitude of ship-dependent errors in the calculated $\hat{s}_j^2$, on which equally substantial geographical variations of the presumably true variance were superposed. Although spatial variations of the variance field were of relatively large scales, there was evidence of undersampling, especially in the meridional direction.

The procedure we have used for estimating the true variance field $s^2(\mathbf{x})$, or actually its square root $s(\mathbf{x})$, is symbolically written as

$$s(\mathbf{x}) = \mathscr{S}(\hat{s}_j \text{ for all } j; L_x, L_y), \qquad (7.5)$$

where $\mathscr{S}$ denotes an operator for mechanical interpolation with a spatial low-pass filter; the operand of $\mathscr{S}$ is the discrete data $\hat{s}_j$ at $\hat{\mathbf{x}}_j$, and L_x and L_y are the cutoff wavelengths in east–west and north–south directions, respectively, of the direction-dependent filter. In section 6, we introduced $\mathscr{S}$ simply as a filtering operator, since the interpolation role of the operator is insignificant when the operand is densely defined. Here, in (7.5), the discrete operand must be interpolated in space as input to the filter. Note that we do not use an overbar with $s(\mathbf{x})$ above, because the purpose of (7.5) is not to filter $s(\mathbf{x})$ but to obtain an estimate of $s(\mathbf{x})$ itself. Although the procedure is equally applicable to $\hat{s}_j^2$ to directly estimate $s^2(\mathbf{x})$, we prefer estimating $s(\mathbf{x})$ and, then, squaring it.

The choice of the filter parameters is the most critical part of the procedure (7.5). The goal is to average out the instrumental differences among stations by choosing L_x and L_y large enough, but not so large as to wipe out the geographical variations. Since the latter were primarily in the meridional direction in the GATE area, the use of a directional filter was necessary to achieve the goal. In the end, (7.5) gives a fairly smooth field of $s(\mathbf{x})$, which generally does not agree with $\hat{s}_j$ at $\hat{\mathbf{x}}_j$. The ratio of the two, $s(\hat{\mathbf{x}}_j){:}\hat{s}_j$, provides us with a correction factor which adjusts individual $\hat{f}_j$ for either too high or too low instrumental response at station $\hat{\mathbf{x}}_j$.

b. Covariance

Unlike the slowly varying variance field, the covariance function $m_j(\mathbf{x})$ must carry the information on small-scale components. As was discussed in section 6, what is needed in the field analysis is the filtered covariance function $\bar{m}_j(\mathbf{x})$ containing the information only of resolvable scales. Even then, the estimation of $\bar{m}_j(\mathbf{x})$ by mechanical interpolation of discrete $\hat{m}_{jj'}$ of (7.1) is "as tricky a maneuver as pulling oneself up by his own bootstraps." In order to reduce the possible aliasing in the interpolated $\bar{m}_j(\mathbf{x})$, the set of data $\hat{m}_{jj'}$ for any fixed j should have a coverage dense enough to resolve the smallest resolvable scale. On the other hand, the lack of such coverage is the very reason for the limited resolvable scales.

The circular dilemma can be broken only by a spatial composite of $\hat{m}_{jj'}$ under certain assumptions. One such assumption is spatial homogeneity. However, when the variance significantly varies in space, the assumption may be applied to the covariance only after it is normalized by the variance. In addition to homogeneity, which implies invariance in parallel translation, the usual practice also assumes isotropy, which means invariance in rotation. We did not assume isotropy in GATE analysis, however. If it were assumed, the covariance functions would have lost about 30 percent of the otherwise recoverable information. It is noted that, for the assumed homogeneity to improve the spatial coverage of the composite data, the geometrical pattern of the observing stations should not be self-repetitive in translation, as would happen with regularly spaced grid points. In this regard, the slightly irregular, double hexagonal arrangement of ships during GATE was helpful.

The normalized sample-covariance (correlation) matrix is defined and calculated by

$$\hat{n}_{jj'} = \hat{s}_j^{-1}\hat{m}_{jj'}\hat{s}_{j'}^{-1}$$

or

$$[\hat{n}] = [\hat{s}]^{-1}[\hat{m}][\hat{s}]^{-1}. \qquad (7.6)$$

The next step, under the assumed homogeneity, is to estimate the normalized, filtered autocovariance function $\bar{\nu}(\mathbf{r})$ in the horizontal space of the relative displacement vector $\mathbf{r}$. For this purpose, every element $\hat{n}_{jj'}$ of (7.6) is assigned to a relative position $\hat{\mathbf{r}}_{jj'}$, defined by

$$\hat{\mathbf{r}}_{jj'} = \hat{\mathbf{x}}_{j'} - \hat{\mathbf{x}}_j, \qquad (7.7)$$

in the $\mathbf{r}$-space. At the origin $\mathbf{r} = 0$, the normalized sample-variance $\hat{n}_{jj}$ is unity for every $j = j'$ and includes contributions from both resolvable and unresolvable signals. Thus, even if random observational errors were absent, the diagonal terms of (7.6) must be excluded as data for estimation of $\bar{\nu}(r)$.

Applying the filtered mechanical-interpolation operator in the $\mathbf{r}$-space to the discrete data (7.6), except for those with $j = j'$, we obtain

$$\bar{\nu}(\mathbf{r}) = \mathscr{S}(\hat{n}_{jj'} \text{ for } j \neq j'; l_c), \tag{7.8}$$

where l_c is the filter parameter that defined a potentially achievable target in (6.1), and its value is to be adjusted for the optimum definition of resolvable scales by the empirical procedure described in section 6. When the optimum filter is decided, $\bar{\nu}(0)$ at the origin will indicate the fractional ratio of the resolvable-signal variance to the total input variance.

Returning to the original **x**-space, we obtain the filtered, normalized covariance functions by

$$\left.\begin{aligned} \bar{n}_j(\mathbf{x}) &= \bar{\nu}(\mathbf{x} - \hat{\mathbf{x}}_j) \\ \{\bar{n}(\mathbf{x})\} &= (\bar{n}_j(\mathbf{x}))_{J\times 1} \end{aligned}\right\}. \tag{7.9}$$

The filtered covariance functions, which were theoretically defined by (6.5), are now actually obtained by

$$\{\bar{m}(\mathbf{x})\} = [\hat{s}]\{\bar{n}(\mathbf{x})\}s(\mathbf{x}). \tag{7.10}$$

With (7.10) and the reverse of (7.6), i.e.,

$$[\hat{m}] = [\hat{s}][\hat{n}][\hat{s}], \tag{7.11}$$

we may rewrite (6.6) for the normalized analysis,

$$s(\mathbf{x})^{-1}\tilde{f}(\mathbf{x}) = \{\bar{n}(\mathbf{x})\}^{\mathrm{T}}[\hat{n}]^{-1}[\hat{s}]^{-1}\{\hat{f}\}, \tag{7.12}$$

which clearly shows that the difference between $[\hat{s}]$ and $s(\mathbf{x})$ is working for equalization of the data with respect to the station-dependent variation of instrumental response.

The assumption of homogeneity, especially that after normalization, incurs a certain loss in the degrees of freedom, so that the recovery of unnormalized estimates is not totally unique. For example, $[\hat{s}]$ in both (7.10) and (7.11) may be replaced by the presumably truer variance matrix with $s(\hat{\mathbf{x}}_j)$ as diagonals. It can be an attractive alternative if the cause of errors in $[\hat{s}]$ is not as certain as we have assumed. For this and other reasons, it *does* happen in practice that the actual $[\hat{m}]$ used in analysis does not exactly match the one theoretically defined by (7.1). Then, the mismatch, however slight it seems, may literally ruin the analysis. Prevention of such disastrous consequences is our next topic.

c. Desensitization

As was mentioned in section 3, the hypersensitivity of statistical interpolation is a serious problem. Unless countered by some means, it worsens when more data are available to define a field and, thus, a better analysis should result. The commonly applied remedy for the problem is an extra "observational error variance" added to the diagonal elements of $[\hat{m}]$. However, the additional amount that is necessary for the cure has no relation to actual observational errors at individual stations. In fact, the cause of the problem is not with individual data, but with the spatial pattern these correlated data make as a set.

As we discuss below, the covariance matrix $[\hat{m}]$ defines a statistically probable range of spatial patterns with which every set of data $\{\hat{f}\}$ should conform. Hypersensitive reaction occurs when the analysis tries to interpret a set of data whose spatial pattern does not fall within the probable range. Although a mismatch of this sort may occur with statistically freakish sets of data, it is more often caused by an assumed, or synthesized, $[\hat{m}]$ which erroneously sets up the probability of expected patterns. Therefore, the removal of unrealistic expectations of $[\hat{m}]$ must be the basis of a rational remedy for hypersensitivity.

Since the sample-covariance matrix (7.1) is symmetric, we can define eigenvalues λ_k and corresponding orthonormal eigenmodes $\{\phi_k\}$, $k = 1, 2, \cdots, J$ such that

$$\left.\begin{aligned} [\hat{m}]\{\phi_k\} &= \lambda_k\{\phi_k\} \\ \{\phi_k\}^{\mathrm{T}}\{\phi_{k'}\} &= \delta_{kk'} \end{aligned}\right\}. \tag{7.13}$$

All of the eigenvalues are positive or zero, and we may assume they are numbered in decreasing order of values, i.e., $\lambda_k \geqslant \lambda_{k+1}$. Each eigenmode $\{\phi_k\}$ is a column matrix of J discrete values, ϕ_{jk}, assignable to stations at $\hat{\mathbf{x}}_j$, $j = 1, \cdots, J$, respectively. Since $\{\phi_k\}$ is not a physical vector or a continuous function in space, we have avoided the use of other familiar names, "eigenvector" and "empirical orthogonal function (EOF)."

The eigenmodes, above, form a complete set of orthogonal bases spanning a J-dimensional linear algebraic manifold. Therefore, any set of J data, $\{\hat{f}\}$, can be expanded in terms of the eigenmodes. Namely,

$$\{\hat{f}\} = \sum_{k=1}^{J} \{\phi_k\}\hat{a}_k, \tag{7.14}$$

where $\hat{a}_k$ is the amplitude of the kth mode. For the ensemble of $\{\hat{f}\}$ that has defined $[\hat{m}]$, it follows that

$$\lambda_k = \langle \hat{a}_k{}^2 \rangle. \tag{7.15}$$

The covariance matrix itself can be expanded, such that

$$[\hat{m}] = \sum_{k=1}^{J} \{\phi_k\}\lambda_k\{\phi_k\}^{\mathrm{T}}, \tag{7.16}$$

and, for the variance, we have

$$\operatorname{Trace}[\hat{m}] = \sum_{j=1}^{J} \hat{s}_j^2 = \sum_{k=1}^{J} \lambda_k. \tag{7.17}$$

We can now discuss the sensitivity of statistical interpolation in clear mathematical terms. If no eigenvalue exactly vanishes, the inverse of (7.16) is given by

$$[\hat{m}]^{-1} = \sum_{k=1}^{J} \{\phi_k\}\lambda_k{}^{-1}\{\phi_k\}^{\mathrm{T}}. \tag{7.18}$$

Thus, with (7.14) for $\{\hat{f}\}$, the analysis (6.6) becomes

$$\tilde{f}(\mathbf{x}) = \{\bar{m}(\mathbf{x})\}^{\mathrm{T}} \sum_{k=1}^{J} \{\phi_k\}(\hat{a}_k/\lambda_k). \tag{7.19}$$

If some eigenvalues *do* vanish, say, $\lambda_k = 0$ for $K < k \leqslant J$, the corresponding $\hat{a}_k$ must be zero due to (7.15), and the summation with respect to k in (7.19) should be terminated at K.

Often, in practice, a particular dataset $\{\hat{f}\}$ to be analyzed is not necessarily a member of the ensemble of $\{\hat{f}\}$ by which (7.1) defined $[\hat{m}]$. In a more extreme practice, the ensemble is treated as a mere theoretical concept so that a mathematical model replaces the ensemble-averaging in defining $[\hat{m}]$. In these cases, (7.15) breaks down and there is no statistical constraint on the magnitude of $\hat{a}_k$ with respect to λ_k. Some eigenvalues for higher k may become zero or even negative, without the amplitudes $\hat{a}_k$ vanishing at those k. Since these eigenvalues are meaningless, the corresponding terms in (7.19) should be discarded. Even for the remaining positive eigenvalues, it may happen by chance that some $\hat{a}_k$ that correspond to very small λ_k are not small enough to keep their ratio within proper bounds, resulting in erroneously large contributions to the summation in (7.19). Thus, if unpredictable hypersensitivity of the analysis is to be avoided, those terms with positive but very small λ_k should also be discarded. The proposed exclusion implies a partial loss of the input signals that may have been correct observationally. However, as long as the retained eigenvalues still account for a major part of the total variance (7.17), the loss is statistically insignificant.

To formalize the above procedure, let us define K as the number of eigenmodes to be retained in the analysis, counting in the order of decreasing eigenvalues. The inverse of $[\hat{m}]$ in the K-dimensional subspace is obtained by

$$[\hat{m};K]^{-1} = \sum_{k=1}^{K} \{\phi_k\}\lambda_k^{-1}\{\phi_k\}^{\mathrm{T}}. \qquad (7.20)$$

The desensitized analysis that includes only the first K terms in summation of (7.19) is given by

$$\tilde{f}(\mathbf{x}) = \{\bar{m}(\mathbf{x})\}^{\mathrm{T}}[\hat{m};K]^{-1}\{\hat{f}\}. \qquad (7.21)$$

The value of K, above, has to be chosen empirically. This can be done either by examining the spectrum of eigenvalues or by combining the selection process with the search for the optimum filter l_c as described in section 6. For the latter, it is only necessary to replace $[\hat{m}]^{-1}$ in the definition of the error variance (6.8) or (6.10), as well as in (6.13), by the truncated inverse (7.20). The eigenmodes with smaller eigenvalues are more oscillatory from one station to another. Thus, the excluded amplitudes of those modes would have contributed mainly to the unresolvable part of input signal that was to be filtered out anyway. Therefore, the analysis (7.21) is not very sensitive to a choice of K within a reasonable range.

8. Field analysis of time-sequenced data

a. The advantage

Until now, we have discussed the field analysis by statistical interpolation in which the judgmental information necessary for de-aliasing is derived from the ensemble-averaged spatial covariance functions. Although the ensemble usually consists of spatial datasets at many different observation times, the chronological order of individual sets is actually irrelevant to the ensemble-averaging process. If all the member sets of the ensemble were rearranged in a random chronological order, the spatial covariance functions would remain exactly the same. Thus, it is obvious that the above method of analysis does not use all the information that is available in the ensemble of properly time-sequenced sets.

Given such sets of data, a human analyst certainly consults with past and future data, looking for timewise continuity of propagating disturbances. It is also a common practice to composite several datasets at neighboring observation times by displacing them in space according to an assumed rate of translation. Objective analysis by statistical interpolation may similarly take advantage of the time-sequenced ensemble by introducing time-lagged spatial covariances. In the post-analysis of such data, it is more efficient to Fourier-transform the ensemble in the time-frequency domain and to apply the statistical knowledge of cross-spectra for combined judgments in space and time.

The proposed method will not alter the limit of resolvable spatial scales discussed in section 6. Within the resolvable scales, however, the phase and coherence of propagating disturbances, if such are statistically present in the data, will be included automatically in the optimization of analysis; thus, spatial phase errors of these disturbances are likely to be reduced. Unlike the subjective composite, the statistical method does not assume or impose rigid, kinematic displacements.

b. Separation of frequency bands

For the new method to be effective, the given ensemble of data must have been sampled frequently in time, so that the risk of temporal aliasing is small. Therefore, in the following discussion, we shall assume a well-sampled time series of data at every station $\hat{\mathbf{x}}_j$ and denote it by $\hat{f}_{\mathrm{TOT},j}(t)$ for $j = 1, \cdots, J$ or, in the column-matrix notation, by $\{\hat{f}_{\mathrm{TOT}}(t)\}$, without explicit indication of discreteness in time.

In our application to the GATE data, a time series in the above form was generated for each ship station by mechanical interpolation of the reported upper-air data in the form of time–height cross section. During this process, the raw data were checked and errors were, if possible, corrected. With nominally three-hourly, though often six-hourly, observations, and after several iterations of data-editing and interpolation, the risk of temporal aliasing was judged acceptable. However, the time series was neither periodic nor long enough (20 days in Phase III observation period) for the application of the exact Fourier transform. For the purpose of deriving the ensemble statistics, only three or four frequency ranges could be considered quite independent. Therefore, a band-filtered analog was more practical than the formal Fourier transform. The band-filtering was achieved by subtracting two low-pass filters at consecutive cutoff frequencies. The low-pass filter was applied to the time–height cross section analysis by the method given in the Appendix, specifying the filters independently in time and in height. Since the filter cutoff was tapered, the consecutive frequency bands slightly overlapped each other.

In the frequency-band analog of the *real* Fourier transform, the total input data may be written as

$$\hat{f}_{\mathrm{TOT},j}(t) = \hat{f}_{N,j}(t) + \hat{f}_j(t) \tag{8.1}$$

with

$$\hat{f}_j(t) = \sum_{q>0} \hat{f}_{q,j}(t), \tag{8.2}$$

where $\hat{f}_{N,j}(t)$ represents the norm-defining data in the lowest frequency band ($q = 0$), which includes the time mean, and $\hat{f}_j(t)$ is the sum of deviations $\hat{f}_{q,j}(t)$ in higher frequency bands, designated by index $q = 1, \cdots, Q$. As before, these band-filtered data, grouped for all the J stations, will be denoted by column matrices, $\{\hat{f}_N(t)\}$, $\{\hat{f}(t)\}$ and $\{\hat{f}_q(t)\}$.

Our goal is to obtain the spatial analysis in the form

$$\tilde{f}_{\mathrm{TOT}}(\mathbf{x}, t) = \tilde{f}_N(\mathbf{x}, t) + \tilde{f}(\mathbf{x}, t). \tag{8.3}$$

The time-dependent norm field, $\tilde{f}_N(\mathbf{x}, t)$, will be obtained by mechanical interpolation of $\{\hat{f}_N(t)\}$ in space, while the deviation field, $\tilde{f}(\mathbf{x}, t)$, will be analyzed by statistical interpolation, utilizing $\{\hat{f}_q(t)\}$ for all $q > 0$ as the input data. Note that the time mean of $\{\hat{f}_q(t)\}$ vanishes. We write the *true* field, again as a rhetorical device, in the same form as above,

$$f_{\mathrm{TOT}}(\mathbf{x}, t) = f_N(\mathbf{x}, t) + f(\mathbf{x}, t). \tag{8.4}$$

The actual selection of the frequency bands must be adjusted to the given data. Although the decision is subjective, the range of possible choices is practically limited. For example, the bandwidth of $\{\hat{f}_N(t)\}$ should be narrow enough to allow mechanical interpolation of the norm field; yet it should be wide enough so that even the slowest variation in the remaining $\{\hat{f}(t)\}$ is sufficiently recurrent within the duration of the time series. Similarly, the bandwidth of each $\{\hat{f}_q(t)\}$ should be narrow enough for resolution of the frequency spectrum and yet wide enough to retain statistical significance in the interstation correlation.

The frequency bands we adopted for GATE analysis were the following: the norm band N ($q = 0$) for all variations with periods longer than eight days, including the mean and trend; the E band ($q = 1$) for periods between eight days and two days, centered at the average period (about four days) of the synoptic-scale easterly wave; and the D band ($q = 2$) for periods between two days and twelve hours, centered at the one-day period of apparent diurnal modulation of convection in the GATE area. The remaining variations at higher frequencies (all periods less than twelve hours) were extremely variable among ships and judged to contain almost pure noise and were not used in spatial analysis.

c. Statistical interpolation by analog cross-spectra

Following Wallace and Dickinson (1972), we define the band analog of the *complex* Fourier transform for each band $q > 0$ by

$$\hat{F}_{q,j}(t) \equiv \frac{1}{2}\left(\hat{f}_{q,j}(t) + (i2\pi\omega_q)^{-1}\frac{d}{dt}\hat{f}_{q,j}(t)\right), \tag{8.5a}$$

and for all j by

$$\{\hat{F}_q(t)\} \equiv (\hat{F}_{q,j}(t))_{J\times 1}, \tag{8.5b}$$

where ω_q is a representative frequency of the qth band and is only used for scaling the time derivative to form the imaginary part; all the frequencies of the band are still present in (8.5). For mathematical convenience of the two-sided transform, we extend (8.5) to negative frequencies by $\omega_{-q} = -\omega_q$ so that

$$\{\hat{F}_{-q}(t)\} = \{\hat{F}_q(t)\}^*. \tag{8.6}$$

Thus,

$$\left.\begin{aligned}\{\hat{f}_q(t)\} &= 2\,\mathrm{Re}\{\hat{F}_q(t)\}\\ &= \{\hat{F}_q(t)\} + \{\hat{F}_{-q}(t)\}\end{aligned}\right\}. \tag{8.7}$$

The general linear form (2.3) for analysis is now generalized by associating complex-valued influence functions $\{\Psi_q(\mathbf{x})\}$ with $\{\hat{F}_q(t)\}$ for every $q \neq 0$. Namely,

$$\tilde{f}(\mathbf{x}, t) = \sum_{q\neq 0} \{\Psi_q(x)\}^{\mathrm{H}}\{\hat{F}_q(t)\}, \tag{8.8}$$

where the summation covers all the positive and negative bands, but excludes the norm ($q = 0$). To ensure the analysis to be real-valued, we should have

$$\{\Psi_{-q}(\mathbf{x})\} = \{\Psi_q(\mathbf{x})\}^*. \tag{8.9}$$

To discuss statistical interpolation further, we apply the minimization principle to analyze all the frequency bands together. Although it is mathematically simpler to handle each band separately, our interest is not in analysis of just a few spectral peaks; we must properly account for all the frequencies, including those covered by overlapping band-filters. For clarity of presentation, we shall use the error variance (3.2) of the traditional definition. However, the result, below, can be easily extended to the targeted error variance (6.2), since our discussion of resolvable spatial scales equally applies to every temporal frequency band. Also for clarity, time t, as argument, will be dropped, although the ensemble average is still with respect to t.

Now, with substitution of (8.8) for $\tilde{f}(\mathbf{x})$, (3.2) is to be minimized at every $\mathbf{x}$ by choosing every element of $\{\Psi_q(\mathbf{x})\}$ for every $q \neq 0$. The minimizing solution, $\{\Psi_q(\mathbf{x})\}$, can be obtained by simultaneously solving $2Q$ matrix equations for $q \neq 0$,

$$\sum_{q' \neq 0} [\hat{M}_{qq'}]\{\Psi_{q'}(\mathbf{x})\} = \{M_q(\mathbf{x})\}, \qquad (8.10)$$

where

$$[\hat{M}_{qq'}] = \langle \{\hat{F}_q\}\{\hat{F}_{q'}\}^{\mathrm{H}} \rangle, \qquad (8.11)$$

$$\{M_q(\mathbf{x})\} = \langle \{\hat{F}_q\} f(\mathbf{x}) \rangle. \qquad (8.12)$$

The typical element of (8.11) for $q = q'$, $\hat{M}_{qq,jj'}$, is the frequency-band analog of the cross-spectrum between stations $\hat{\mathbf{x}}_j$ and $\hat{\mathbf{x}}_{j'}$. For $q \neq q'$, the elements of (8.11) should be small but do not necessarily vanish, since two frequency bands partially share the same frequencies in the overlap. Even if q and q' are of opposite signs, there are still some contributions to (8.11) due to nonperiodicity of the band-filtered data. In (8.12), $f(\mathbf{x})$ contains all the frequency bands except the norm, so that the data for empirical estimation of $\{M_q(\mathbf{x})\}$, the *true* cross-spectrum function, are the sum of (8.11) with respect to all $q' \neq 0$.

In the ideal case in which all the interband contributions are negligible, (8.10) is reduced to

$$[\hat{M}_{qq}]\{\Psi_q\} = \{M_q(\mathbf{x})\}. \qquad (8.13)$$

With the solution of the above for every q, the optimum analysis is given by

$$\tilde{f}(\mathbf{x}) = \sum_{q \neq 0} \{M_q(\mathbf{x})\}^{\mathrm{H}} [\hat{M}_{qq}]^{-1} \{\hat{F}_q\}, \qquad (8.14)$$

and the minimized error variance by

$$E(\mathbf{x}) = \langle f(\mathbf{x})^2 \rangle - \sum_{q \neq 0} \{M_q(\mathbf{x})\}^{\mathrm{H}} [\hat{M}_{qq}]^{-1} \{M_q(\mathbf{x})\}. \quad (8.15)$$

In our GATE analysis, we did not depend on this simplification, but solved the full form of (8.10). Although we do not give the explicit expression for $\tilde{f}(\mathbf{x})$ here, there was no computational problem in calculating the analysis with the desensitization procedure explained in section 7. The targeted error variance (6.2) was actually used for simultaneous analysis of both the E and D bands. By monitoring the spatial variation of the error variance, $l_c = 450$ km was chosen as the cutoff wavelength of the optimum spatial filter.

If two or more frequency bands are analyzed simultaneously, one may wonder why it is necessary to separate them. The reason lies in differences in their spatial structures. We may surmise that the low-frequency E-band data are generally associated with large spatial scales, and the high-frequency D-band data with small spatial scales. This does not imply that the D-band data would allow us to analyze smaller spatial scales than the E-band data would. The resolvable spatial scales are determined by the spatial placement of ships and are the same for either frequency band. What we may expect from the time–space scale association is that the D-band data would contain proportionally more unresolvable signals than would the E-band data. Thus, in order to de-alias the resolvable scales as accurately as possible, the required judgmental information must be adjusted to the different spatial structure for each frequency band.

d. Analysis of the norm fields

Since our interest is in analysis of the total field (8.3), the norm field is as important as the deviation. Unfortunately, however, there is no statistical help to de-alias the spatial analysis of the norm field; it must be accomplished primarily by mechanical interpolation of data at discrete stations. Ideally, therefore, the norm-defining data should only represent sufficiently large spatial scales. Since slow variations in time are usually associated with large spatial scales, separation of the norm data by temporal filtering may reduce the risk of spatial aliasing. However, the association is far from being perfect; even the spatial set of the station time means may still be undersampling the norm field. If this is the case, only subjective judgments can help the analysis of the norm, but without a guarantee of success. As we mentioned in section 3, the operational analysis of global data for numerical models does not even attempt to analyze the norm field from the data.

With the GATE upper-air data, we indeed encountered serious difficulties. Although our norm-defining data, $\{\hat{f}_N\}$, included slow variations in time, the sign of spatial undersampling, coupled with ship-dependent instrumental biases, was apparent in the time-mean data. In the analysis of the mean wind field, troubles were most evident in the field of vertical motion. As was noted by many earlier analysts of the data, the mean vertical motion over the ship array, calculated by the usual kinematic method, naturally balanced out to a very small value at the 200 mb level and above. However, when similar calculations were made separately for the eastern and western halves of the array, the mean vertical motions in the two halves resulted in a huge imbalance of opposite signs. The main cause of this problem was in the sampling: the strong latitudinal variation in the mean wind was undersampled by the outer ships, while it was better captured by a greater number of ships along the central meridian of the array.

Once the cause was diagnosed, directional filtering became an obvious choice as a remedy. The actual procedure was similar to (7.5) except that the data were $\{\hat{f}_N\}$. Although the filter wavelengths, L_x and L_y, had to be determined by subjective trial, we monitored the resulting vertical motion field, especially the spatial distribution of imbalance at the top, to decide on the best filter combination. The final values were $L_x = 1800$ km and $L_y = 900$ km, under the boundary condition Type 2 (see Table A1). The stronger zonal filter extended the influence of the better-sampled central meridian to the east and west, while the weaker meridional filter preserved the likely true latitudinal variation of the mean wind. In analysis of the temperature field, the ship-dependent biases in the norm data were too large to be corrected by spatial filtering alone. Therefore, the analyzed norm wind field was used, through the thermal-wind relation, to set horizontal temperature gradients as part of the boundary conditions on the norm temperature field.

9. Analysis of a vector field

a. Scalars versus vectors

The purpose of this section is to discuss the analysis of a vector variable or, to be specific, the horizontal wind. Since the essence of our preceding discussions for a scalar variable equally applies to a vector variable, we actually need to explain only a few differences between them. Before going into technical details, however, it may be prudent to clarify our position with regard to the vector analysis.

The horizontal wind, $\mathbf{u}$, is a vector that can be represented by two scalar components (u, v) in any Cartesian coordinates (x, y) that have been chosen to represent the horizontal space. If a given set of wind data is a sufficiently dense sample of the field, each component may be analyzed separately by mechanical interpolation. This is equivalent to a parallel analysis of two independent scalar variables; it does not matter whether or not the scalars are components of a vector. If we decide on a simultaneous analysis of the two components to take advantage of the possible correlation between them, it can still be formulated as a multivariate analysis of two scalars. Thus, in multivariate statistical interpolation of wind data at J stations, the required covariance matrix will be that of $2J$ scalar components.

However, the wind vector is not merely a set of scalars, but is a physical vector, implying that its scalar components must satisfy a definite rule of coordinate transformation. In other words, the wind vector is one and the same physical entity, while its representation in components varies with the choice of coordinates. Since the sphericity of the earth was ignorable in the GATE analysis, the transformation we are concerned with is the rotation of the (x, y)-coordinates about the vertical axis. In analysis of the wind data, we naturally wish the analysis result to be a physical vector also. In other words, the analysis method itself must be invariant to rotation of the coordinates. In statistical interpolation, then, the covariance between two vectors should be a tensor, and the covariance matrix becomes a matrix of tensors. A tensor, in the present context, is representable by four Cartesian components, and the representation changes with rotation of the coordinates. However, the tensor, like the vector, is an invariant physical entity.

Now, confusion may follow from the fact that, in any chosen coordinates, the representation of the tensor covariance matrix by tensor components is mathematically identical to the multivariate covariance matrix in terms of vector components. Why is it necessary to talk of tensors, especially if we do not intend to rotate the coordinates for curiosity's sake or otherwise? It is not necessary, *unless* we manipulate the covariance matrix for very practical reasons.

A familiar example of such manipulation is simplification of the covariance by the assumption of isotropy, i.e., rotational invariance of the representation. Although this assumption was not used in our analysis, it is definable only in the context of tensor covariance (Buell, 1972). Another example, which is more relevant to us, arises when we attempt to normalize the covariance matrix. "There are a variety of possible approaches" (Wallace and Dickinson, 1972) describes well the problem one faces in normalization of the multivariate covariances. As we shall see, there is only one correct way to normalize the tensor covariances. Therefore, it is not for academic pedantry but for practical necessity that we develop vector-analysis procedures in the tensor-invariant form.

b. Extended notation

The matrix notation of section 2, for ordered sets of scalar data, is extended to similarly ordered sets of vector data. Thus, the concurrent wind data at J stations are denoted by a column matrix of J vector elements,

$$\{\hat{\mathbf{u}}\} \equiv (\hat{\mathbf{u}}_j)_{J\times 1}, \tag{9.1a}$$

where each vector wind $\hat{\mathbf{u}}_j$ at station $\hat{\mathbf{x}}_j$ is considered to be internally represented by a 2-by-1 column matrix of Cartesian components,

$$\hat{\mathbf{u}}_j = \begin{pmatrix} \hat{u}_j \\ \hat{v}_j \end{pmatrix}, \tag{9.1b}$$

so that, for the purpose of performing matrix operations by the conventional rules, $\{\hat{\mathbf{u}}\}$ is a column matrix of $2J$ scalar elements. Thus, the transposed row matrix of (9.1a) is understood to be

$$\{\hat{\mathbf{u}}\}^{\mathrm{T}} \equiv (\hat{\mathbf{u}}_j^{\mathrm{T}})_{1\times J}. \tag{9.2}$$

Corresponding to (3.4) for scalar data, the covariance matrix for an ensemble of vector data is written as a $J \times J$ matrix of tensor covariances,

$$[\hat{\mathbf{m}}] \equiv (\hat{\mathbf{m}}_{jj'})_{J\times J} = \langle \{\hat{\mathbf{u}}\}\{\hat{\mathbf{u}}\}^{\mathrm{T}} \rangle, \tag{9.3a}$$

and each tensor covariance, $\hat{\mathbf{m}}_{jj'}$, is internally represented by a 2-by-2 matrix of Cartesian components,

$$\hat{\mathbf{m}}_{jj'} = \langle \hat{\mathbf{u}}_j \hat{\mathbf{u}}_{j'}{}^{\mathrm{T}} \rangle = \begin{pmatrix} \langle \hat{u}_j \hat{u}_{j'} \rangle & \langle \hat{u}_j \hat{v}_{j'} \rangle \\ \langle \hat{v}_j \hat{u}_{j'} \rangle & \langle \hat{v}_j \hat{v}_{j'} \rangle \end{pmatrix}. \qquad (9.3b)$$

For the rhetorical true wind field, $\mathbf{u}(\mathbf{x})$, the true covariance functions are written as a column matrix of J tensor functions,

$$\{\mathbf{m}(\mathbf{x})\} \equiv (\mathbf{m}_j(\mathbf{x}))_{J\times 1} = \langle \{\hat{\mathbf{u}}\} \mathbf{u}(\mathbf{x})^{\mathrm{T}} \rangle. \qquad (9.4)$$

The tensor notation is similarly extended to the cross-spectra of time-sequenced wind data. Applying consecutive frequency-band filters to both components of the wind in time series, we obtain the band analog of the real Fourier transform, $\{\hat{\mathbf{u}}_q(t)\}$. Then, applying the procedure defined by (8.5a) to both vector components, we obtain the band analog of the complex Fourier transform, $\{\hat{\mathbf{U}}_q(t)\}$, for $q \neq 0$: The matrix of analog cross-spectra is a J-by-J matrix of complex-valued tensors,

$$[\hat{\mathbf{M}}_{qq'}] \equiv (\hat{\mathbf{M}}_{qq',jj'})_{J\times J} = \langle \{\hat{\mathbf{U}}_q\} \{\hat{\mathbf{U}}_{q'}\}^{\mathrm{H}} \rangle, \qquad (9.5a)$$

and each tensor element is

$$\hat{\mathbf{M}}_{qq',jj'} = \langle \hat{\mathbf{U}}_{q,j} \hat{\mathbf{U}}_{q',j'}{}^{\mathrm{H}} \rangle, \qquad (9.5b)$$

which can be internally represented, as in (9.3b), by four complex-valued Cartesian components. (Human perception of the tensor by components is a formidable problem. See comments later in this section.)

The extension of empirical estimation procedures of section 7 to the tensor statistics requires normalization by tensor variances of vector data. Contrary to the multivariate approach, in which the Cartesian components of a variance tensor are arbitrarily dismembered, the correct handling of the variance requires that all the components of the tensor are kept together. Thus, without any ambiguity, the variance matrix for all the stations is

$$[\hat{\mathbf{s}}^2] \equiv (\hat{\mathbf{s}}_j{}^2 \delta_{jj'})_{J\times J}, \qquad (9.6a)$$

and for each station,

$$\hat{\mathbf{s}}_j{}^2 = \hat{\mathbf{m}}_{jj} = \langle \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j{}^{\mathrm{T}} \rangle. \qquad (9.6b)$$

The "square root" of the variance and its inverse are also tensors. These are denoted, respectively, by

$$\begin{aligned} [\hat{\mathbf{s}}] &\equiv (\hat{\mathbf{s}}_j \delta_{jj'})_{J\times J} \\ [\hat{\mathbf{s}}]^{-1} &\equiv (\hat{\mathbf{s}}_j{}^{-1} \delta_{jj'})_{J\times J} \end{aligned} \qquad (9.7)$$

and are uniquely definable by

$$\left. \begin{aligned} [\hat{\mathbf{s}}][\hat{\mathbf{s}}] &= [\hat{\mathbf{s}}^2] \\ [\hat{\mathbf{s}}][\hat{\mathbf{s}}]^{-1} = [\hat{\mathbf{s}}]^{-1}[\hat{\mathbf{s}}] &= [1] \end{aligned} \right\}, \qquad (9.8)$$

where [1] is now the matrix of unit tensors on the diagonal.

In the empirical estimation of cross-spectra of wind data, normalization is made by tensor cospectra (only for $q = q'$), which are defined by

$$[\hat{\mathbf{S}}_q{}^2] \equiv (\hat{\mathbf{S}}_{q,j}{}^2 \delta_{jj'})_{J\times J}, \qquad (9.9a)$$

$$\hat{\mathbf{S}}_{q,j}{}^2 = \hat{\mathbf{M}}_{qq,jj} = \langle \hat{\mathbf{U}}_{q,j} \hat{\mathbf{U}}_{q,j}{}^{\mathrm{H}} \rangle. \qquad (9.9b)$$

The "square root" $[\hat{\mathbf{S}}_q]$ and its inverse $[\hat{\mathbf{S}}_q]^{-1}$ are also definable for these complex-valued tensors in the same way as their real-valued counterparts are by (9.7) and (9.8). A practical algorithm to calculate them is given below.

By definition, tensor (9.9b) is representable, in terms of Cartesian components, by a self-adjoint matrix,

$$\hat{\mathbf{S}}_{q,j}{}^2 = \begin{pmatrix} A & C \\ C^* & B \end{pmatrix}, \qquad (9.10)$$

where A and B are real numbers, while C is generally complex. Also by definition, the matrix is positive-definite, except for the special case mentioned below. Thus, we can generally define two positive real numbers,

$$D = (AB - CC^*)^{1/2},$$
$$X = (A + B + 2D)^{1/2},$$

with which we further define

$$a = \frac{1}{2}\left(X + \frac{A - B}{X}\right),$$
$$b = \frac{1}{2}\left(X - \frac{A - B}{X}\right),$$
$$c = \frac{C}{X},$$

where a and b are positive real numbers and c is complex. Then, the "square root" of (9.10) is a self-adjoint, positive-definite matrix, given by

$$\hat{\mathbf{S}}_{q,j} = \begin{pmatrix} a & c \\ c^* & b \end{pmatrix}, \qquad (9.11)$$

and its inverse by

$$\hat{\mathbf{S}}_{q,j}{}^{-1} = \frac{1}{D} \begin{pmatrix} b & -c \\ -c^* & a \end{pmatrix}. \qquad (9.12)$$

The above algorithm also applies to the real-valued tensors in (9.8), in which C and c are real numbers.

In the special case in which the two components of wind data, $\hat{u}_j$ and $\hat{v}_j$, are perfectly correlated in the entire ensemble, (9.11) is still valid but (9.12) will fail because D vanishes. However, if such exceptional data have to be included in the analysis, the inverse tensor can still be defined in the sense of truncated inverse that was discussed in section 7.

c. *Wind analysis by statistical interpolation*

In order to apply the minimization principle of statistical interpolation to wind analysis, we must properly define the error variance. The fact that two definitions are possible requires clarification. For this purpose, we shall take the real form of the analysis as example. Thus, extending the linear form (2.3) to a vector variable, we write the analyzed vector field as

$$\tilde{\mathbf{u}}(\mathbf{x}) = \{\boldsymbol{\psi}(\mathbf{x})\}^{\mathrm{T}}\{\hat{\mathbf{u}}\}, \tag{9.13}$$

where

$$\{\boldsymbol{\psi}(\mathbf{x})\} \equiv (\boldsymbol{\psi}_j(\mathbf{x}))_{J\times 1}. \tag{9.14a}$$

Since each influence function, $\boldsymbol{\psi}_j(\mathbf{x})$, specifies the contribution of a vector at $\hat{\mathbf{x}}_j$ to another vector at $\mathbf{x}$, it must be a tensor and is representable by four Cartesian components, each of which is a function of $\mathbf{x}$. Namely,

$$\boldsymbol{\psi}_j(\mathbf{x}) = \begin{pmatrix} \psi_{j,uu}(\mathbf{x}) & \psi_{j,uv}(\mathbf{x}) \\ \psi_{j,vu}(\mathbf{x}) & \psi_{j,vv}(\mathbf{x}) \end{pmatrix}. \tag{9.14b}$$

The variance of the analysis error vector, $\tilde{\mathbf{u}}(\mathbf{x}) - \mathbf{u}(\mathbf{x})$, may be defined either by the tensor product of the error vector or by the scalar product. Thus, the traditional error variance (3.2) can be extended, for vector analysis, to either

$$\mathscr{E}(\mathbf{x}) \equiv \langle(\tilde{\mathbf{u}}(\mathbf{x}) - \mathbf{u}(\mathbf{x}))(\tilde{\mathbf{u}}(\mathbf{x}) - \mathbf{u}(\mathbf{x}))^{\mathrm{T}}\rangle \tag{9.15}$$

or

$$\mathscr{E}(\mathbf{x}) \equiv \langle(\tilde{\mathbf{u}}(\mathbf{x}) - \mathbf{u}(\mathbf{x}))^{\mathrm{T}}(\tilde{\mathbf{u}}(\mathbf{x}) - \mathbf{u}(\mathbf{x}))\rangle. \tag{9.16}$$

The statistically optimum $\{\boldsymbol{\psi}(\mathbf{x})\}$ is to be determined by minimizing the error variance at each $\mathbf{x}$ with respect to any possible choice of every component of (9.14b) for all j. However, since (9.15) is a tensor, it cannot be simply "minimized." Instead, we shall require all the Cartesian components of (9.15) to be stationary in the sense of variational calculus. On the other hand, (9.16) is a scalar, i.e., not a scalar component but a tensor of order zero. Thus, it is tensor-invariant and can be minimized in the usual sense. In both cases, however, the results are identical. In other words, the tensor and scalar forms of the error variance are equivalent in deriving the tensor-matrix equation for $\{\boldsymbol{\psi}(\mathbf{x})\}$,

$$[\hat{\mathbf{m}}]\{\boldsymbol{\psi}(\mathbf{x})\} = \{\mathbf{m}(\mathbf{x})\}, \tag{9.17}$$

which defines the statistically optimum analysis.

The above conclusion, which also applies to the targeted error variance of section 6, is important to the resolvable-scale analysis of wind fields. In sections 6 through 8, we have discussed the entire analysis procedure for a scalar variable. The computational aspects of the procedure can easily be rewritten for the wind analysis with the vector–tensor notation of this section. However, the procedure also calls for decisions on the optimum spatial filter, defining the resolvable scales, and on the number of eigenmodes in truncated matrix inversion. Because of the equivalence shown above, the decision-making process can still be monitored by the scalar form of the error variance.

In actual execution of the analysis, all the necessary calculations are made in terms of scalars and scalar components. To a computer, vector–tensor calculations mean nothing but an increase in clerical complexity. Even in empirical estimation of covariances or cross-spectra, our procedure does not depend on human interpretation of the calculated results. Nevertheless, objectivity is not a guarantee of satisfaction. If we should believe in the final wind analysis, we must convince ourselves that the statistical fields we have used contain meaningful information. However, the tensor, especially the complex-valued tensor, defies human perception or understanding when it is represented by Cartesian components.

To meet this human need to watch what the computer does, we have derived from the tensor property of rotational invariance a new way to represent the tensor by pairs of a magnitude and an angle. The covariance tensor requires two such pairs, and the complex-valued cross-spectrum tensor requires four. Each pair is easily and independently interpretable and amenable to direct human perception almost like a vector. This polar representation of the tensor, as we may call it, was an indispensable tool to monitor the progress of our GATE wind analysis and should be useful in other statistical studies of two-dimensional vector fields, as was demonstrated by Shapiro (1986). Although a full description of the polar representation is intended for publication elsewhere, a condensed version is available in Ooyama (1985).

By Helmholtz's theorem (e.g., see Daley, 1985), the two-dimensional vector can be expressed in terms of two scalar functions, the streamfunction and velocity potential. The tensor covariance of vector winds, therefore, may be expressed in terms of the scalar covariances of these functions plus the cross-correlation between them. If the tensor covariance is assumed to be homogeneous and isotropic, the scalar forms may also be simplified as proposed by Daley (1985). Hollingsworth and Lönnberg (1986) have calculated the wind statistics in the scalar forms (actually, the statistics of forecast errors) from the FGGE data without assuming isotropy. In the most general form, the scalar formulation is theoretically equivalent to the tensor. However, neither the streamfunction nor the velocity potential is directly measurable by observations; they are related to the observable winds through differential relations. Thus, the determination of the scalar-form statistics involves spatial integration with appropriate boundary conditions which are often additional assumptions.

Unlike the global data Hollingsworth and Lönnberg (1986) processed, the GATE data were available only in a small tropical region, so that the need for boundary conditions made the scalar formulation very unattractive. As we mentioned earlier, the perceived difficulty of interpretation has been the greatest hindrance to a wider adoption of the tensor formulation for the wind statistics. The representation in terms of the longitudinal and transversal components (Buell, 1972) is widely used for the isotropic part of the tensor. Although the anisotropic part can be represented by the correlation between the longitudinal and transversal components (Buell, 1971), its interpretation requires a considerable mental effort. The proposed polar representation by Ooyama (1985) solves this difficulty by representing the general tensor covariance in terms of independently interpretable, invariant scalar functions, *without* invoking spatial differentiation or integration.

10. Conclusions

The present study was motivated by our desire to analyze the wind fields over the GATE ship-array. Since the derived fields of vorticity, divergence and vertical motion were of vital importance to intended scale-interaction studies, a reliable representation of horizontal scales in the wind analysis was a matter of utmost concern. Although GATE represented an unprecedented concentration of observations over a maritime tropical region, every effort to analyze the upper-air dataset had to face severe problems of undersampling and mixed data quality. Our attempt to overcome the problems was a judgmental objective analysis by statistical interpolation in which the informational basis of necessary judgments was statistically derived from the dataset itself.

It was found that in order to achieve our goal the traditional premise of statistical interpolation had to be reexamined in terms of correctly resolvable spatial scales. The main conclusions of this theoretical inquiry are (i) the resolvable scales are determined by the geometrical distribution of observing stations; (ii) the knowledge of true statistics can improve the analysis of resolvable scales by de-aliasing those signals that belong to resolvable scales, but has no effect on the definition of resolvable scales; (iii) residual effects of unresolvable signals on the analysis are removable by a spatial filter and should be so removed; and (iv) de-aliasing applies only to the wave amplitudes of resolvable scales, and the wave phases in space may still be in error.

On the basis of these conclusions, we have developed objective analysis procedures that are targeted for the best achievable analysis of resolvable scales. The procedures include an adequate estimation of "true" statistical fields from the given ensemble of data, a search for the optimum spatial filter by practical criteria, and a method of desensitizing the analysis to statistically errant data by the truncated inversion of the covariance matrix. In order to reduce the spatial phase error of propagating disturbances, we have taken advantage of the GATE data being time-sequenced; the statistical analysis method has been generalized to interpolate, in space, the timewise Fourier-transformed data in two frequency bands. Since the wind is a physical vector, we have written the entire procedure in the tensor-invariant form. This approach is not only theoretically correct, but decidedly advantageous in very practical terms; it eliminates notorious ambiguities encountered in the multivariate approach, and it also leads to a *humane* method of perceiving statistical tensors.

As for the result of our attempt at the GATE data, the analysis of wind fields, including vorticity, divergence and vertical motion, was completed in 1980 by the author, and the analysis of temperature and relative humidity in 1983 by S. K. Esbensen of Oregon State University. The tabulated results of both analyses are archived at the National Center for Atmospheric Research and are available for use by any interested scientist. The method of access to the datasets and some details of the temperature–humidity analysis procedures are described in an unpublished report by S. K. Esbensen and K. V. Ooyama (1983): "An objective analysis of temperature and relative humidity data over the *B* and *A*/*B* ship arrays during Phase III of GATE." A copy of the report is available from the senior author.

Although the quality of analysis by statistical interpolation may be assessed by the error variance, it is a measure relative to the given data. In objective analysis, no method, however elaborate, can recreate what was missed by the original observations. Only by meteorological interpretation of the analyzed results through diagnostic and prognostic tests can the real worth of both the data and analysis be ascertained. It is regrettable that the author, under altered circumstances, could not personally continue the intended work.

However, the result of the wind analysis, when seen on time-lapse movie film, shows fascinating propagation and evolution of disturbances that have not been detected by other low-resolution analyses. One particularly striking phenomenon is the appearance of vorticity couplets in the outflow layer of several cloud clusters. Quantitative studies of this and related phenomena have been published by Esbensen et al. (1982), Tollerud and Esbensen (1983) and Sui and Yanai (1986). Somewhat unexpected use of the analyzed data has been reported by Krishnamurti et al. (1983).

The present paper also describes a method of filtered mechanical interpolation in the Appendix. It was developed as an integral part of our analysis procedures. Since the method accepts a variety of boundary conditions and applies optional filters up to the boundary, it is suitable for general application, especially when the analysis domain is artificially bounded. By generalizing the method to accept inhomogeneous boundary conditions, one can develop an analysis method on several nested domains, and even a prognostic numerical model, with very clean interface conditions.

Work on hurricane analysis and prediction is in progress on this basis.

Finally, we would like to conclude the paper with a philosophical note. The power of creative inference is the driving force of science, but must be checked and nourished by factual evidence. We saw in GATE a rare opportunity to probe the physical linkage between convective systems and their environment and tried to establish factual bases, at first, by deductive processes only. After having explored every possible avenue to extract "facts" from the observational data, the author cannot hide his empathy with Bernard Trevisan (alchemist, 1406–1490) who uttered with his last breath his conviction: "To make gold, one must start with gold" (quotation from Jaffe, 1976). Nevertheless, it has been, and will be, the task of a meteorological analyst to rectify whatever shortcomings may exist in the available data, by whatever means the end justifies. In this regard, the approach of this paper is not pragmatic enough for general application. As an extreme example of the deductive approach to judgmental analysis, however, it should be of interest to the designers of their own analysis procedures and, it is hoped, to the planners of new field experiments.

Acknowledgments. The author wishes to express a sincere appreciation to Steve Esbensen, his partner in the analysis project, for his unfailing encouragement and valuable contributions to the completion of this manuscript; to Steve Lord, Lloyd Shapiro, and George Soukup for helpful comments on the manuscript; to Jan-Hwa Chu for his enthusiastic participation at several stages of the project; to Ed Zipser for his trust in this work during difficult times at NCAR; and to Stan Rosenthal for his interest and administrative support. The author specially thanks Dave Fulker of NCAR for the initial guidance on use of the cubic B-splines.

APPENDIX

Filtered Mechanical Interpolation of Irregularly Distributed Data in a Finite Domain

1. Representation of a field by cubic splines

A continuous field in a domain can be continuously represented by a linear combination of basis functions. A gridpoint representation is not continuous since the bases are defined only at discrete points. Common examples of continuous bases are *n*th-degree polynomials, Legendre orthogonal polynomials, trigonometric and other harmonic functions, Chebyshev polynomials, Hermite interpolation polynomials, linear and cubic splines. Any of these sets of basis functions may yield a satisfactory representation, if an arbitrarily large number of bases are allowed. In reality, we use only a finite number of bases of a set, since the field to be represented has a limited amount of information. Then, the advantage and disadvantage of any particular set must be weighted against the goals of intended applications.

The goals we set here in the order of priority are (i) spatial uniformity of representation, (ii) flexibility in choosing boundary conditions, (iii) differentiability, and (iv) computational efficiency. The domain we consider is finite not because the atmosphere is so confined, but because the data for analysis are available only in that limited area. In this regard, Legendre and Chebyshev polynomials severely violate the uniformity requirement, since orthogonality of the bases in either case is achieved by a spatially variable weight that favors more detailed representation near the domain boundary than in the interior. The absence of data outside the domain causes another problem; it may adversely affect the analysis inside, unless the latter is controlled by appropriate boundary conditions. The representation by a Fourier series, though perfect in uniformity, allows only the periodic condition. We need a flexible representation that accepts more general boundary conditions of our own choosing.

The basis functions we have adopted are the finite elements of C_2 continuity. Specifically, the element is the cubic B-spline (De Boor, 1972; Lyche and Schumaker, 1973), which is made of four smoothly joined segments of cubic polynomials and identically vanishes outside the four basic intervals. It is defined on the nondimensional coordinate ξ for the basic interval of unity by

$$\Phi(\xi) \equiv \begin{cases} 0, & \text{if} \quad |\xi| \geqslant 2, \\ \frac{1}{4}(2-|\xi|)^3, & \text{if} \quad 2 \geqslant |\xi| \geqslant 1, \\ \frac{1}{4}(2-|\xi|)^3 - (1-|\xi|)^3, & \text{if} \quad 1 \geqslant |\xi| \geqslant 0. \end{cases} \quad \text{(A1)}$$

Let us first consider a one-dimensional domain (x_0, x_M), which is to be divided in M equal intervals of width $\Delta x = (x_M - x_0)/M$. The dividing points, x_m for $m = 0, 1, \cdots, M$, including the end points of the domain, will serve as nodes of cubic segments. Any positive integer, preferably not less than 4, may be taken as M, which defines the desired degrees of representational freedom. For the ease of mathematical notation, two outside points, x_{-1} and x_{M+1}, are also defined as auxiliary nodes, one interval away from the boundary nodes, x_0 and x_M, respectively. Thus, the nodes we consider are at

$$x_m = x_0 + m\Delta x, \quad m = -1, 0, 1, \cdots, M, M+1. \quad \text{(A2)}$$

To each node x_m, we assign a basis function $\phi_m(x)$, derived from (A1) by shifting the origin to x_m and by scaling the basic interval to Δx, i.e.,

$$\phi_m(x) \equiv \Phi((x - x_m)/\Delta x). \quad \text{(A3)}$$

The kth derivative with respect to x for $k = 0, 1, 2$ or 3 will be denoted by

$$\phi_m{}^{(k)}(x) = d^k\phi_m(x)/dx^k. \tag{A4}$$

The third derivative is not defined at the nodes, but still is integrable over the domain. Any function $u(x)$ that is representable in the domain is defined by a linear combination of $\phi_m(x)$,

$$u(x) = \sum_{m=-1}^{M+1} a_m\phi_m(x), \tag{A5}$$

where a_m, $m = -1, 0, \cdots, M, M + 1$, are the amplitudes of the nodal B-spline bases. For the first three nodes, $m = -1, 0, 1$, and similarly for the last three nodes, $m = M - 1, M, M + 1$, the definition (A3) extends a nontrivial part of $\phi_m(x)$ outside the domain. However, it is important to note that the representation (A5) defines $u(x)$ only for the domain, $x_0 \leqslant x \leqslant x_M$. Any interpretation of $u(x)$ outside the domain is invalid. We shall call (A5) the *open* form of representation, in contrast to another form to be discussed in next section.

By the definition of the cubic B-spline (A1), the representation $u(x)$ is continuous and continuously differentiable up to the second order. The third derivative is piecewise continuous with finite discontinuities at the nodes. (Exception: any quadratic polynomial function, including a constant and a straight line, is *exactly* representable over the domain.) By the virtue of equally spaced nodes, the representation is macroscopically uniform in the sense that every nodal interval within the domain shares an identical capability of representing a field. In microscopic scales of Δx or less, spatial uniformity is not maintained. For example, the $2\Delta x$ cosine wave with maxima and minima at alternating nodes is approximately representable, while the $2\Delta x$ sine wave with maxima and minima between the nodes has no representation. If these and still shorter waves need an accurate representation, Δx must be reduced by increasing M.

The extension of the above to a rectangular two-dimensional domain is straightforward. The domain is now bounded by (x_0, x_M) and (y_0, y_N), where N is the number of the basic nodal intervals, Δy, in the y-direction. It is not necessary that Δx and Δy are equal. The nodal basis functions in y, similar to (A3), are defined by

$$\phi_n(y) = \Phi((y - y_n)/\Delta y), \tag{A6}$$

centered at $y_n = y_0 + n\Delta y$, $n = -1, 0, 0, \cdots, N, N + 1$. Similarly to (A4), the kth derivative with respect to y will be denoted by $\phi_n{}^{(k)}(y)$, for $k = 0, 1, 2$ and 3. A representable function $u(x, y)$ in the rectangular domain is defined by a bilinear combination of $\phi_m(x)$ and $\phi_n(y)$,

$$u(x, y) = \sum_{m=-1}^{M+1}\sum_{n=-1}^{N+1} a_{mn}\phi_m(x)\phi_n(y), \tag{A7}$$

where a_{mn} is the amplitude of the bilinear basis function at the node (x_m, y_n). The representation is valid only within the rectangular domain that includes the boundary lines.

2. Homogeneous boundary conditions

We return to the one-dimensional case and demonstrate the boundary conditions at the left end of the domain, $x = x_0$. The conditions at the right end, $x = x_M$, are similar. Since (A5) is valid at the boundary point, u, u_x and u_{xx} are defined at x_0, where the subscript x denotes differentiation with respect to x. Let us generally define a second-order linear differential operator $G[u]$ by

$$G[u] \equiv g_0 u + g_1 u_x + g_2 u_{xx}, \tag{A8}$$

where g_k, $k = 0, 1, 2$, are constant coefficients, not all of which are zero. The general form of the boundary condition we may impose on $u(x)$ is

$$G[u] = \hat{g}, \quad \text{at} \quad x = x_0, \tag{A9}$$

where $\hat{g}$ is a given constant. If $\hat{g} = 0$, the condition is called homogeneous; otherwise, it is inhomogeneous. Although inhomogeneous conditions can be implemented, they were not used in the GATE data analysis. In this paper, therefore, we shall discuss only the homogeneous conditions.

Substituting (A5) for u in (A9) with $\hat{g} = 0$, and noting that $\phi_m(x)$ for $m \geqslant 2$ has no participation at x_0, we have the homogeneous condition expressed in terms of the first three amplitudes,

$$a_{-1} = \beta_0 a_0 + \beta_1 a_1, \tag{A10}$$

where, for $m = 0$ and 1,

$$\beta_m = -(\sum_{k=0}^{2} g_k\phi_m{}^{(k)}(x_0))/(\sum_{k=0}^{2} g_k\phi_{-1}{}^{(k)}(x_0)). \tag{A11}$$

In strict terms of formality, the choice of g_k in (A8) must be restricted so that the denominator in the above would not vanish. However, such a possibility does not occur unless the homogeneous condition attempts to emulate an *outwardly* increasing exponential function, or similarly ill-advised behaviors of $u(x)$, at the boundary. In practice, therefore, we may assume (A11) always defines β_m. Similarly at x_M, though not necessarily for the same coefficients g_k, the homogeneous boundary condition is reduced to

$$a_{M+1} = \beta_M a_M + \beta_{M-1} a_{M-1}. \tag{A12}$$

The coefficients β_m for a common variety of boundary conditions are listed in Table A1. The type 10 condition forces an outward exponential decay of $u(x)$ to zero, with a scale length λ. In the type 21, it is the gradient of $u(x)$ that decays to zero.

When the boundary conditions at both x_0 and x_M are chosen, the representation (A5) with (A10) and (A12) may be written in the *closed* form,

TABLE A1. Common examples of homogeneous boundary conditions. (In the second column, $+\lambda$ for B.C. at x_0, and $-\lambda$ for B.C. at x_M; λ is a positive scale length for the outward exponential decay.)

Type	B.C. at x_0, or at x_M	β_0, β_M	β_1, β_{M-1}
0	$u = 0$	-4	-1
1	$u_x = 0$	0	1
2	$u_{xx} = 0$	2	-1
10	$u = \pm\lambda u_x$	$\dfrac{-4\Delta x}{2\lambda + \Delta x}$	$\dfrac{2\lambda - \Delta x}{2\lambda + \Delta x}$
21	$u_x = \pm\lambda u_{xx}$	$\dfrac{6\lambda}{3\lambda + \Delta x}$	$\dfrac{-3\lambda + \Delta x}{3\lambda + \Delta x}$

$$u(x) = \sum_{m=0}^{M} a_m \psi_m(x), \tag{A13}$$

where $\psi_m(x)$, for $m = 0, 1, \cdots, M$, are the closed basis functions defined by

$$\left.\begin{aligned} &\psi_m(x) = \phi_m(x) + \beta_m \phi_{-1}(x), \\ &\qquad\qquad \text{for} \quad m = 0 \quad \text{and} \quad 1 \\ &\psi_m(x) = \phi_m(x) + \beta_m \phi_{M+1}(x), \\ &\qquad\qquad \text{for} \quad m = M \quad \text{and} \quad M-1 \\ &\psi_m(x) = \phi_m(x), \quad \text{for} \quad m = 2, 3, \cdots, M-2 \end{aligned}\right\} . \tag{A14}$$

For the two-dimensional rectangular domain, boundary conditions must be chosen on all four sides, although the four conditions need not be of the same type. If they are all homogeneous, the sides at y_0 and y_N are closed by introducing $\psi_n(y)$ in the same way as the sides at x_0 and x_M are closed by (A14). Then, the closed form of the bilinear representation is given by

$$u(x, y) = \sum_{m=0}^{M} \sum_{n=0}^{N} a_{mn} \psi_m(x) \psi_n(y). \tag{A15}$$

3. Least-squares fitting with a derivative constraint

Before discussing interpolation of discrete data, we shall first discuss the filtered representation of a given continuous function $\hat{u}(x)$ by the closed form (A13) in the one-dimensional domain (x_0, x_M). In general, $\hat{u}(x)$ is not exactly representable on the cubic spline bases or does not necessarily satisfy the assumed boundary conditions. Therefore, the representation is approximate. We shall define the best approximation, $u(x)$, by minimizing the squared differences between $u(x)$ and $\hat{u}(x)$ over the domain. We may also include in the definition a certain derivative constraint, which will be shown to act on $u(x)$ as a low-pass filter.

The mathematical problem is to determine amplitudes a_m of (A13) such that

$$\int_{x_0}^{x_M} \{(\hat{u}(x) - u(x))^2 + \alpha D_k[u(x)]^2\} dx = \min, \tag{A16}$$

where α is a disposable constant, and D_k is a differential operator which, in the one-dimensional case, may simply be the kth derivative, i.e.,

$$D_k[u] = d^k u / dx^k. \tag{A17}$$

The allowable order of the constraint, k, is 1, 2 or 3. As noted earlier, the constraint is integrable even at $k = 3$. The spectral response of the constraint as a filter will be discussed shortly.

Substituting (A13) for $u(x)$ in (A16), we obtain the equations for a_m as

$$\sum_{m'=0}^{M} (p_{mm'} + \alpha q_{mm'}) a_{m'} = b_m, \qquad \text{for} \quad m = 0, 1, \cdots, M, \tag{A18}$$

where

$$\left.\begin{aligned} b_m &= \int_{x_0}^{x_M} \psi_m(x) \hat{u}(x) dx \\ p_{mm'} &= \int_{x_0}^{x_M} \psi_m(x) \psi_{m'}(x) dx \\ q_{mm'} &= \int_{x_0}^{x_M} \psi_m^{(k)}(x) \psi_{m'}^{(k)}(x) dx \equiv q_{mm'}^{(k)} \end{aligned}\right\} . \tag{A19}$$

The coefficients $p_{mm'}$ and $q_{mm'}$, for all m and m', form $(M + 1) \times (M + 1)$ square matrices P and Q, respectively. Here P is positive definite and Q semidefinite, and both are banded matrices of seven diagonals due to the finite width of the basis functions. Thus, with a recursive formula,

$$a_m = e_m a_{m+1} + e'_m a_{m+2} + e''_m a_{m+3} + f_m, \tag{A20}$$

(A18) can be efficiently solved by calculating e_m, e'_m, e''_m and f_m in a forward sweep with respect to m, and, then, a_m in a backward sweep.

The simplest way to find the effect of the derivative constraint on the representation is to apply the calculus of variations to the integral (A16). Treating $u(x)$ as if it were an analytic function, and ignoring boundary effects for a sufficiently large domain, we obtain the Euler–Lagrange equation of the problem as

$$(-1)^k \alpha (d^{2k} u / dx^{2k}) + u = \hat{u}. \tag{A21}$$

If the given $\hat{u}$ is a trigonometric function of wave length l, the same function will be the solution of (A21) except that its amplitude is reduced by a factor

$$r_\alpha(l) = (1 + (l_c/l)^{2k})^{-1}, \tag{A22}$$

provided that we set

$$\alpha = (l_c/2\pi)^{2k}. \tag{A23}$$

Thus, the kth-order derivative acts as a low-pass filter with a $(2k)$th-degree taper in the spectral response, and l_c is the cutoff wavelength where the amplitude response is a half. Since the actual $u(x)$ by (A13) is not analytic as assumed in the above, the response function (A22) is not correct if l_c is too close to $2\Delta x$ or smaller. However, more precise calculations show that (A22) is a good approximation of the true response for $k = 2$ and 3, if l_c is about equal to, or greater than, $4\Delta x$.

The extension of the above to two-dimensional fitting of $\hat{u}(x, y)$ by $u(x, y)$ of (A15) is, in principle, straightforward, but there are a few variations to consider. One method that is frequently practiced is to execute the two-dimensional fitting as the direct product of two one-dimensional processes, first in x, then in y. However, this method is obviously not suitable to interpolation of irregularly distributed data. Therefore, we shall summarize, below, a fully two-dimensional approach.

The minimization problem is essentially the same as (A16), except that the integration covers the rectangular domain. A major difference occurs in a wider option in choosing the form of derivative constraints. If the coordinates x and y represent physically unrelated variables, such as time and height, the filter in x and that in y may be independently specified. On the other hand, if x and y are the Cartesian coordinates on a horizontal plane, we may wish that no directionality imposed by the filter. Examples of D_k for such an isotropic filter are

$$\left.\begin{aligned} D_1[u] &= \nabla u \\ D_2[u] &= \nabla^2 u \\ D_3[u] &= \nabla(\nabla^2 u) \end{aligned}\right\}, \qquad \text{(A24)}$$

where ∇ is the two-dimensional del-operator. Since D_1 and D_3 are vectors, their squaring should be by the dot product.

The equations for amplitudes a_{mn} can be written as

$$\sum_{m'=0}^{M} (P_{mm'} + \alpha Q_{mm'}) A_{m'} = B_m,$$
$$\text{for} \quad m = 0, 1, \cdots, M, \qquad \text{(A25)}$$

where A_m, as a column matrix of $N + 1$ elements, represents a subset of a_{mn} for $n = 0, 1, \cdots, N$; B_m is similarly a subset of b_{mn}; $P_{mm'}$ is a $(N + 1) \times (N + 1)$ square submatrix of $p_{mnm'n'}$ for $n, n' = 0, 1, \cdots, N$; $Q_{mm'}$ is similarly a submatrix of $q_{mnm'n'}$; and

$$\left.\begin{aligned} b_{mn} &= \int_{x_0}^{x_M} \int_{y_0}^{y_N} \psi_m(x)\psi_n(y)\hat{u}(x, y)dydx \\ p_{mnm'n'} &= p_{mm'}p_{nn'} \end{aligned}\right\}, \qquad \text{(A26)}$$

but for $D_3[u]$, as an example,

$$q_{mnm'n'} = q_{mm'}{}^{(3)}q_{nn'}{}^{(0)} + 3q_{mm'}{}^{(2)}q_{nn'}{}^{(1)}$$
$$+ 3q_{mm'}{}^{(1)}q_{nn'}{}^{(2)} + q_{mm'}{}^{(0)}q_{nn'}{}^{(3)}. \qquad \text{(A27)}$$

In the above, $p_{nn'}$ and $q_{nn'}{}^{(k)}$ are similar to those in (A19) but with $\psi_n(y)$. Note that boundary terms, which may arise during integration of the derivative constraint by parts, are not shown in (A27), although they have been included in our computer program. The solution of (A25) is obtainable through a recursive formula similar to (A20) but in terms of A_m and with $(N + 1) \times (N + 1)$ matrix coefficients.

4. Filtered interpolation of discrete data

The problem of interpolating a set of discrete data is almost identical to that of representing a given continuous function, if the data coverage over the domain is reasonably dense (e.g., a few data points in each nodal cell). In fact, the integrals defined in the previous section are actually calculated by summation of discrete values. In the case of sparse coverage, in which many nodal cells are found devoid of data, the present method is still mathematically well defined, with minor exceptions explained below. The real problem with sparse data, the paucity of information, has been discussed in the main text of this paper. The purpose of this section is to summarize the mathematical aspect of interpolation.

Dealing directly with the two-dimensional case, we assume a given set of discrete data $\hat{u}_j$ at $\hat{\mathbf{x}}_j = (\hat{x}_j, \hat{y}_j)$, respectively, for $j = 1, 2, \cdots, J$. All the data points must be within the domain or on the boundary, but their indexing order is immaterial. Some data may even be collocated. The boundary conditions for the closed representation (A15), the form of a derivative constraint, and the filter cutoff wavelength, l_c, must be decided. Then, the minimization problem for the discrete data is defined by

$$\sum_{j=1}^{J} (\hat{u}_j - u(\hat{x}_j, \hat{y}_j))^2 w_j \Delta x \Delta y$$
$$+ \alpha \int_{x_0}^{x_M} \int_{y_0}^{y_N} D_k[u]^2 dydx = \min, \qquad \text{(A28)}$$

where w_j is an optional weight factor assigned to each data point $\hat{\mathbf{x}}_j$. Although we still call the present method of interpolation *mechanical*, the boundary conditions, the filter and the weight factors are judgmental information to be supplied by the analyst; further comments on this point are found at the end of this section.

With substitution of (A15) for $u(x, y)$, the minimizing solution in terms of A_m is obtainable from the same form of equations as (A25), except for the changes noted below. Since the derivative constraint is applied not to the given data but to the represented field $u(x, y)$, there is no need to change the integral definition of $q_{mm'}{}^{(k)}$ that leads to the definition, such as (A27), of $q_{mnm'n'}$ and eventually to $Q_{mm'}$. On the other hand, the elements that constitute B_m and $P_{mm'}$ must be defined by summation Thus, in lieu of (A26), we have

$$\left.\begin{aligned} b_{mn} &= \sum_{j=1}^{J} \psi_m(\hat{x}_j)\psi_n(\hat{y}_j)\hat{u}_j w_j \Delta x \Delta y \\ p_{mnm'n'} &= \sum_{j=1}^{J} \psi_m(\hat{x}_j)\psi_{m'}(\hat{x}_j)\psi_n(\hat{y}_j)\psi_{n'}(\hat{y}_j) w_j \Delta x \Delta y \end{aligned}\right\}. \quad \text{(A29)}$$

It is important that the same w_j be used in the both definitions, above.

The recursive algorithm of the preceding section can be applied to solve (A25) with the newly defined matrix coefficients (A29). The algorithm is stable and accurate in a wide range of applications, with certain, obviously recognizable, exceptions. To explain these exceptions, we write (A25) in a more abstract form

$$(P + \alpha Q)A = B, \qquad \text{(A30)}$$

where A and B are super column matrices whose elements are column matrices A_m and B_m, respectively, and P and Q are super square matrices whose elements are matrices $P_{mm'}$ and $Q_{mm'}$, respectively. The solution A of (A30) is uniquely determinable, if the null space of P and that of Q do not intersect each other. In the case of dense data, P may be positive definite (with an empty null space). Then, (A30) is solvable even with $\alpha = 0$ (no filter).

In the case of sparse data, in which P is likely to be singular, α must be positive however small it may be. The null space of P for irregularly distributed data is hard to define in general terms. On the other hand, the null space of Q is independent of the data and has a simple structure that is easily recognizable by inspection of its definition. For example, let us consider the case in which the boundary conditions on all the four sides are of the type 2 (Table A1) and the derivative constraint is D_3 of (A24). Then, the null space of Q contains only those A that define $u(x, y)$ to be a linear surface in terms of x and y. On the other hand, if the given set of $\hat{\mathbf{x}}_j$ contains at least three points that are not on a straight line, a linear surface does not belong to the null space of P. Thus, regardless of the size of M or N for the representational purpose, all that is required by (A30) for uniqueness are three non-colinear data points. If the boundary conditions are of either type 1 or type 21, only one data point is necessary. Without further examples, we may conclude that the unique solvability of (A30) is practically assured to any number and distribution of discrete data that are worth analyzing.

Unfortunately, the mathematical niceties have no relation to the question of whether or not the resulting field of mechanical interpolation would produce an acceptable analysis. If the information in the given data is not sufficient to produce an acceptable result, expectations of a human analyst must be codified and put into the analysis as additional information. The objective method of combining statistical expectations with the observational data has been discussed earlier in this paper. The present method of mechanical interpolation also allows some degree of control over the result, by adjusting optional features according to the analyst's heuristic judgment.

Of the optional features, the choice of weight factors, w_j, is not a sensitive or effective means of affecting the interpolation result; it is not worth the effort to mull over the precise value of w_j at every data point. If the data coverage is reasonably dense but uneven, w_j may be chosen to be inversely proportional to the local density of the data points. If the data coverage is sparse and most nodal cells have only one data or none, the derivative constraint takes an assertive control over the result; all w_j may simply be set to unity.

The choice of boundary conditions, however, can be very important to the analysis in a finite domain. In meteorological applications, the boundaries of the domain are often drawn in the middle of a continuing physical space for artificial reasons. The boundary conditions, then, are a substitute for the absence of hard data outside the domain and must be set by the analyst according to his best knowledge of physical circumstances. When the type 2 condition is assumed on two adjacent sides of the rectangular domain, it is advisable to impose an additional condition, $u_{xy} = 0$, at the corner, in order to prevent the "dog-eared" appearance of the interpolated field near the corner. This corner condition can be added to (A28) with a Lagrange multiplier.

If the data coverage is reasonably dense, the derivative constraint can be used as a spectral filter. The order of the constraint gives a choice of steepness of the cutoff taper, and the coefficient, through (A23), specifies the cutoff wavelength. If a filter with a steeper cutoff than the sixth degree is desired, a Fourier filter or other commonly available digital filter may be used on the interpolated field. However, the present method has a decided advantage over others in application to the nonperiodic finite domain. Combined with appropriate boundary conditions, the derivative constraint is uniformly applied over the domain up to the boundary, so that there is no need for special treatment near the boundary or for the removal of the trend in advance.

If the data are sparse, the derivative constraint may be used for controlling the overshooting that often plagues the interpolation of irregularly distributed data. As for aliasing, no method can correct it without additional information on the structure of the true field.

REFERENCES

Baker, E. H., and T. E. Rosmond, 1985: Short course on objective data analysis and vector computers. *Bull. Amer. Meteor. Soc.*, **66**, 1153–1161.

Blackman, R. B., and J. W. Tukey, 1958: *The Measurement of Power Spectra.* Dover, 190 pp.

Buell, C. E., 1971: Two-point wind correlations on an isobaric surface in a nonhomogeneous non-isotropic atmosphere. *J. Appl. Meteor.*, **10**, 1266–1274.

——, 1972: Correlation functions for wind and geopotential on isobaric surfaces. *J. Appl. Meteor.*, **11**, 51–59.

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

Daley, R., 1985: The analysis of synoptic scale divergence by a statistical interpolation procedure. *Mon. Wea. Rev.*, **113**, 1066–1079.

De Boor, C., 1972: On calculating with B-splines. *J. Approximation Theory*, **6**, 50–62.

Eddy, A., 1967: The statistical objective analysis of scalar data fields. *J. Appl. Meteor.*, **6**, 597–609.

Esbensen, S. K., E. I. Tollerud and J.-H. Chu, 1982: Cloud-cluster-scale circulations and the vorticity budget of synoptic-scale waves over the eastern Atlantic intertropical convergence zone. *Mon. Wea. Rev.*, **110**, 1677–1692.

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields.* Gidrometeor. Izdat., Leningrad. [Translated from Russian, Israel Program for Scientific Translations, Jerusalem 1965, 242 pp.]

Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A**, 111–136.

Houze, R. A., Jr., and A. K. Betts, 1981: Convection in GATE. *Rev. Geophys. Space Phys.*, **19**, 541–576.

Jaffe, B., 1976: *Crucibles: The Story of Chemistry.* fourth ed., Dover, 368 pp.

Krishnamurti, T. N., S. V. Low-Nam and R. Pasch, 1983: Cumulus parameterization and rainfall rates II. *Mon. Wea. Rev.*, **111**, 815–828.

Lorenc, A. C., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.*, **109**, 701–721.

Lyche, T., and L. L. Schumaker, 1973: Computation of smoothing and interpolating natural splines via local bases. *SIAM J. Numer. Anal.*, **10**, 1027–1038.

Ooyama, K. V., 1985: The polar representation of tensor cross-spectra of winds. *Ext. Abstracts Vol., 16th Conf. on Hurricanes and Tropical Meteorology*, Houston, Amer. Meteor. Soc., 182–183.

Petersen, D. P., 1973: Static and dynamic constraints on the estimation of space–time covariance and wavenumber-frequency spectral fields. *J. Atmos. Sci.*, **30**, 1252–1266.

Rutherford, I. D., 1972: Data assimilation by statistical interpolation of forecast error fields. *J. Atmos. Sci.*, **29**, 809–815.

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

Shapiro, L. J., 1986: The three-dimensional structure of synoptic-scale disturbances over the tropical Atlantic. *Mon. Wea. Rev.*, **114**, 1876–1891.

Sui, C.-H., and M. Yanai, 1986: Cumulus ensemble effects on the large-scale vorticity and momentum fields of GATE. Part I: Observational evidence. *J. Atmos. Sci.*, **43**, 1618–1642.

Thiebaux, H. J., 1973: Maximally stable estimation of meteorological parameters at grid points. *J. Atmos. Sci.*, **30**, 1710–1714.

——, and R. M. Passi, 1976: Linear combinations of dependent meteorological estimates: A synopsis. *Mon. Wea. Rev.*, **104**, 1172–1174.

Thompson, R. M., Jr., S. W. Payne, E. E. Recker and R. J. Reed, 1979: Structure and properties of synoptic-scale wave disturbances in the intertropical convergence zone of the eastern Atlantic. *J. Atmos. Sci.*, **36**, 53–72.

Tollerud, E. I., and S. K. Esbensen, 1983: An observational study of the upper tropospheric vorticity fields in GATE cloud clusters. *Mon. Wea. Rev.*, **111**, 2161–2175.

Wallace, J. M., and R. E. Dickinson, 1972: Empirical orthogonal representation of time series in the frequency domain. Part I: Theoretical considerations. *J. Appl. Meteor.*, **11**, 887–892.