INVERSE METHODS AND RESULTS FROM THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT

by

BRUCE DOUGLAS CORNUELLE

B.A., Pomona College
(1978)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

April, 1983

Signature of author.....
Joint Program in Oceanography, Massachusetts Institute of Technology - Woods Hole Oceanographic Institution, April 1983

Certified by .....
Thesis supervisor.

Accepted by .....
Chairman, Joint Committee for Physical Oceanography.
Massachusetts Institute of Technology - Woods Hole Oceanographic Institution.

INVERSE METHODS AND RESULTS FROM THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT

by

BRUCE DOUGLAS CORNUELLE

Submitted to the Massachusetts Institute of Technology -- Woods Hole Oceanographic Institution Joint Program in Oceanography on April 7, 1983 in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

ABSTRACT

Ocean acoustic tomography was proposed in 1978 by Munk and Wunsch as a possible technique for monitoring the evolution of temperature, density, and current fields over large regions. In 1981, the Ocean Tomography Group deployed four 224 Hz acoustic sources and five receivers in an array which fit within a box 300 km. on a side centered on 26°N, 70°W (southwest of Bermuda). The experiment was intended both to demonstrate the practicality of tomography as an observation tool and to extend the understanding of mesoscale evolution in the low-energy region far from the strong Gulf Stream recirculation.

The propagation of 224 Hz sound energy in the ocean can be described as a set of rays travelling from source to receiver, with each ray taking a different path through the ocean in a vertical plane connecting the source and receiver. The sources transmitted a phase-coded signal which was processed at the receiver to produce a pulse at the time of arrival of the signal. Rays can be distinguished by their different pulse travel times, and these travel times change in response to variations in sound speed and current in the ocean through which the rays passed.

In order to reconstruct the ocean variations from the observed travel time changes, it is necessary to specify models for both the variations and their effect on the travel times. The dependence of travel time on the oceanic sound speed and current fields can be calculated using ray paths traced by computer. The vertical structure of the sound speed and current fields in the ocean were modelled as a combination of Empirical Orthogonal Functions (EOFs) from MODE. The horizontal structure was continuous, but was constrained to have a gaussian covariance with a 100 km. e-folding scale. The resulting estimator closely

resembles objective mapping as used in meteorology and physical oceanography. The tomographic system has at present only been used to estimate sound speed structure for comparison with the traditional measurements, especially the first two NOAA CTD surveys, but the method provides means for estimating density, temperature or velocity fields, and these will be produced in the future.

The sound speed estimates made using the tomographic system match the traditional measurements to within the associated arror bars, and there are several possibilities for improving the signal to noise ratio of the data. Given high-precision data, tomographic systems can resolve ocean structures at small scales, such as in the Gulf Stream, or at large scales, over entire ocean basins. Work is in progress to evaluate the usefulness of tomography as an observation tool in these applications.

Thesis Supervisor: Dr. Carl Wunsch
Cecil and Ida Green Professor of
Physical Oceanography, Massachusetts
Institute of Technology, Cambridge, MA.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND HISTORICAL SKETCH

1.1 INTRODUCTION

One of the principle difficulties plaguing physical oceanographers is the shortage of ocean data. The oceans are large, and the important processes have scales of tens to hundreds to thousands of kilometers (Richman, Wunsch, and Hogg (1977)). The two major means of observation are ship-borne measurement systems such as the Conductivity-Temperature-Depth probe (CTD) which records temperature (T) and salinity (S) as a function of depth during lowerings from a stationary ship, and moored instruments, such as current meters and temperature-pressure (T-P) recorders which are deployed along cables stretched between an anchor on the bottom and buoyant floats at or below the sea surface. CTD lowerings require upwards of 3 hours, but produce extremely detailed records permitting small-scale resolution of the vertical T and S structures. Moored instruments can sample rapidly in time, and their vertical resolution is only limited by the spacing between sensors, although usually no more than about 10 instruments are placed on a 5000 meter mooring. Each mooring or CTD cast samples at a single horizontal (x,y) location, so that area coverage is limited by the expense of moorings or by ship steaming time.

With the increasing sophistication of ocean models, the need for data has become much greater than during the early exploration period when the large-scale structures of the oceans were being defined. The early exploration cruises pictured the ocean as having steady, large-scale, surface current systems with a rapid decrease in strength with increasing depth. The deep ocean was thought to be nearly at rest, with a few very large, slow currents. Once the major current systems had been mapped, interest shifted from exploration to understanding the mechanisms which controlled the observed features. The more data oceanographers took, the more complicated the pictures became, and the simplicity of the large-scale steady currents was replaced by a complex of interacting and intermittent motions, no less varied than the weather in the atmosphere.

When moorings carrying current meters became available, much of the ocean kinetic energy was found to reside in "mesoscale" motions, with horizontal scales of order 100 km. (O(100 km.)), and time scales of O(50 days) (Richman, Wunsch, and Hogg, 1977). The dynamics of these motions are analogous to those of weather in the atmosphere. Oceanographers now face the same problems that meteorologists have been struggling with--obtaining adequate sampling in space and time to resolve the mesoscale motions, i.e. a "synoptic" data set.

Meteorological data systems now include satellites in a global network of pressure and radiosonde measurements, but the oceanographic observation systems have not kept pace. The oceans are opaque to electromagnetic radiation, so that satellite measurements cannot observe beyond the sea surface, and the open ocean is an extremely inhospitable environment for instruments, so that mechanically complicated systems present tremendous engineering difficulties. Munk and Wunsch suggested a solution to the data-acquisition problem (Munk and Wunsch, 1979) (called MW in the following) with a proposal to monitor the oceans using remote sensing by sound energy. They called the technique "Ocean Acoustic Tomography" because of its similarity to medical tomography (Swindell and Barrett (1977)) which uses X-rays transmitted along many paths through a patient to reconstruct a 2 or 3 dimensional picture of the region through which they passed. Low frequency sound transmitted from a source to a receiver moored at depth in the ocean propagates along distinct ray paths as well, and Munk and Wunsch proposed to use the travel times for pulses following different ray paths to infer the structure of the intervening ocean.

1.2 BRIEF HISTORY

The tomography proposal built on an existing body of work on ocean acoustics, bringing together a number of ideas and techniques which had been developed for other applications. The possibility of long-range transmission of low-frequency sound in the ocean had been known since the 1940's, and a scheme for locating downed fliers by triangulating on the sound from TNT charges had been proposed (Ewing and Worzel 1948). Porter, Spindel, and Jaffee (1973) developed a mooring tracking system which used the travel times of acoustic transmissions to monitor the motion of a mooring. By 1977, low-frequency sound transmissions were being used to track neutrally bouyant "SOFAR" floats over long distances (Webb (1977), Spindel, Porter, and Webb (1977), or see Baker (1981)). Steinberg and Birdsall (1966) transmitted continuous wave (CW) sound across the Florida straits using a 406 Hz sound source, and a later experiment transmitted CW sound over 1250 km. (Clark and Kronengold, 1974). The early transmission experiments were mounted to study the intensity of sound transmitted over long distances, while the phase structure was found to be very unstable, due in part to internal wave variations.

Sound speed in the ocean is most sensitive to temperature and pressure effects, and decreasing temperature with depth produces a decrease of sound speed with depth in the upper ocean (in most areas) while the increasing pressure eventually more than balances this effect, resulting in a sound speed minimum at about 1 km. depth in the North Atlantic. (Figure 1.1). The acoustic waveguide is called the SOFAR channel, which tends to refract sound energy toward the axis. This waveguide, coupled with the fact that mechanical absorbtion decreases with decreasing frequency, permits long-range sound transmissions using sources with finite energy. Sound transmitted from a source to a receiver can be described theoretically as a set of "rays" (by analogy with light rays in optics) each of which follows a different path (Figure 1.2). A single pulse leaving the transmitter will be received as a set of "image" pulses, one for each distinct ray (Figure 1.3). The travel time for a given pulse depends on the length of the path it took and the sound speed along that path. These travel times can be computed, given the path and the sound speed profile, by solving the so-called "forward problem". The solution of the forward problem describes the dependence of the pulse travel time along a particular path, Γ_i, on the sound speed field of the ocean, $C(\underline{x},t)$.

FIGURE 1.1 AVERAGED SOUND SPEED PROFILE FOR N. ATLANTIC SOUTHWEST OF BERMUDA

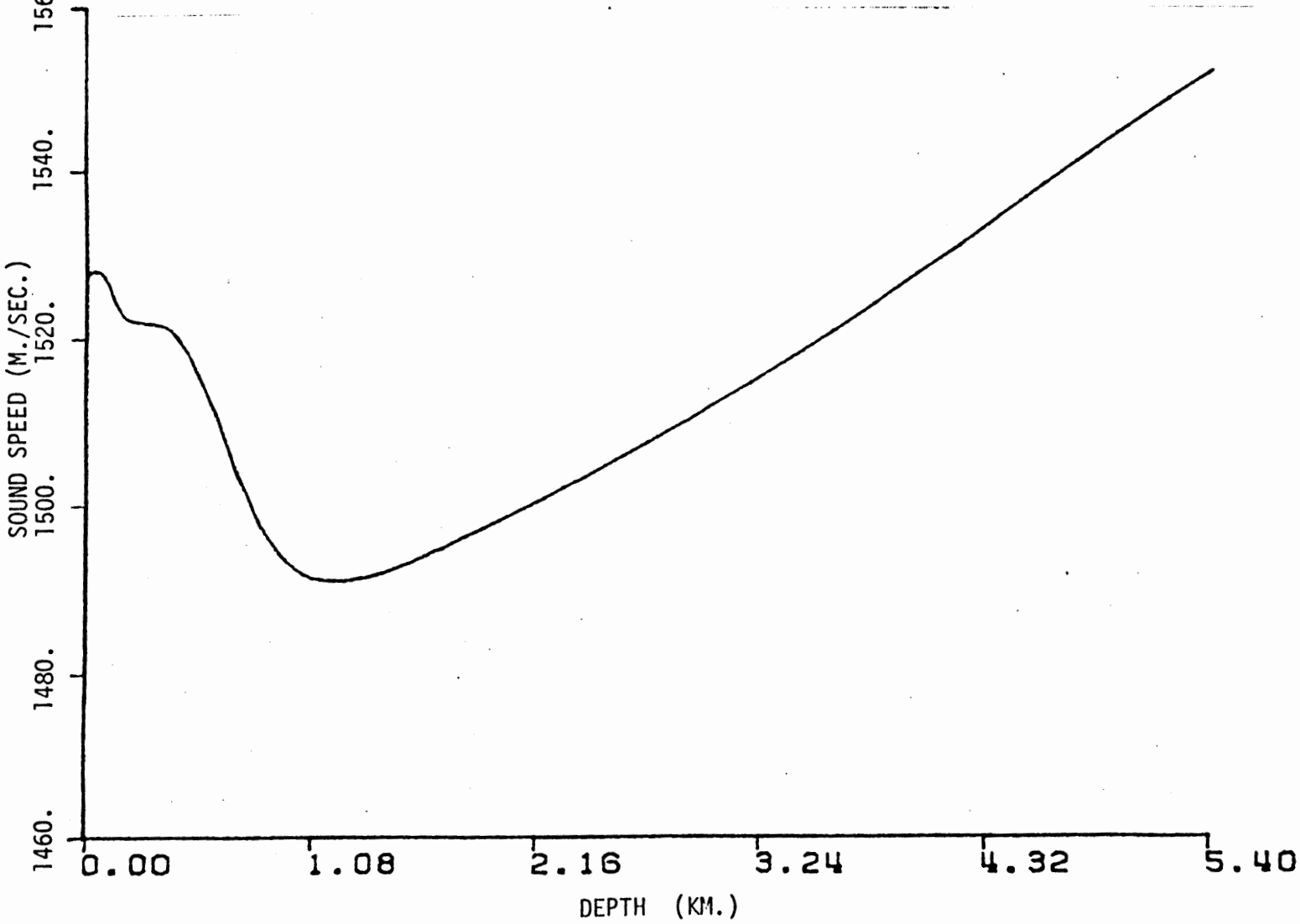


FIGURE 1.2 TWO NUMERICALLY TRACED ACOUSTIC RAY PATHS FOR THE PROFILE SHOWN AS FIGURE 1.1

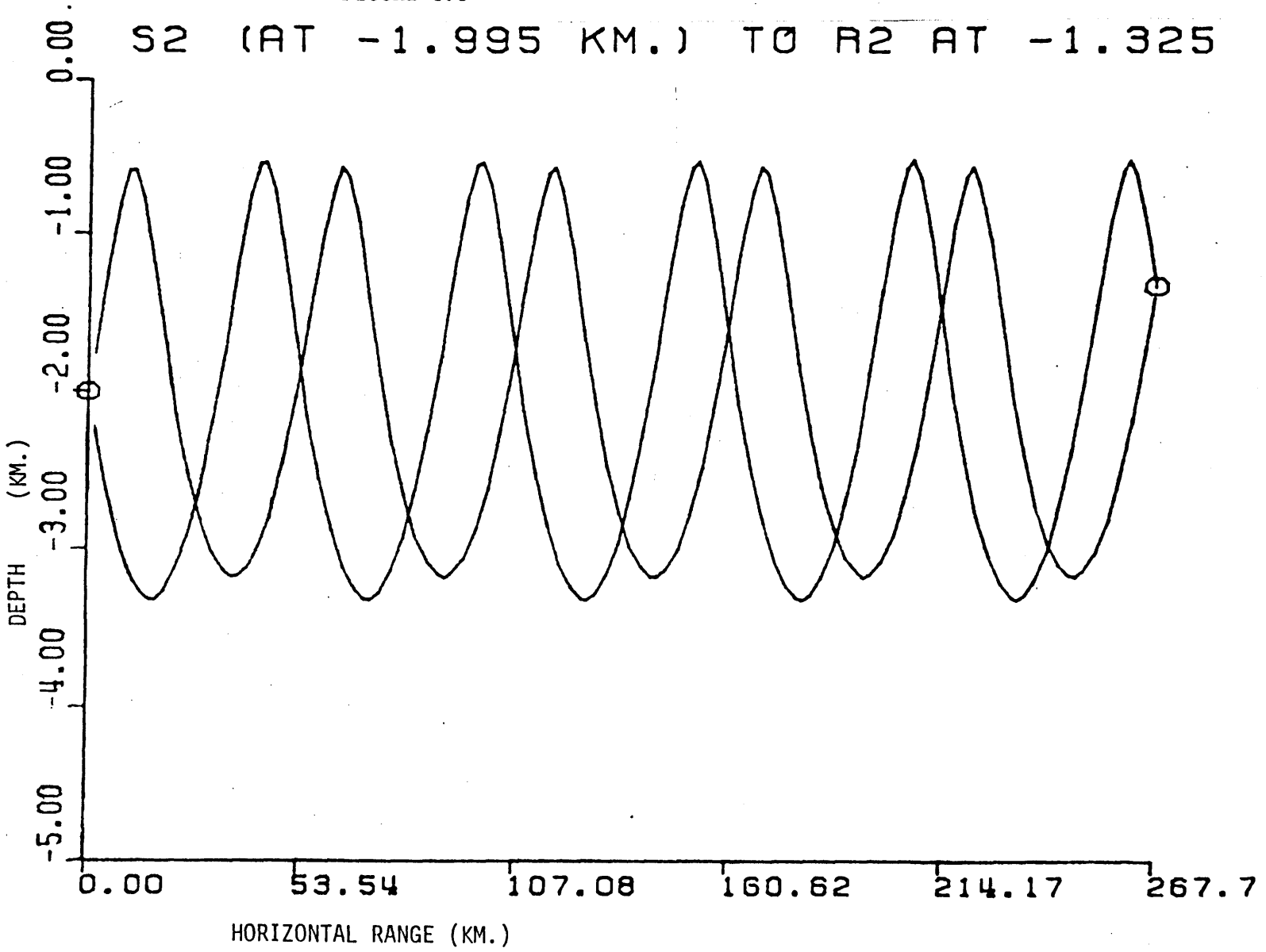
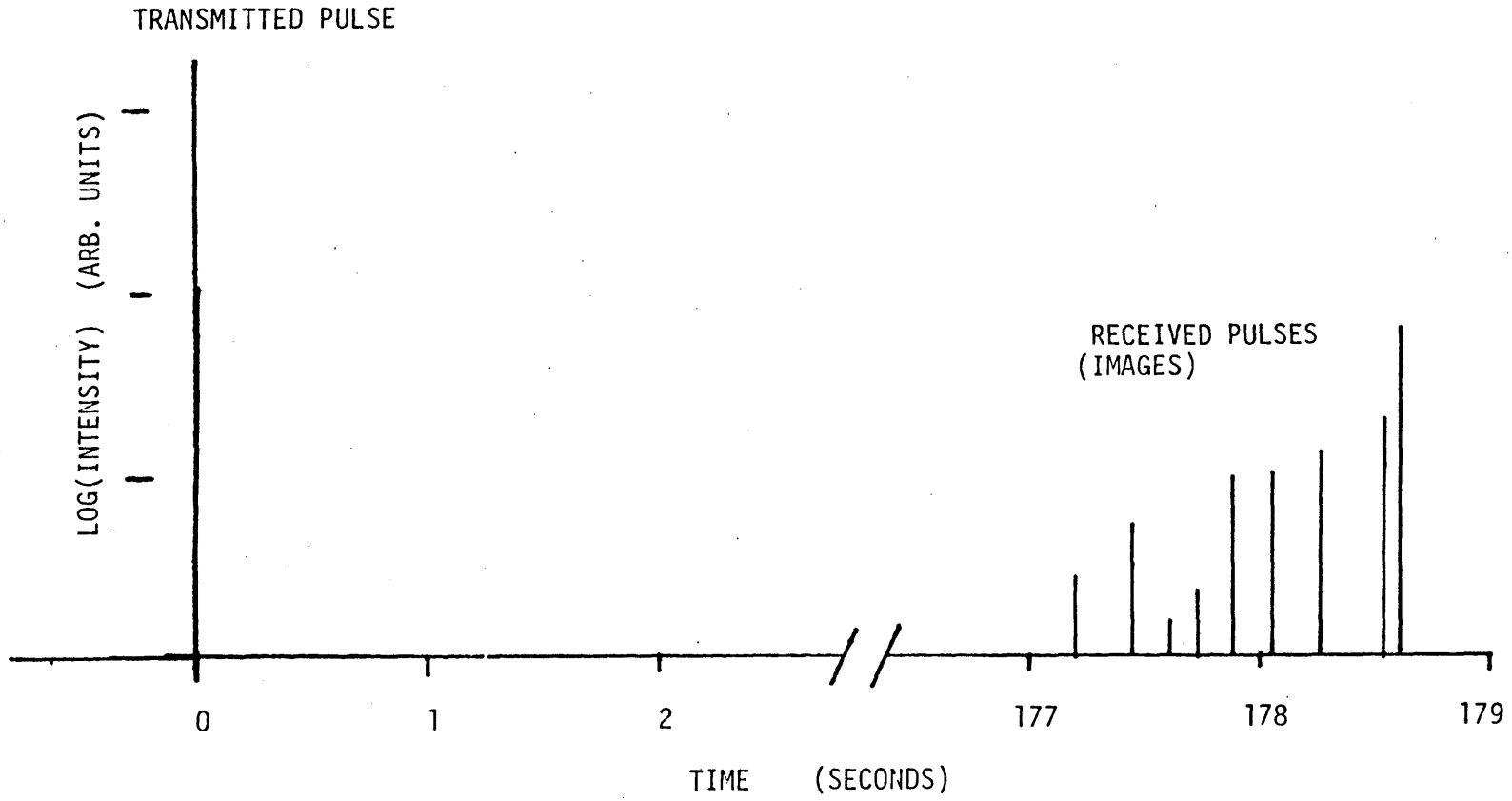


FIGURE 1.3 SCHEMATIC OF TRANSMITTED PULSE AND RECEIVED PULSE "IMAGES" REPRESENTING THE ARRIVALS OF PULSES WHICH TRAVELLED ALONG DIFFERENT RAY PATH. THE RECEIVED PULSE IMAGES ARE CALLED "MULTIPATH ARRIVALS".



The earliest experiments were mounted to gain information on how sound propagated in the ocean. Once the theory describing ocean acoustics ("the forward problem") was understood and verified, investigators began to consider the "inverse problem"--observing propagation and inferring ocean structure. LaCasce and Beckerle (1975) suggested (vaguely) that pulse transmissions might be used to "monitor the periodicities of Rossby waves", on the basis of a simple explosion-monitoring experiment southwest of Bermuda. Porter and Spindel, in 1977, proposed a specific way to monitor eddies using transmissions of 220 Hz pulses, based on their already considerable experience.

Munk and Worcester (1976) had also suggested that oceanographic information might be obtained from acoustic moorings, while an experiment by Peter Worcester (1977), along with Munk and Birdsall, tested the practicality of acoustic measurements of current over relatively short range. Worcester transmitted sound between transceivers suspended from two ships 25 km. apart, and used differences in pulse travel times between reciprocal ray paths to infer current velocity averaged along the ray paths, but encountered problems, such as untracked source and receiver motion. The currents produced arrival time shifts on the order of milliseconds, while the drifting and heaving ships introduced travel time changes two orders of magnitude

larger. The experiment used 2 kHz sources to achieve enough bandwidth to transmit pulses, so that it would have been difficult to work at longer range, and the "inverse problem" of unscrambling the averaging along ray paths had not been attacked.

Hugo Bezdek put Worcester, Munk, and Birdsall in touch with Spindel and Porter, as a result of their experience with mooring tracking, and the common interest of observing the ocean acoustically. Spindel and Munk went to sea together in 1978 to deploy the 2 kHz sources on a mooring with tracking. Spindel also deployed the first source that sent coded signals at 220 Hz--using signal processing techniques to make long-range pulse arrival time measurements possible. The success of this add-on test by Spindel was the real beginning of the recognition that long-range acoustic ocean monitoring was truly possible.

If the travel times for pulses following different paths can be reliably distinguished, then slice reconstruction, as in medical tomography, should be possible, although the medical algorithms are not applicable, due to the complicated geometry and incomplete sampling. Theoretical calculations for the North Atlantic (MW) predicted that many different rays should be resolvable, providing a potentially large amount of information, but it was not known whether the paths or the pulse arrival patterns would be stable enough to reliably observe any shifts in travel time along a particular path.

On the basis of Fermat's principle (that sound propagates along paths which extremize the travel time for a given sound speed field) and a careful analysis of internal wave effects, MW predicted that the paths should be stable, so that changes due to the evolution of the ocean mesoscale would be resolvable.

The need to determine pulse arrival times requires a narrow pulse, and therefore a wide bandwidth of the transmitted signal. This is not a problem if explosives are used as the source, but is difficult for a low-frequency, low-power self-contained source such as would be needed on a long-duration mooring. The early low-frequency acoustic transmissions were CW, as mentioned above, as phases (travel times) were regarded as too unstable to be resolved, particularly given the limited bandwidths. The 270 Hz sources developed by Doug Webb for the SOFAR float program (Webb, 1977), were modified to send CW signals at 220 Hz (Spindel, Porter, and Webb, 1977). Later, digital signal processing techniques made possible by burgeoning computer technology were employed to send wider band, coded signals at 220 Hz (Spindel 1979) and 224 Hz (Spindel 1980). The source that Spindel deployed in 1978 which showed that accurate long-range arrival times were attainable in principle was of this type. The sources were derived from the SOFAR float program, but were modified to be part of a mooring and were larger and heavier than the original sources on the floats.

The 224 Hz sources used in the 1981 Tomography experiment use piezoelectric transducers to drive 4 large resonant tubes, resembling organ pipes, for efficient coupling to the water, and have bandwidths of 20 Hz. They transmitted a phase-coded digital signal which was phase-matched filtered (Birdsall, 1976) at the receiver to produce coherence peaks at lags where the received signal closely matched a stored replica of the transmitted signal. These peaks can be thought of as representing the arrivals of short packets of energy from the source, simulating ray arrivals from a broadband explosive pulse. The travel times for these "pseudo pulses" can be measured accurately enough to discriminate between different multipath arrivals. It thus became possible to test the conjecture that the arrivals would be stable enough to use as data in an ocean observation program.

Two tests were mounted, one over a 900 km. path near Bermuda (Spiesberger, Spindel, and Metzger, 1980), and another over 300 km. paths (Spindel and Speisberger, 1981). Both experiments confirmed MW's predictions, in fact surpassing their expectations, showing clearly resolvable paths which shifted in response to oceanic changes while preserving a stable pattern of arrival times. It was also learned that variations in arrival time for the final cutoff of a set of acoustic pulses from underwater explosions had been observed in the early 1960's (Hamilton, 1977).

Given the stability and resolvability of several different paths, consider the "inverse problem" of converting observed shifts in travel time for the different rays into maps of sound speed changes in the intervening ocean. In medical tomography, the X-rays pass directly through the patient and are transmitted from a nearly continuous set of points around the perimeter of the region to be imaged, so that transform techniques may be used in the reconstruction. Ocean acoustic tomography relies on a relatively small set of complicated ray paths (Figure 1.2) which imperfectly and inhomogeneously sample the ocean. Reconstructions require geophysical inverse theory, one form of which was developed by Backus and Gilbert (1967) to treat imperfect and incomplete data.

In the paper which introduced tomography, Munk and Wunsch presented a solution of the inverse problem for the 2 dimensional problem with several sources and receivers distributed around a square region divided into boxes. These preliminary simulations suggested that data from 4 sources and 4 receivers could provide 16 independent pieces of information and adequately resolve a 1000 km. by 1000 km. region divided into 16 boxes. If more boxes were used, the ability to resolve any given box declined, but given the simplicity of the initial case, there were many prospects for improvement.

1.3 THE 1981 EXPERIMENT BY THE OCEAN TOMOGRAPHY GROUP

On the basis of these calculations and the transmission experiments mentioned above, the researchers involved in the various aspects of the problems came together as The Ocean Tomography Group and designed an experiment to demonstrate tomography as a practical observation technique (Ocean Tomography Group, 1982). This experiment was carried out during the first half of 1981, and much of the work described in this thesis was focussed on the particular application of tomography embodied by the 1981 experiment.

The 1981 experiment was designed to emulate MODE, (MODE Group, 1978), with interest focused on the dynamical evolution of mesocale features in a region south west of Bermuda. This location was chosen because a main purpose of the experiment was to demonstrate the utility of acoustic tomography as an oceanographic observation tool. It was thought best to avoid unexplored regions, in order to optimize the design of the array with archived data. In any case, the description of apparently new phenomena by the acoustics alone would have been regarded as questionable. The region was chosen to be out of the energetic Gulf Stream near field, so that the eddy energy would be moderate to weak, in order to avoid problems with important nonlinearities in the acoustics or dangerous mooring movement.

The experiment has been described in the paper by the Ocean Tomography Group (1982) but will be summarized here to fix ideas. The experimental layout is shown in Figure 1.4. 4 224 Hz sources and 5 WHOI and SIO receivers were moored in an array within a 300 km. by 300 km. box centered on 26 N, 70 W. The experimental array also included 2 conventional oceanographic moorings with current meters and temperature-pressure (T-P) recorders. During the course of the experiment, 3 CTD and bottle hydrographic surveys were made by NOAA ships in the region, and several AXBT flights were made by the Navy, in order to have traditional measurements in the region for comparison with the tomography results.

A typical sound speed profile for this region is shown as Figure 1.5, showing the strong waveguide with the axis at about 1300 meters depth. The sources and receivers were mounted on subsurface moorings to reduce leaning in currents. Instrument depths were nominally 2000 meters, well below the sound speed minimum. When both source and receiver are located on the sound channel axis, pairs of rays with equal, even numbers of turning points but opposite launch angle sign have identical travel times if the profile is range indepedent. The actual ocean is range-dependent, but the degeneracy can still impede peak resolution and identification. Off-axis geometry breaks this degeneracy. Moving source and receiver off the sound channel axis also decreased the number of rays received, but did not greatly reduce the number of useful rays. Most

FIGURE 1.4 SKETCH OF THE LAYOUT OF THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT. (TAKEN FROM NATURE 299 BY THE OCEAN ACOUSTIC TOMOGRAPHY GROUP, 1982).

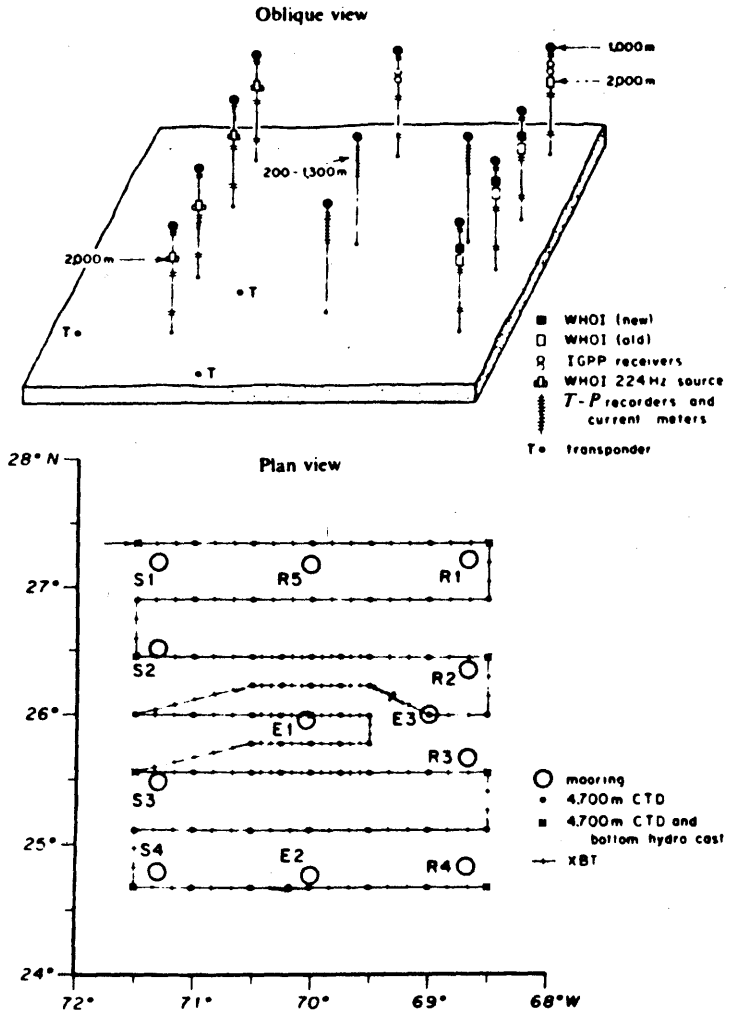
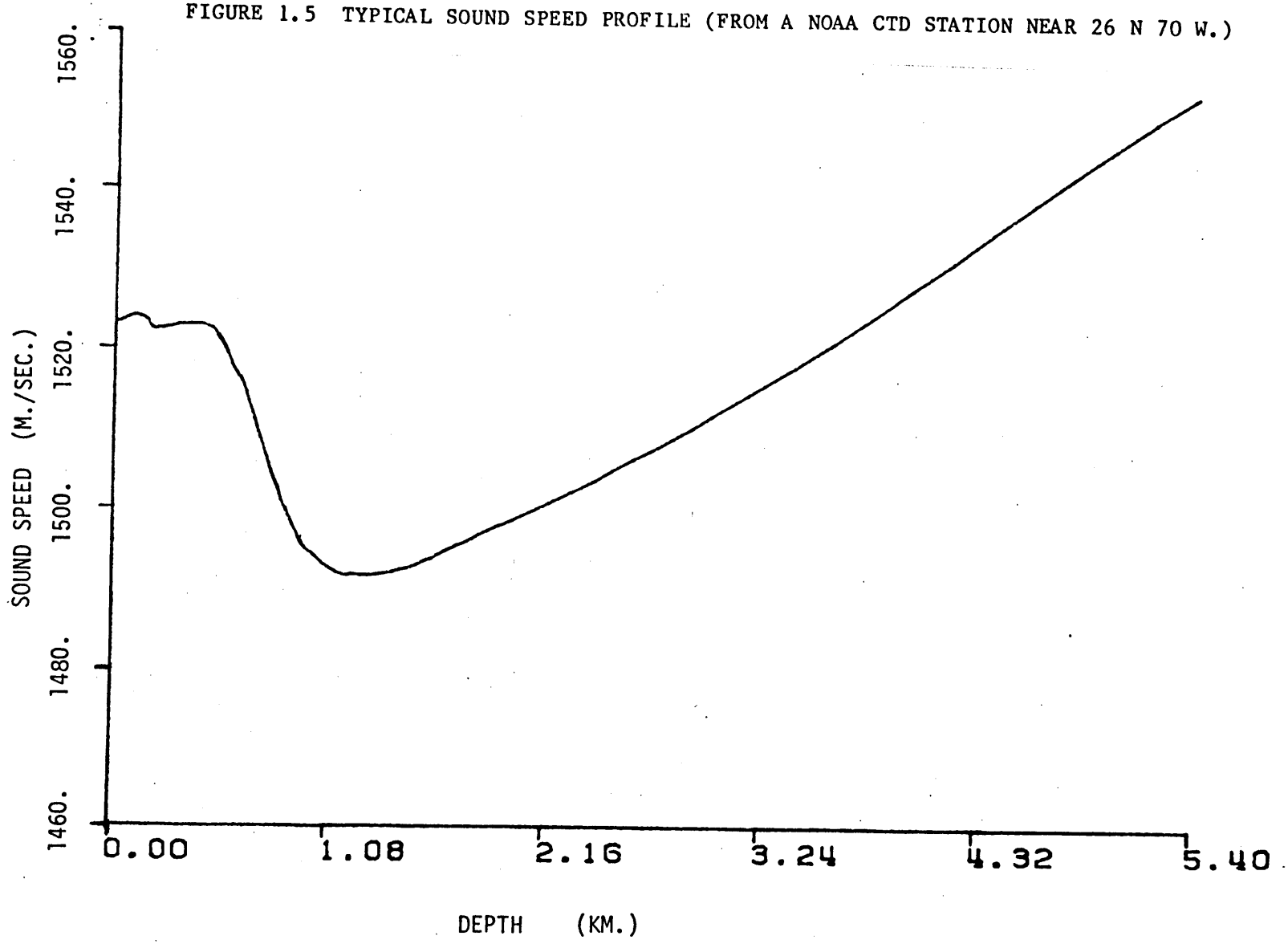


FIGURE 1.5 TYPICAL SOUND SPEED PROFILE (FROM A NOAA CTD STATION NEAR 26 N 70 W.)



of the rays eliminated by this position shift stay close to the channel axis, and have nearly identical travel times, indistinguishable by the practical system. Each source-receiver pair defines a vertical plane through the box along which the rays which leave that source and reach that receiver propagate. Figure 1.2 shows a typical source-receiver path with a number of rays, while Figure 1.6 shows the time evolution of an arrival pattern for one of the source-receiver pairs during the 1981 experiment.

Changes in the arrival pattern can be caused by several mechanisms besides the variation of the ocean sound speed. For the system to be useful, these other sources of variance must be considered as noise, and must be reduced to levels far below the mesoscale travel time changes. As a basis on which to design the 1981 experiment, MW estimated the sound speed variations for the mesoscale at about 200 msec, requiring a noise level somewhere below 10 msec. After the experiment was in the water, comprehensive calculations of rms expected variations based on the data from the MODE experiment revised the original estimate downward to about 40 msec, making the error requirements far more stringent.

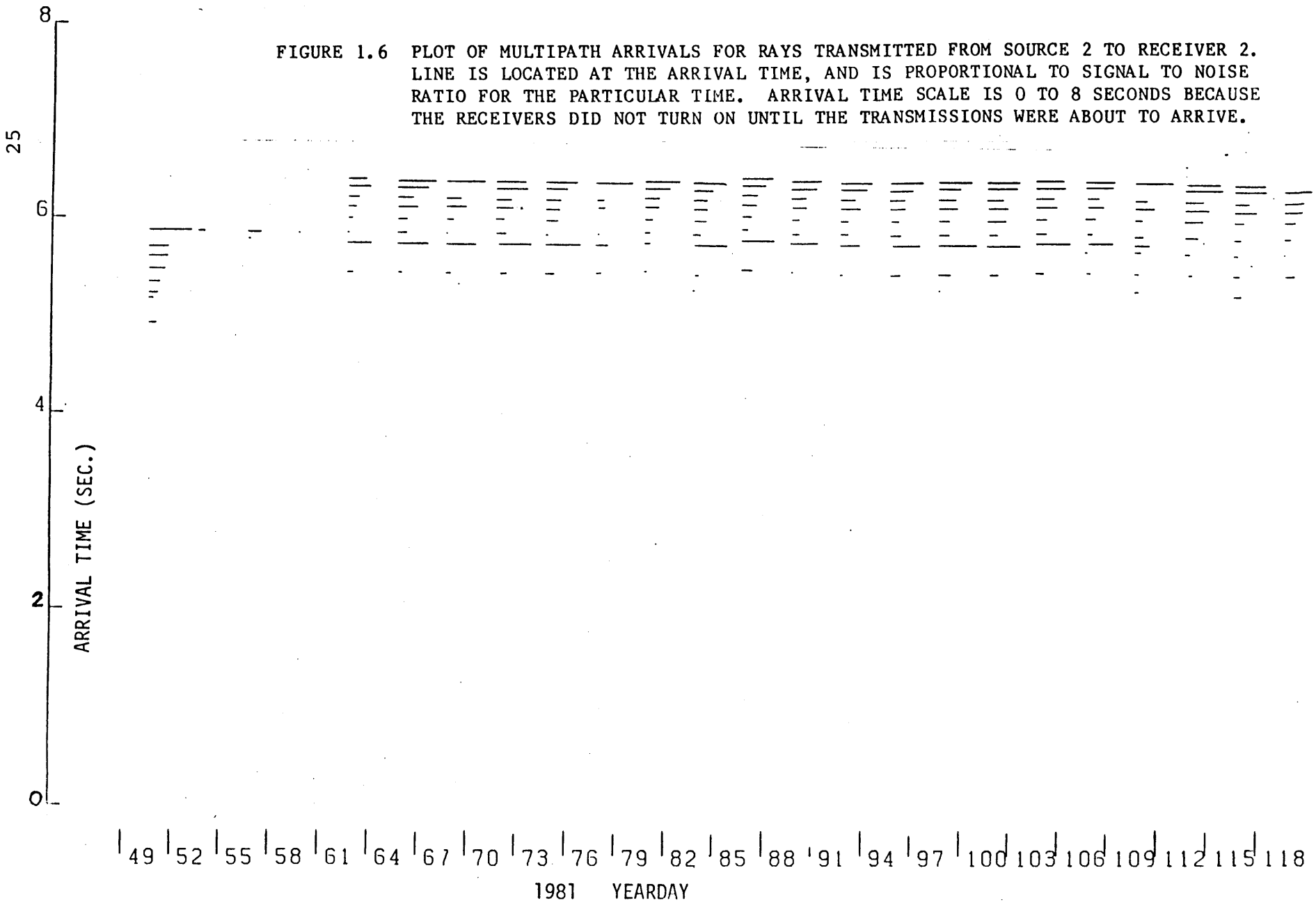


FIGURE 1.6 PLOT OF MULTIPATH ARRIVALS FOR RAYS TRANSMITTED FROM SOURCE 2 TO RECEIVER 2. LINE IS LOCATED AT THE ARRIVAL TIME, AND IS PROPORTIONAL TO SIGNAL TO NOISE RATIO FOR THE PARTICULAR TIME. ARRIVAL TIME SCALE IS 0 TO 8 SECONDS BECAUSE THE RECEIVERS DID NOT TURN ON UNTIL THE TRANSMISSIONS WERE ABOUT TO ARRIVE.

Because tomography is based on transmissions from sources to receivers, the data are very sensitive to errors in mooring position. Given a typical oceanic sound speed of 1500 m/sec, 15 meters of error in the length of a ray adds 10 msec. of travel time error. This is important when compared with 40 msec., the expected level of travel time changes due to the mesoscale field. Knowing the positions of the moorings is thus much more critical than with a conventional array of moorings. In addition, moorings can move around, leaning in response to ocean currents, so that horizontal position changes of 1000 meters are not unexpected for the top of a standard mooring in 5000 meters of water. The tomography moorings were subsurface, meaning that the tops of the moorings were syntactic foam floats or steel spheres at about 750 to 1000 meters depth, (see Figure 1.4), and were moderately taut in order to reduce the amplitude of the mooring motion. In spite of this design, instrument position shifts of 500 meters in the horizontal and 100 meters in the vertical were expected.

Tomography also requires a high degree of clock precision and accuracy over a long (4 months in the 1981 experiment) underwater deployment. The sources and receivers are autonomous, so it is possible for the clocks in each instrument to drift independently, adding errors to the travel time measurements. If these errors are to be

kept to 1 msec over the course of the experiment, that means 1 millisecond in 4 months, or one part in 10^{10}. The quartz crystal oscillators available today cannot meet that standard, especially if they are subjected to the rapid temperature changes associated with mooring deployment. Rubidium oscillators can attain this accuracy, but consume far too much power, given the limitations to the battery power available at present.

The problem of mooring motion was solved by using a refined version of the mooring tracking system developed at Woods Hole Oceanographic Institution by Spindel, Porter, and Jaffee (1973). The system uses three transponders installed on the ocean bottom in a triangle surrounding the mooring, which are interrogated by another transponder on the mooring. The travel times for the pulses sent between these instruments can be converted to mooring position, allowing continuous tracking of the transponder on the mooring with an accuracy of about 1.5 meter. A model of the mooring is then used to estimate the motion of the source or receiver given the motion of the level at which the transponder was located. For this system to operate most accurately, the relative positions of the mooring and the three transponders must be surveyed (to within a few meters) relative to the mooring to be tracked. Tomography adds another complication, because the direction of the

displacement relative to the other moorings is very important. Once the mooring shifts from some arbitrary initial position were known, the time base of the received signal was shifted by $\Delta T = \Delta R/C$, where ΔR is the shift in mooring position converted to extra horizontal range for the source-receiver pair in question, and C is an averaged sound speed at the level of the receiver.

The problem of clock drift was also solved by Spindel, by using a rubidium clock as a frequency standard, checking for drift of the quartz oscillators. The rubidium standards were turned on daily, and after they had time to stabilize, they were used to compute the relative frequency shifts of the quartz with respect to the rubidium. These shifts were recorded in the receiver. Using this record, the time base of each instrument could be adjusted later, bringing practical clock accuracy up to about 2 msec.

The need to measure these quantities, while not particularly onerous, does add complication and expense to both the acoustic instrumentation and mooring deployment. It likewise multiplies the number of systems which may fail. During the 1981 experiment, some of the mooring motion transponders returned incomplete data sets, making it impossible to apply mooring motion corrections to part of the data. For these reasons, extensions to the inverse techniques were developed to permit mapping using uncorrected acoustic data. These procedures may perhaps obviate complicated correction logging in future experiments.

1.4 PREVIEW OF THESIS CONTENTS AND GOALS

Given that the engineering problems of obtaining the data for the mesoscale have been solved, the usefulness of the tomographic system as an observing tool depends on how much information can be extracted from the data. In this thesis I will describe a complete system for treating the acoustic data to construct estimates of the ocean structure. The formalism I will present serves three purposes: 1) To demonstrate and evaluate a specific application of tomography: the 1981 experiment; 2) To provide an analytical and numerical basis for understanding and designing furture experiments, tomographic or otherwise; and 3) To compare and contrast the common linear inverse methods.

Chapter 2 contains a discussion of the ocean acoustics necessary for understanding how the sources and receivers sample the ocean. Chapter 3 covers the quasi-geostrophic equations of geophysical fluid mechanics, which form the basis for the models used in the acoustic forward problem and the inverse solution. Chapter 4 is a general discussion of inverse techniques, while Chapter 5 is an intercomparison of many existing inverse methods. Chapter 6 is devoted to inverse techniques as applied to acoustic tomography, and incorporates results from Chapters 2 and 3.

Chapter 7 is concerned with the specific problems which arise when the tomographic system includes moored instruments, as in the 1981 experiment. Chapter 8 discusses the preliminary data reduction for the 1981 experiment,while Chapter 9 describes the details of the inverse techniques applied to the 1981 data. Chapter 10 discusses the results of these inverse techniques and examines the capabilities of tomography, both as applied in 1981 and in the future.

The reader who is not interested in the oceanographic theory or inverse methods may wish to skip to chapters 9 and 10 for the results of the 1981 experiment. In any case, the reader must recognize that the 1981 tomography experiment produced a completely novel data set, so that much time has been required for each stage of data processing. For this reason, the maps and numbers presented here are by no means final or optimal, but represent a "first-pass" look at the capabilities of tomography.

CHAPTER 2

ELEMENTARY OCEAN ACOUSTICS

2.1 THE GEOMETRICAL OPTICS APPROXIMATION: ACOUSTIC RAYS

The attenuation of sound in the ocean is proportional to frequency so that sound with a frequency of about 200 Hz can be transmitted usefully over several thousand kilometers before being swamped by noise. The SOFAR floats (Baker, 1981 in Warren & Wunsch) use this low-loss frequency range coupled with the acoustic waveguide typically found in the North Atlantic (See Figure 1.1) to allow tracking of floats over long distances using relatively low-energy, battery powered sources. The first ocean acoustic tomography experiment used similar sources, operating at a center frequency of 224 Hz and transmitting a phase-coded signal suitable for travel time measurement (The Ocean Tomography Group, 1982).

At 200 Hz, sound in the ocean has a wavelength of about 7.5 meters, small when compared with typical scales for the sound-speed structure of either the basic climatological state or the mesoscale fluctuations (Figure 2.1), but large compared to vertical microstructure and most fine structure (see Gregg (1977) for spectra). The

FIGURE 2.1 A: SOUND SPEED ANOMALY AT 700 METERS DEPTH RELATIVE TO THE AVERAGED SOUND SPEED PROFILE SHOWN IN FIGURE 1.1 CALCULATED FROM THE FIRST NOAA CTD SURVEY DURING 1981 YEARDAY 66 TO 85. CONTOURS ARE SOUND SPEED IN METERS PER SECOND, CONTOUR INTERVAL IS 1.0 M/SEC.

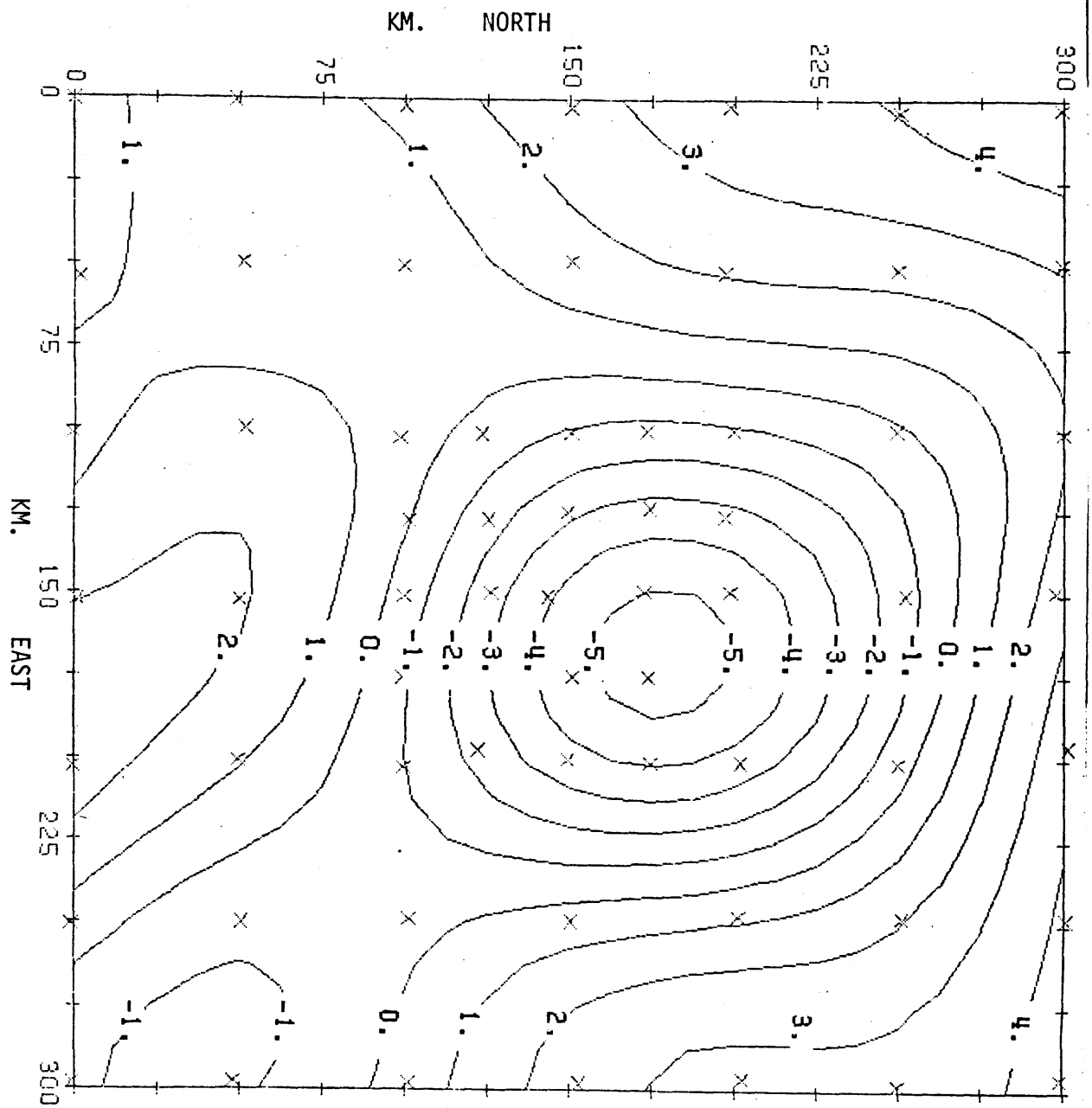
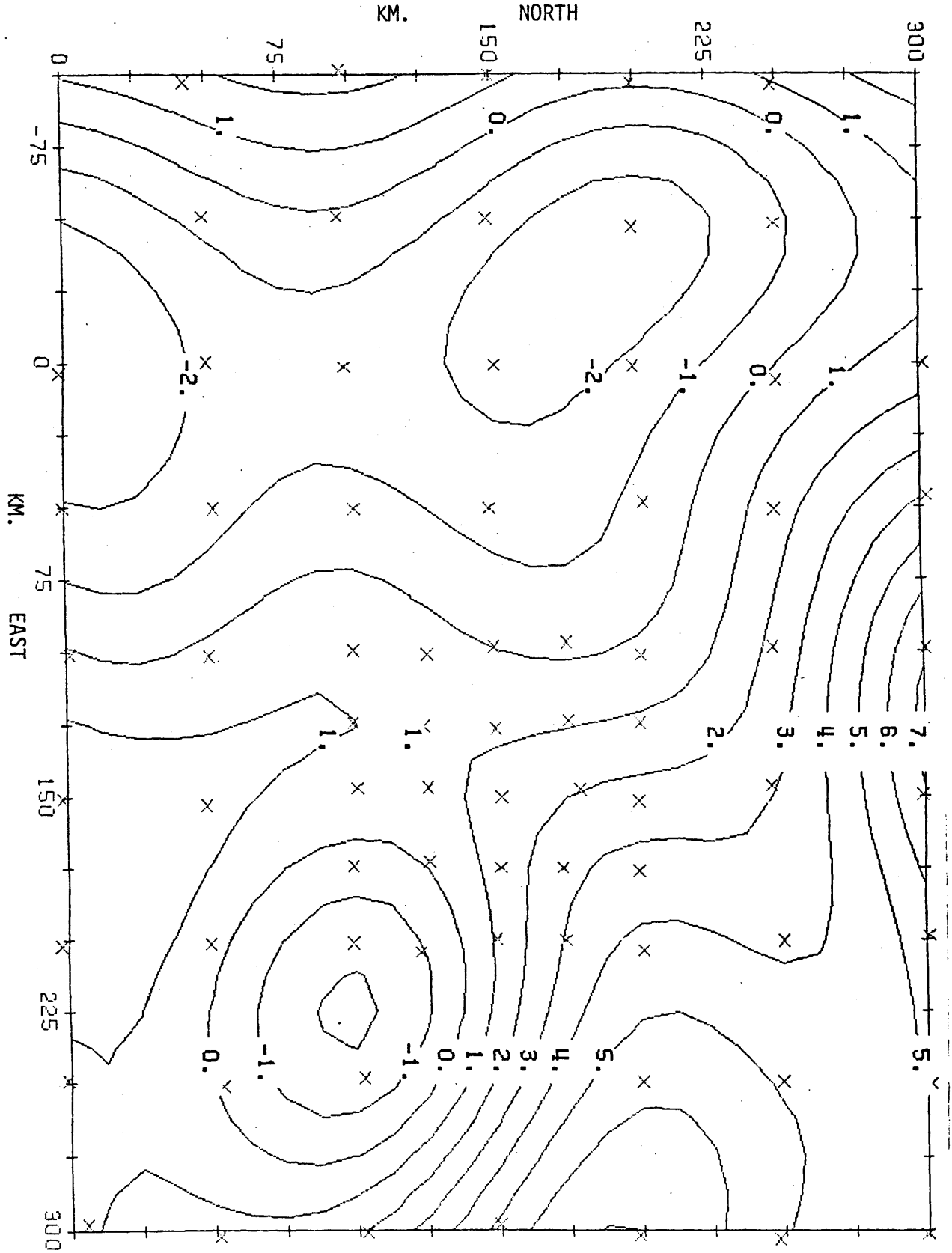


FIGURE 2.1 B: SOUND SPEED ANOMALY AT 700 METERS DEPTH, AS IN FIGURE 2.1 A, FROM SECOND NOAA CTD SURVEY DURING 1981 YEARDAY 120 TO 139.



slow variation of the interesting structures when compared with the sound wavelengths allows a simplification of the acoustic wave equation called the geometrical optics approximation, using the concept of acoustic rays. Other and better approximations may be used to derive different physical pictures, most notably the physical optics extensions to the acoustic ray theory or the use of modes as an alternate description of the propagation of sound. The geometrical optics approximation is simple, but adequate for many needs, including the analysis for the 1981 tomography experiment, so it will be described in greatest detail, although it is not always sufficiently accurate for many applications. The development here will follow Officer (1958).

Let $\phi(\underline{x},t)$ be sound pressure in a resting ocean (or sea bottom). The wave equation for sound is:

$$\nabla^2\phi = \frac{1}{C(\underline{x})^2}\frac{\partial^2\phi}{\partial t^2} \qquad (1)$$

$C(\underline{x})$ is the sound speed, and is considered constant with respect to the time of propagation of the sound energy. Suppose that there is a source of angular frequency ω, then let

$$\phi(\underline{x},t) = \phi_o\exp(i[S(\underline{x}) - \omega t]) \qquad (2)$$

$S(\underline{x})$ is phase as a function of distance. Constraining S to be real, so that amplitude variations are ignored, substitution of (2) into (1) yields

$$\left(\frac{\partial S}{\partial x}\right)2 + \left(\frac{\partial S}{\partial y}\right)2 + \left(\frac{\partial S}{\partial z}\right)2 \equiv$$

$$S_x^2 + S_y^2 + S_z^2 = \omega^2/C(\underline{x})^2 \equiv n^2(\underline{x}) \qquad (3)$$

$\left(\frac{\partial S}{\partial x}\right),\left(\frac{\partial S}{\partial y}\right),\left(\frac{\partial S}{\partial z}\right)$ are the local wavenumbers:

$$\phi(\underline{x},t) = \phi_o\exp(i[S_x\cdot x + S_y\cdot y + S_z\cdot z - \omega t]) \qquad (4)$$

$$\phi(\underline{x},t) = \phi_o\exp(i[\underline{\nabla S}\cdot\underline{x} - \omega t]) \qquad (5)$$

and vary slowly over the scale of a wavelength in the same way that $C(\underline{x})$ does.

The gradient of phase, $\underline{\nabla S} = (S_x, S_y, S_z)$, is normal to the acoustic phase fronts, and in the resting ocean, this is the direction of the local tangent to the ray path, defining the ray path. For s ≡ arc length along a ray,

$$\frac{dx}{ds} = \frac{S_x}{n(\underline{x})} = S_x \cdot C(\underline{x})/\omega \qquad (6a)$$

$$\frac{dy}{ds} = \frac{S_y}{n(\underline{x})} \qquad (6b)$$

$$\frac{dz}{ds} = \frac{S_z}{n(\underline{x})} \qquad (6c)$$

Call $\underline{\nabla S} \equiv \underline{k}(\underline{x})$, the local wavenumber vector:

$$\phi(x,t) = \phi_o \exp(i[\underline{k}(\underline{x}) \cdot \underline{x} - \omega t]$$

Taking d/ds of [6(a,b,c)] yields (Officer, 1958):

$$\frac{d}{ds}\left(n(\underline{x}(s))\frac{dx}{ds}\right) = \frac{\partial n}{\partial x} \qquad (7a)$$

$$\frac{d}{ds}\left(n(\underline{x}(s))\frac{dy}{ds}\right) = \frac{\partial n}{\partial y} \qquad (7b)$$

$$\frac{d}{ds}\left(n(\underline{x}(s))\frac{dz}{ds}\right) = \frac{\partial n}{\partial z} \qquad (7c)$$

These are the equations that are integrated by most ray-tracing programs to determine $\Gamma_i = \underline{x}(s)$, the ith ray path, given an initial location, launch angle, and direction. Normally, the sources are assumed to radiate with spherical symmetry, so that we only consider propagation in the vertical plane between source and receiver, so that instead of $\underline{x}(s)$, we use $(r(s),z(s))$, where r is horizontal range.

If $n = n(z)$ only, which is approximately true for the ocean, then $\partial n/\partial r = 0$ and so (7a,b,c) become:

$$\frac{d}{ds}\left(n(\underline{x}(s))\frac{dr}{ds}\right) = \frac{\partial n}{\partial r} = 0 \qquad (8a)$$

$$\frac{d}{ds}\left(n(\underline{x}(s))\frac{dz}{ds}\right) = \frac{\partial n}{\partial z} = \frac{dn}{dz} \qquad (8b)$$

8(a) is a statement of Snell's law: that the horizontal component of the wavenumber is conserved when the sound speed varies only as a function of z, or

$$\left(n(\underline{x}(s))\frac{dr}{ds}\right) = \text{constant.} \qquad (9)$$

If θ is the angle that the ray makes with the with the horizontal, then $dr/ds = \cos(\theta)$, and we get

$$\frac{\cos(\theta)}{C(z)} = \text{constant along a ray path.} \qquad (10)$$

Cos(θ)/C(z) is sometimes called P, the "ray parameter", so that (9) becomes:

$$dP/ds=0 \quad \text{along } \Gamma_i, \tag{11}$$

expressing the conservation of ray parameter along ray paths. Ray-tracing programs may be range-independent (C = C(z)), or range-dependent in two or three dimensions (C = C(x,z) or C = C(x,y,z)). The ray tracing code used for the calculations in this thesis was originally written to be range-independent, but was modified to trace rays in a succession of locally range independent sound speed profiles, making it crudely range-dependent in two dimensions (r,z). The ray is assumed to travel in a vertical plane oriented along a line between source and receiver, ignoring any bending due to horizontal sound speed gradients.

For most mesoscale features these gradients are small compared to the vertical gradients and so the horizontal ray bending has been ignored, although Munk (1980) has treated horizontal ray bending in detail for simulated mesoscale eddies and Gulf Stream rings in two dimensions (horizontal plane). He finds that the maximum deflection angle is proportional to ν, the fractional change in sound speed ($\nu \equiv C'/C_o$):

$$\text{Maximum deflection angle} = 2\theta_{max} = .664 \cdot \nu$$

for a circularly symmetric eddy. If the feature is

equidistant from source and receiver, then the ray geometry can be approximated by an isosceles triangle (Figure 2.2).

The extra ray arc length is thus

$$\Delta R = R/\cos(\theta) - R = R(1/\cos(\theta) - 1)$$

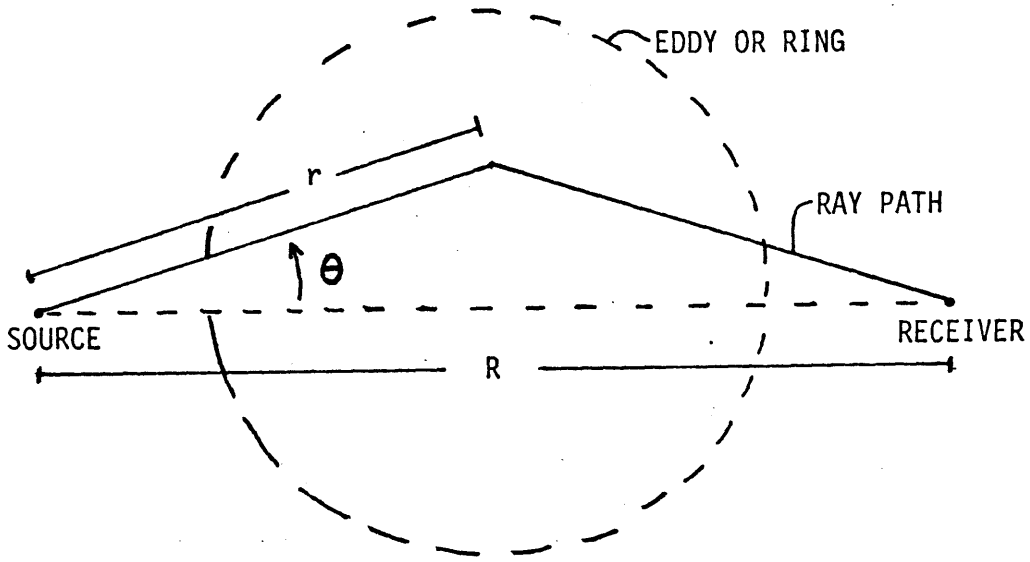
For a 15 m/sec eddy amplitude,

$$\nu = 1. \times 10^{-2}$$

$$\Delta R/R = 5.5 \times 10^{-6}$$

This would cause an error of 1 msec at 300 km range, but most eddies would not have the proper configuration, and the expected rms error is much smaller.

FIGURE 2.2 SCHEMATIC OF HORIZONTAL RAY PATH DEFLECTION FOR A CIRCULAR EDDY

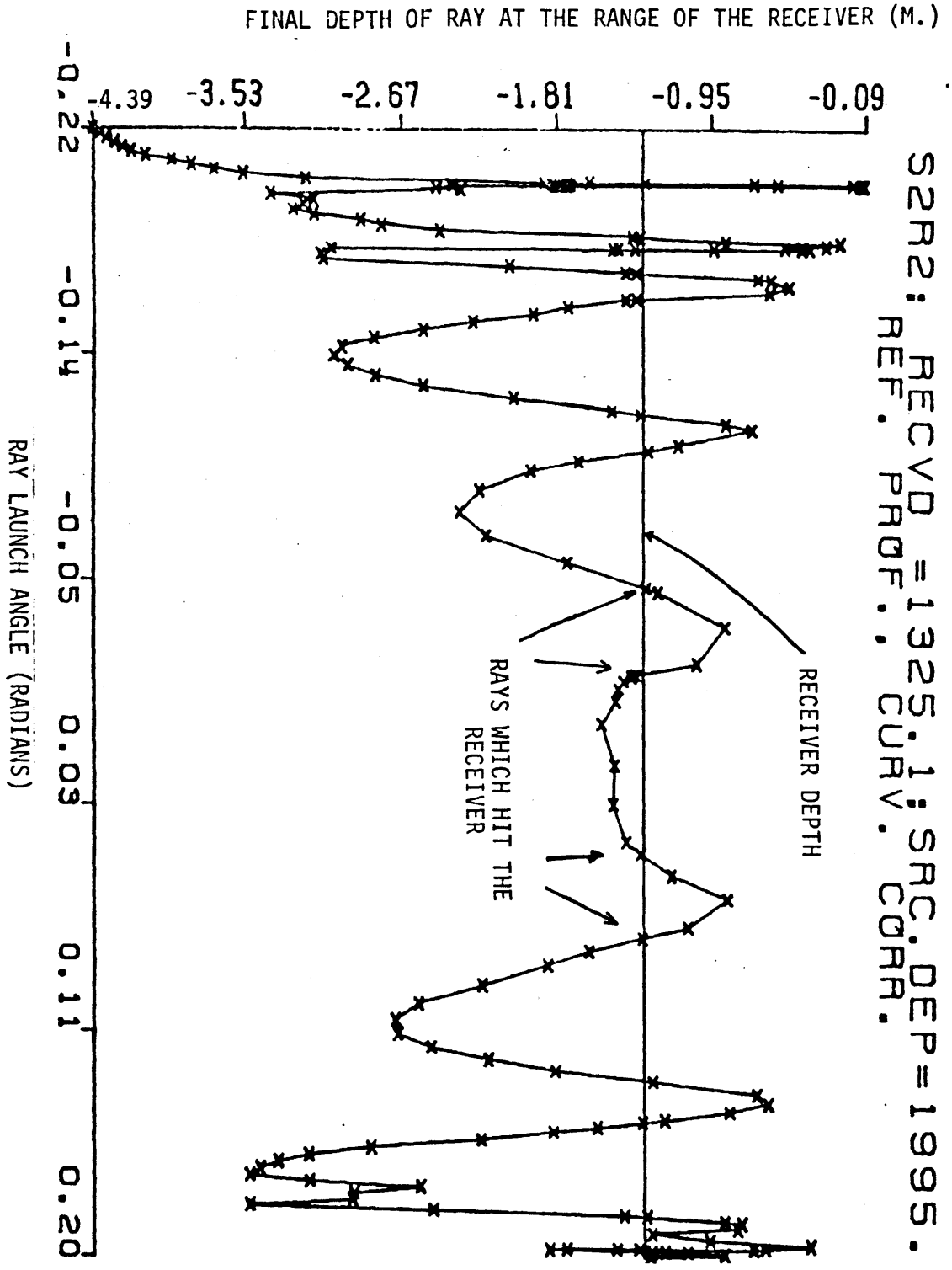


ACTUAL RAY PATH = $2r = R/(\cos\theta)$

2.2 ACOUSTIC RAY TRACING: THE EIGENRAY PROBLEM

Although the rays have been assumed to travel in a vertical plane between source and receiver, only a few of the many possible launch angles from a given source will yield a ray which intersects the receiver (Figure 1.2). The rays that hit the receiver are called eigenrays and are solutions of an eigenvalue problem, as demonstrated for a simple case by Munk and Wunsch (1982). In the case of a complicated or range dependent sound speed profile, analytical solutions to this eigenvalue problem become impossible, and numerical techniques for determining eigenrays must be sought. The most obvious, and perhaps least efficient method merely searchs through a range of launch angles, repeatedly tracing rays out to the range of the receiver and converging on and saving as solutions those rays which pass close enough to be considered as having hit the receiver (Figure 2.3). This technique works whether the code is range dependent or independent, for any sound speed profile or bottom topography which can be treated by the program.

FIGURE 2.3 DEPTH OF ACOUSTIC RAY AT THE RANGE OF THE RECEIVER AS A FUNCTION OF LAUNCH ANGLE FROM THE SOURCE. THE LINE MARKS THE RECEIVER DEPTH. EIGENRAYS ARE FOUND AT ANGLES WHERE THE RAY DEPTH CURVE INTERSECTS THE DEPTH OF THE RECEIVER.



Efficient techniques for determining the sound field at the receiver exist in the seismic literature, and have been successfully applied to the oceanic problem (Brown, 1982). These methods involve keeping more terms in the WKB approximation applied to the propagation equation, and producing "synthetic seismograms" which predict both the amplitude and phase (arrival time) of the sound waves reaching the receiver. These techniques have the advantage that they predict "diffracted arrivals", sound energy leaking from rays which do not intersect the receiver, in the geometric optics sense, but which have turning points at the range of the receiver. The amplitude of the sound pressue field is large at the turning point (∞ is predicted by the geometrical optics approximation) and if the receiver is within a few hundred meters, the exponentially decaying leakage field may remain large enough to be detected as a ray arrival. This is analogous to tunelling in quantum mechanics.

Purely refracted rays are usually labelled by the number of turning points and the sign of the launch angle, thus a +11 RR ray has 11 turning points, a positive launch angle, and is refracted both above and below. Rays which hit the sea surface or bottom are reflected by the discontinuity in sound speed at the boundary, and may still

be received. These are also identified by the number of turning points, including the surface and bottom bounces, and the sign of the launch angle, as in +12 SRBR (both surface and bottom reflected) or -9 RSR (reflected from the surface, refracted at the lower turning point).

2.3 THE FORWARD PROBLEM: TRAVEL TIMES IN THE OCEAN

Once the path of a ray, call it ray i, has been traced from the source to the receiver, it is possible to calculate the travel time, T_i, by integrating along the ray path, Γ_i:

$$T_i = \int_{\Gamma_i} \frac{ds}{C(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}} \qquad (12)$$

s is arclength along the ray, $\underline{\tau}$ is a unit vector tangent to the ray, and the ocean is assumed to change negligibly during the time the ray is propagating. Each eigenray has a unique launch angle, and, therefore, a unique path through the ocean, sampling the sound speed field differently from other eigenrays. Because the sound speed profile changes strongly with depth, the total travel time for a ray which has much of its arclength in high-speed regions will be smaller than for a ray with the same path length but in low-speed regions. Different rays can usually be distinguished at the receiver by differing travel times, (see Figure 1.3). The pattern of ray arrivals is dependent on the sound speed profile.

The velocity term in the denominator of the integrand,

$$\underline{u}(\underline{x},t)\cdot\underline{\tau}, \qquad (13)$$

accounts for changes in the apparent speed of sound due to current, provided local shear can be ignored (Hamilton, et al., 1980). Currents have been ignored in the ray tracing because the magnitude of the current shear in the ocean is typically

$$\frac{10 \text{ cm/sec}}{1000 \text{ meters}} = O(10^{-4})$$

the typical sound speed gradient is stronger:

$$\frac{\partial C}{\partial z} = \frac{4 \text{ m/sec}}{100 \text{ meters}} = O(10^{-2}).$$

Sound speed gradients thus dominate ray bending, except perhaps when the rays pass parallel to frontal zones such as the Gulf Stream.

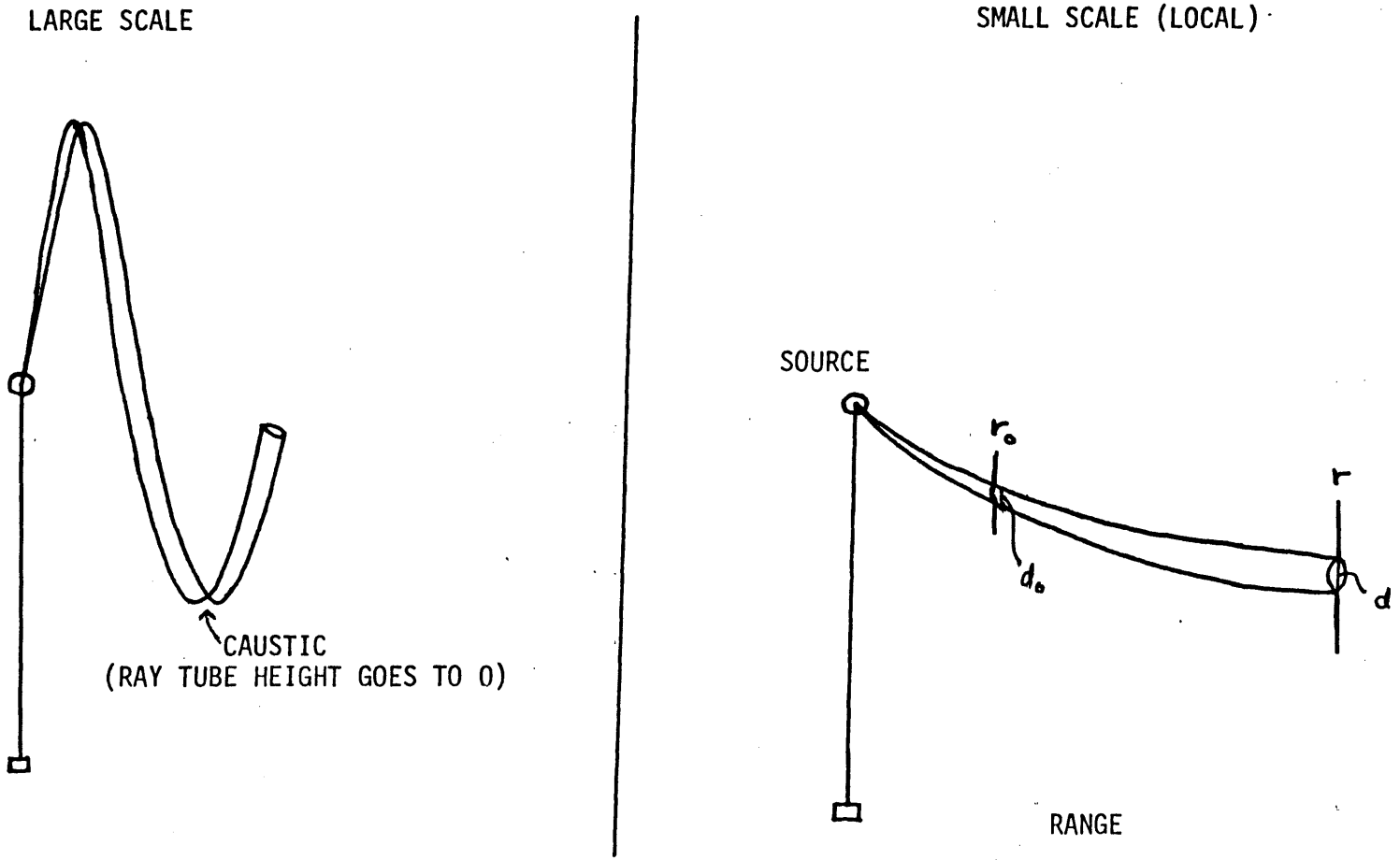
Internal waves produce both sound speed gradients and current shear at scales on the order of meters. These features are comparable in scale to an acoustic wavelength, and tend to scatter the sound, blurring the simple ray paths calculated for the large-scale refraction into ensembles of micro-multipaths which change with the

internal waves. These shifting paths interfere with one another, producing variations in overall travel time for the path and significant changes in the intensity of the received sound. There is a rich literature on the physics of these interactions (see, for example Flatté, et. al., 1979), and much information on the statistics of the internal wave field can be gained from examining the short-time changes in amplitude and phase. It would be very interesting to extend the tomographic inverse techniques to use the many crossing paths to resolve spatial structure of the internal wave field in the same way that they are now used to observe the mesoscale. Unfortunately, the approximations used above do not apply to the internal wave scales, so a separate development is required, and is outside the scope of this thesis. The shifts produced by the internal waves have been treated as noise in the inversions for the mesoscale field, so adding the physics of internal wave scattering to the inversion would improve the estimates of the mesoscale, even if the information about the internal waves was not directly useful.

Equation (12) describes the dependence of travel time on the sound speed and current fields in the region through which the ith ray travels, and is referred to as the solution to the "forward problem", a general term for describing the dependence of the observed data on the unknown. Solving the forward problem for amplitude presents more of a difficulty, because the geometrical optics approximation ignores amplitudes. Heuristic amplitude estimates may be made by considering two rays differing by a small amount in launch angle. The area between the two rays forms a "ray tube" (Figure 2.4). The acoustic energy propagates along the rays and therefore does not pass through the sides of the tube, so energy flux is conserved along the tube. The intensity is then inversely proportional to the area of the tube. For a radially symmetric source, neglecting dissipation, let I_o be the initial intensity, d_o the initial vertical seperation of the two rays, and r_o the range at which these two were specified. At some greater range, r, the seperation will be d, and the intensity will be I, but the energy flux will be conserved:

$$F_o = I_o \cdot d_o \cdot 2\pi r_o = I \cdot d \cdot 2\pi r = F \qquad (14)$$

FIGURE 2.4 SKETCH OF "RAY TUBE" (NOT SHOWING CYLINDRICAL SYMMETRY.)



The intensity at this range must then be

$$I = \frac{I_o \cdot d_o \cdot 2\pi r_o}{d \cdot 2\pi r} = \frac{I_o \cdot d_o \cdot r_o}{d \cdot r} \qquad (15)$$

The low-order character of the oceanic sound speed profile is that of a waveguide (see Figure 1.1), so two rays initially differing by a small angle will follow similar paths, and the vertical separation between the walls of the ray tube will generally increase relatively slowly. The intensity loss is therefore due almost entirely to the range increase in equation (15), which corresponds to cylindrical spreading. This is one of the reasons that long range acoustic transmissions are possible at reasonable power.

This crude amplitude estimate has little to recommend it besides simplicity. It becomes infinite at caustics (the points where rays cross, such as at turning points, so that the ray tube height goes to 0) and ignores the often dominant effect of multipath interference due to changes in the sound speed induced by internal waves (Flatte, et al., 1978). The amplitude fluctuations produced by internal waves can dominate those produced by the mesoscale physics, but averaging over many internal wave periods can eliminate much of the variation.

Mike Brown has considered techniques for estimating sound speed field structure using amplitude data, (Brown, 1982), and concluded that the amplitude data was not particularly useful for the 1981 experiment. Amplitude data require a more rigorous treatment of the acoustic propagation than geometric optics, and this thesis will not treat amplitude explicitly. Given an adequate solution to the forward problem, the inverse techniques presented below can be adapted to the use of amplitude data, although they may no longer be the most convenient forms.

2.4 LINEARIZATION OF THE FORWARD PROBLEM

The forward problem for travel time (equation (12)) is nonlinear in the sound speed field, and although methods exist to invert non-linear problems, solutions can be found efficiently if the forward problem can be linearized. Suppose we pick a reference state, $C_o(\underline{x},t)$, with $\underline{u}(\underline{x},t) = 0$, and express the observed ocean sound speed as a perturbation to this basic state:

$$C(\underline{x},t) = C'(\underline{x},t) + C_o(\underline{x},t) \tag{16}$$

For the ocean, $C_o(\underline{x},t)$ is large, O(1500 m/sec), and

$$|C'(\underline{x},t)| << |C_o(\underline{x},t)| \tag{17}$$

so that the integrand of (12) may be expanded:

$$\begin{aligned} T_i &= \int_{\Gamma_i} \frac{ds}{C(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}} \\ &= \int_{\Gamma_i} \frac{ds}{C_o(\underline{x}(s),t) + C'(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}} \\ &= \int_{\Gamma_i} \frac{ds}{C_o(\underline{x}(s),t)} - \int_{\Gamma_i} \frac{[C'(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}]ds}{C_o(\underline{x}(s),t)^2} \\ &\quad + \text{terms } O(C'^2/C_o^3) \end{aligned} \tag{18}$$

For the oceanic mesoscale, $C'/C_o \approx 0.01$ usually, so the linearization in (18) should be good to one part in 10^4. Unfortunately, the path of the integral is also dependent on the sound speed profile, and the effect of sound speed changes on the ray path and thus on the travel time are not easy to parameterize. Hamilton et al. (1980) have made calculations that show that these changes are exactly zero for small perturbation, as a result of Fermat's principle, so that the changes in path due to small changes in the sound speed do not affect the calculation of travel time.

Internal waves induce small-scale fluctuations in the sound speed field through their often large vertical velocities, stretching and compressing the smooth profile. These changes, on scales comparable to the wavelength of sound, cause the acoustic energy to scatter into micro-multipaths, bundles of paths following the "main" path calculated for the mesoscale variations, but blurring its outlines. The sound ray averages the positive and negative perturbations from any given wave, but each micro-multipath will have a slightly different travel time, introducing the possibilty of phase cancellation when the many small paths re-combine. For this reason, internal-wave induced fluctuations affect the amplitude of the sound arrivals more strongly than the travel time,

making travel time a robust datum. Note that Hamilton, et. al. did not prove that the path remains the same, but that the contributions to the travel time from ray path deformation tend to cancel out.

The integral used to calculate travel time for the perturbed ocean can therefore be taken over the unperturbed ray paths, Γ_{oi}, computed for $C_o(\underline{x},t)$, provided

$$|C_o(\underline{x},t)| \gg |C'(\underline{x},t)| \qquad (19)$$

In this case, the linearized forward problem is:

$$T_i = \int_{\Gamma_{oi}} \frac{ds}{C_o(\underline{x}(s),t)} - \int_{\Gamma_{oi}} \frac{[C'(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}]}{C_o(\underline{x}(s),t)^2} ds \qquad (20)$$

or

$$T_i = T_{oi} + T'_i \quad . \qquad (21)$$

Mercer and Booker (1982) have done calculations which produced examples of this relation for Gulf Stream rings of varying energies, and point out that perturbations to the paths affect the sampling of the sound speed field by the ray. In examining their plots of ray travel times vs. ring strength, one is struck by the linearity of the relationship over a large range, although the extremes of the curves are clearly bent. Rings are among the most intense sound speed features encountered in the N. Atlantic, and the experimental region was chosen to reduce the probability of encountering rings, with their attendent complications, in the demonstration experiment.

2.5 THE TRAVEL TIME EFFECTS OF OCEAN CURRENTS

Equation (20) takes into account travel time perturbations that result from both sound speed and ocean currents. This means that, in principle, a tomographic system can produce sound speed, density, and velocity maps without ambiguities due to the "reference level" problem or uncertainty in the T-S relation. In practice, high quality travel time data is necessary in order to distinguish current velocity from sound speed anomalies since the two are averaged together along each ray. The area coverage and error levels must be such that the inverse procedure can identify and separate the two fields. The effects of currents on ray travel times are weaker than those due to sound speed, as can be seen simply by calculating the magnitudes:

$$|C'| \sim O(10 \text{ m/sec}); \quad |u| \sim O(10 \text{ cm/sec})$$

The perturbations due to velocity are thus only a few percent of the total travel time signal. Peter Worcester has pioneered a technique called "reciprocal shooting" (Worcester, 1977), which can greatly improve the current resolving power of the acoustic data by taking advantage of the relative weakness of the effect of current on the sound

rays. If two transceivers transmit to each other in an area with typical currents the ray paths are approximately independent of the direction of travel. For a given ray path, Γ_i, transmitted from Transceiver A to Transciever B, for example, there will exist an oppositely directed path, Γ_j, (from Tr_B to Tr_A), that is identical in all other respects. The linearized forward problem for travel time perturbations can then be written as:

$$T'_i = \int_{\Gamma_{oi}} \frac{[C'(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}]}{C_o(\underline{x}(s),t)^2} ds$$

$$T'_j = \int_{\Gamma_{oj}} \frac{[C'(\underline{x}(s),t) + \underline{u}(\underline{x},t)\cdot\underline{\tau}]}{C_o(\underline{x}(s),t)^2} ds$$

$$= \int_{\Gamma_{oi}} \frac{[C'(\underline{x}(s),t) - \underline{u}(\underline{x},t)\cdot\underline{\tau}]}{C_o(\underline{x}(s),t)^2} ds$$

Taking the difference, $T'_i - T'_j$:

$$T'_i - T'_j = 2\cdot\int_{\Gamma_{oi}} \frac{[\underline{u}(\underline{x},t)\cdot\underline{\tau}]}{C_o(\underline{x}(s),t)^2} ds \qquad (23)$$

and the sum:

$$T'_i + T'_j = 2\cdot\int_{\Gamma_{oi}} \frac{[C'(\underline{x}(s),t)]}{C_o(\underline{x}(s),t)^2} ds \qquad (24)$$

This shows analytically how the use of transceivers instead of single sources or receivers will greatly improve the current resolving power of the acoustic data without adding extra moorings. For a more comprehensive discussion, see Worcester and Cornuelle, (1982), which evaluates the utility of tomography as a current measurement tool. Reciprocal transmissions do not present any special problems in the data processing or inverse techniques outlined below.

2.6 NON-LINEARITY

If the perturbation field calculated by the inverse, $C'(\underline{x},t)$, is large, then (20) may no longer hold accurately, and it is necessary to iterate by choosing a new reference state,

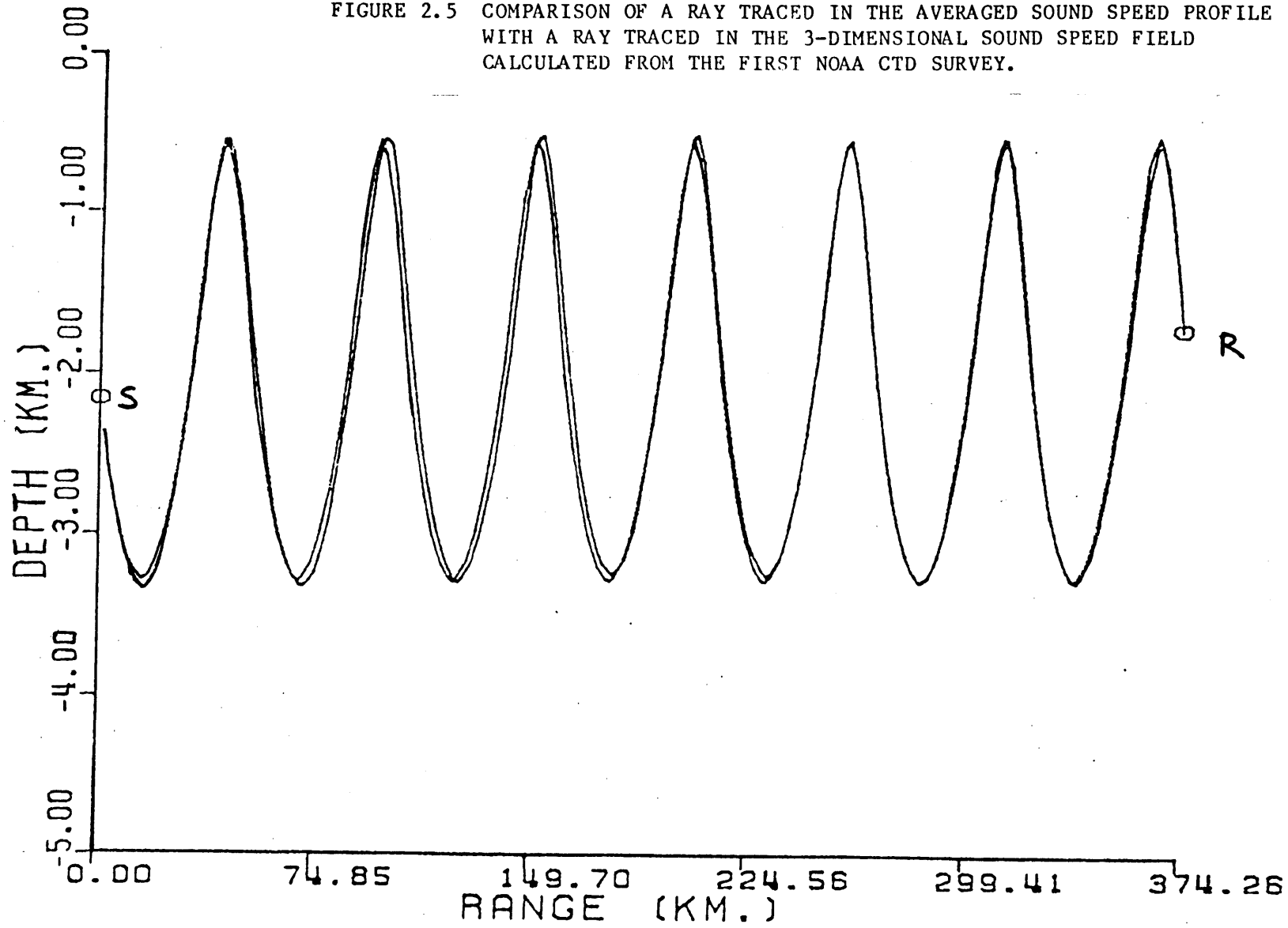
$$C_1(\underline{x},t) = C_o(\underline{x},t) + C'(\underline{x},t) \qquad (25)$$

presumably closer to the true field, $C(\underline{x},t)$, than $C_o(\underline{x},t)$ was. Such iteration is necessary when the assumptions which led to (20) become invalid. The travel time calculations are not as sensitive to the size of $C'(\underline{x},t)$ as the detailed ray path is, since the path deformation has little effect on the travel time calculation (Hamilton, et. al. (1980)). This means that an important criterion for deciding when iteration is required comes from the inversion, not the forward problem. Difficulties will occur when the ray paths are deformed by amounts significant on the scale of the oceanic structures under study.

One can estimate, for the mesoscale experiment described in detail below, that problems will begin to be felt when the perturbed (true) ray path Γ_i and the

unperturbed path, Γ_{oi}, differ by more than O(100 m) vertically or O(5 km) horizontally for a significant fraction [O(10%)] of the range. This estimate is not rigorous, and is given purely to fix ideas; the perturbations observed in the MODE experiment and the 1981 OAT experiment were not sufficient to perturb the ray paths appreciably (see Figure 2.5), so careful numerical calculations of sensitivity have not been made. Since Γ_i is unknown, linearity can be checked by tracing rays in the sound speed field estimated by the inverse, and comparing those paths to the original paths, Γ_{oi}. If these paths differ significantly from the ray paths used in the inversion, then iteration is probably necessary. The convergence of these iteration methods depends on the error and resolving power of the inversion and the linearity of the forward problem, but given adequate resolving power, the robust linearity of the forward problem should lead to rapid convergence.

FIGURE 2.5 COMPARISON OF A RAY TRACED IN THE AVERAGED SOUND SPEED PROFILE WITH A RAY TRACED IN THE 3-DIMENSIONAL SOUND SPEED FIELD CALCULATED FROM THE FIRST NOAA CTD SURVEY.



2.7 RAY IDENTIFICATION

The identification of the rays may present the most difficult problem when strong perturbations are introduced. In order to use the acoustic data in an inversion, the travel times observed in the data must be matched to ray paths traced by the computer and used in the construction of the inverse operator. For example, the latest peak in an arrival pattern may be found to correspond to a +12 RR ray, the next-to-last arrival may be the -11 RR ray, and so on. The ray identifier labels a ray path stored in the computer, which determines how the ray samples the ocean, and is therefore necessary for the calculation of the inverse operator. The process of arriving at the proper match-ups is called "ray identification".

Both the "pulse" arrivals observed in the data and the travel times calculated numerically form patterns (see Figure 2.6), and, provided the differences between the sound speed fields in the two cases are small enough, the two patterns will be comparable. One can then select out observed arrivals which correspond to numerically traced rays. The arrival times for individual rays change nearly linearly with increasing strength of the perturbation but at different rates, so that the overall arrival pattern deforms. The structure of these patterns is an important part of the criteria used to match each observed ray

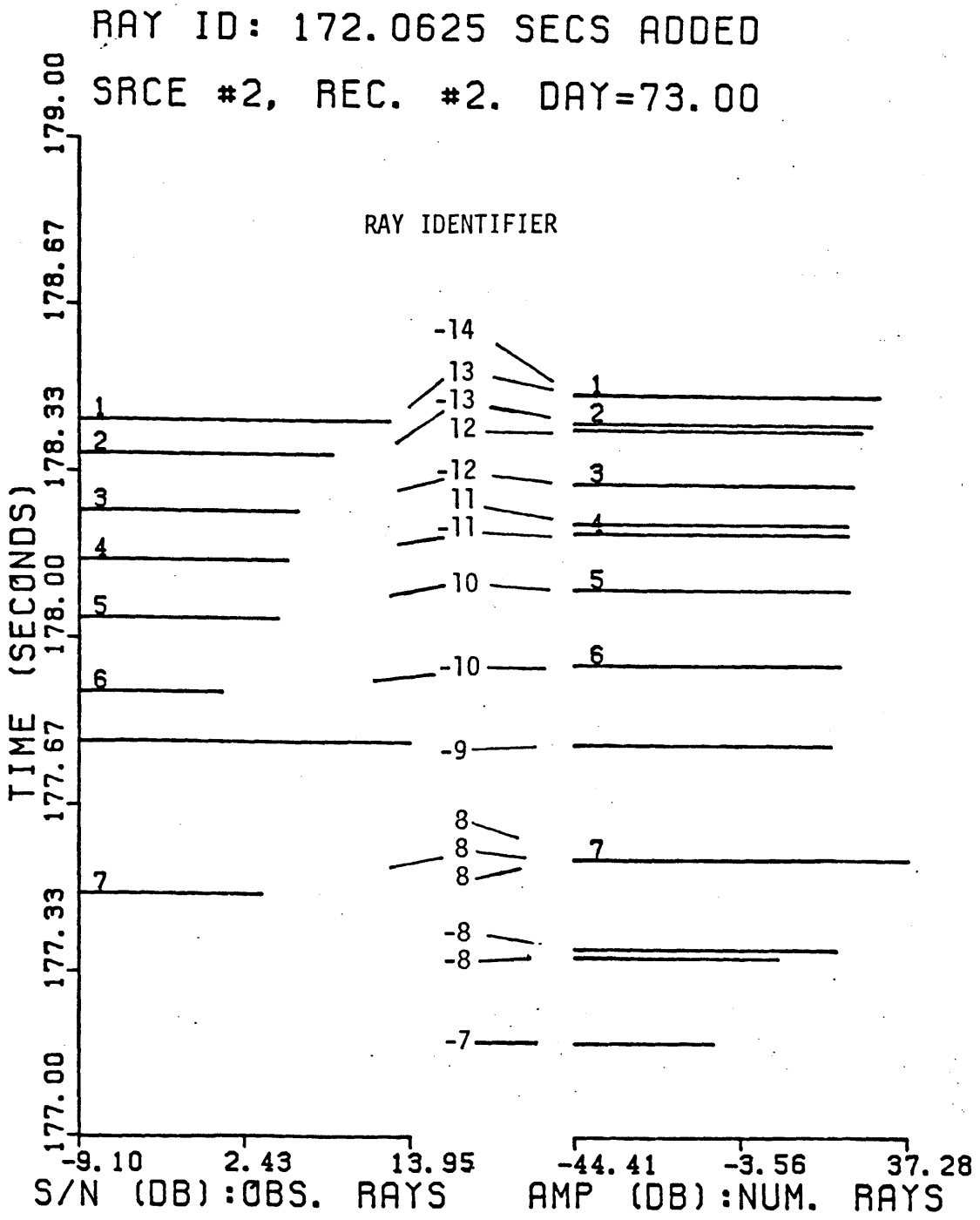


FIGURE 2.6 PLOT OF A RAY IDENTIFICATION SHOWING PATTERN MATCHING BETWEEN NUMERICALLY TRACED RAYS (LEFT) AND OBSERVED RAYS (RIGHT). THE SMALL NUMBERS ON THE LINES MARK RAYS WHICH MATCH.

arrival with the correct ray path, a process called "ray identification". The most stringent bound on the size of the perturbations allowed without iteration could, then, come from the ability to make correct identification. Vertical arrays of hydrophones, such as employed in the receivers constructed by Peter Worcester at Scripps (Worcester 1981) add arrival angle information to the travel time data, improving both the resolution of the receiver and the reliability of the identification. Once again, for the 1981 tomography experiment, pattern shifts were never extreme enough to require re-identification, particularly given the continuity of the arrival pattern over the 3-day sampling interval, which was short compared to the 30 day mesoscale evolution timescale (Figure 1.6).

It was this continuity of ray travel time patterns between a fixed source and receiver over weeks and months that first demonstrated the practicality of acoustic tomography. "Traditional" ocean acoustics had until the late 1960's concentrated on intensity measurements ("propagation loss") for continuous wave sources. The travel time measurements, corresponding to phase information in the CW case, were thought to be too unstable to hold useful information. Landmark experiments using

equipment and techniques developed by Spindel and Webb demonstrated the stability of the pulse arrival pattern over long periods, as predicted by MW. As a result of the original tomography proposal, Spindel, T. Birdsall, and K. Metzger developed sophisticated signal processing to filter out the rapid shifts due to internal waves, leaving the slower changes due to the mesoscale.

2.8 EXTENSIONS OF RAY THEORY: NORMAL MODES

While the ray formulation is simple and useful, it is by no means perfect, and an alternate description of sound propagation involving modes of acoustic pressure has several advantages, and is analytically simple for regions of weak range dependence.

Re-writing (1) for cylindrical coordinates, assuming radial symmetry, and C = C(z) only, yields a separable equation:

$$\nabla^2\phi = \frac{1}{C(z)^2}\frac{\partial^2\phi}{\partial t^2}$$

$$\phi_{rr} + \frac{1}{r}\cdot\phi_r + \phi_{zz} = \frac{1}{C(z)^2}\cdot\phi_{tt} \qquad (26)$$

$$\text{Let} \quad \phi(r,z,t) = \phi_o\cdot R(r)\cdot P(z)\cdot\exp(i\omega t) \qquad (27)$$

Then (16) becomes

$$R''(r) + \frac{R'(r)}{r} + k_h^2\cdot R(r) = 0 \qquad (28)$$

and

$$P''(z) + [(\omega^2/C(z)^2) - k_h^2]\cdot P(z) = 0 \qquad (29)$$

here $-k_h^2$ = separation constant.

Solving (28) with a radiation condition--outgoing waves only (Tolstoy and Clay, 1960):

$$R(r) = H_o^{(1)}(k_h \cdot r) \qquad (30)$$

In the far field, where $k_h \cdot r >> 1$,

$$H_o^{(1)}(k_h \cdot r) \simeq (\pi \cdot k_h \cdot r/2)^{-1/2} \cdot \exp(i[k_h \cdot r + \pi/4]) \qquad (31)$$

k_h may be interpreted as the horizontal wavenumber for the propagation of the modes.

Equation (29) determines the vertical structure of each mode, showing "turning points" at

$$z_T: C(z_T) = \omega/k_h \qquad (32)$$

by analogy with the quantum mechanical problem (Bender and Orszag, 1979).

It may be solved using WKB approximations within each region, or a uniformly valid solution can be obtained using Langer's method (Munk and Wunsch, 1983). Using 2 turning point WKB analysis (Bender and Orszag, 1979) the turning points must satisfy:

$$\int_{z_T-}^{z_T+} [(\omega^2/C(z)^2) - k_h^2]^{1/2} \, dz = (n + 1/2)\pi \qquad (33)$$

From (32) $k_h = \omega/C(z_T^{\mp})$, so (33) becomes

$$\omega \cdot \int_{z_T^-}^{z_T^+} [(1/C(z)^2) - 1/C(z_T^{\mp})^2]^{1/2} \, dz = (n + 1/2)\pi \qquad (34)$$

For fixed n, this yields a dispersion relation, $\omega(k_h)$, because the turning points, $z_T^{\mp}$, are functions of k_h. Equation (34) allows the calculation of horizontal group velocity for mode n:

$$C_g = \frac{\partial \omega}{\partial k_h} \qquad (35)$$

From the expression for group velocity, one can calculate the arrival time of a given mode n with frequency ω as a functional of the C(z) field, providing an alternate form of the forward problem for the modes. Although modes and rays are theoretically interchangeable expressions for the acoustic pressure field, there are cases where mode arrivals may be resolved while ray arrivals cluster too closely, so that a complete extraction of information could use both ray and mode arrival data (Munk and Wunsch, 1982a). At present, only ray arrivals have been used, but modes are to be investigated further in later experiments.

CHAPTER 3

THE QUASI-GEOSTROPHIC APPROXIMATION

3.1 BASIC ASSUMPTIONS

The oceans support motions with a rich range of space and time scales, from acoustic waves at the small scales to the thermohaline circulation, which extends over all the oceans, and evolves on time scales of years to centuries. A large share of the observed energy belongs to a band of motion between these extremes, the "mesoscale". Most of the kinetic energy observed by current meters results from these motions, and they have therefore been of great interest to oceanographers during the past decade.

The theory describing these motions is now quite well-developed, and there are several datasets which give specific realizations of the ocean on adequate space and time scales. Mesoscale features have length scales of order 100 km (O(100 km): meaning between 10 km and 1000 km) current speeds of O(10 cm/sec), and time scales of O(50 days). Non-dimensionalizing the Navier-Stokes equation based on these scales and dropping small terms leads to the quasigeostrophic equations, which are used here in a form based on several other assumptions:

1) The area being modelled is small enough so that the spherical earth can be described locally by cartesian coordinates, leaving the meridional variation of the Coriolis parameter as the only remaining effect of sphericity:

$$f = f_o + \beta_o y \qquad (1)$$

where Θ_o = latitude at which the coordinate system is centered Ω = earth's rotation rate, $f_o = 2\Omega\sin\theta_o$, and $\beta_o = 2\Omega\cos\theta_o/R_e$.

2) The dynamics of interest are perturbations to a motionless rest state in which the ocean is locally in hydrostatic equilibrium. Thus, if $p(\underline{x},t)$ = pressure at a point, and $\rho(\underline{x},t)$ is potential density, then

$$p(\underline{x},t) = p_s(z) + p_m(\underline{x},t) \qquad (2)$$

and

$$\rho(\underline{x},t) = \rho_s(z) + \rho_m(\underline{x},t) \qquad (3)$$

where

$$\frac{\partial p_s}{\partial z}(z) = -\rho_s(z)\cdot g \qquad (4)$$

3) p_m and ρ_m are pressure and potential density perturbations due to the presence of mesoscale motion with current velocities $(u,v,w) = \underline{u}$, and these are nearly in geostrophic and hydrostatic equilibrium:

$$\frac{\partial p_m}{\partial z}(x,t) = -\rho_m(x,t) \cdot g \tag{5}$$

$$f_o \cdot u(\underline{x},t) = \frac{-1}{\rho_s(z)} \cdot \frac{\partial p_m}{\partial y}(\underline{x},t) \tag{6}$$

$$f_o \cdot v(\underline{x},t) = \frac{1}{\rho_s(z)} \cdot \frac{\partial p_m}{\partial x}(\underline{x},t) \tag{7}$$

This final assumption has been examined empirically using some of the datasets mentioned above, notably by the MODE Group (1976), and seems to hold to within experimental error.

Using this basis, Pedlosky (1979) develops the quasigeostrophic approximation rigorously, and I will use the result of his analyses to build theoretical relationships between many of the variables which may be considered as part of the forward or inverse problem.

3.2 DESCRIPTION OF THE (MESOSCALE) PERTURBATION FIELDS

Define a streamfunction:

$$\Psi(\underline{x},t) = p_m(\underline{x},t)/\rho_s(z) \tag{8}$$

as the basic quantity from which other quantities may be derived on the basis of the theory. For instance,

$$\rho_m = -\frac{\rho_s(z)}{g} \cdot \frac{\partial \Psi}{\partial z} \tag{9}$$

$$f_o \cdot v = \frac{\partial \Psi}{\partial x} \tag{10}$$

$$f_o \cdot u = -\frac{\partial \Psi}{\partial y} \tag{11}$$

$$w = \frac{-1}{N^2(z)} \cdot \frac{\partial^2 \Psi}{\partial t \partial z} \tag{12}$$

Here N(z) is buoyancy frequency.

The quasigeostrophic theory yields a dynamic equation for predicting the evolution of these fields which expresses conservation of potential vorticity along fluid trajectories in the absence of viscosity or heating:

$$\left[\frac{\partial}{\partial t} + u\cdot\frac{\partial}{\partial x} + v\cdot\frac{\partial}{\partial y}\right]\cdot\left[\nabla^2\Psi + \frac{\partial}{\partial z}\cdot\left(\frac{f_o^2}{N^2(z)}\cdot\frac{\partial \Psi}{\partial z}\right)\right] + \beta_o\cdot\frac{\partial \Psi}{\partial x} = 0 \tag{13}$$

With boundary conditions

$$w = 0 \text{ at } z = -D \quad \text{(flat bottom)} \tag{14}$$

$$w = \frac{1}{g}\cdot\frac{\partial \Psi}{\partial t} \text{ at } z = 0 \quad \text{(free surface)} \tag{15}$$

For the scaling we have used, the equation (13) is non-linear to lowest order, but it is useful to linearize it to obtain a set of formal relations between variables. For the linearization, the advective terms are dropped, leaving

$$\left[\frac{\partial}{\partial t} \right] \cdot \left[\nabla^2 \Psi + \frac{\partial}{\partial z} \cdot \frac{f_o^2}{N^2(z)} \cdot \frac{\partial \Psi}{\partial z} \right] + \beta_o \cdot \frac{\partial \Psi}{\partial x} \Big] = 0 \qquad (16)$$

which is separable. Let

$$\Psi(\underline{x},t) = \phi(x,y,t) \cdot G(z) \qquad (17)$$

and split (16) in two parts using α^2 as the separation constant:

$$\frac{\partial}{\partial t}[\nabla^2 \phi - \alpha^2 \phi] + \beta \cdot \frac{\partial \phi}{\partial x} = 0 \qquad (18)$$

$$\frac{d}{dz} \left[\frac{1}{N^2(z)} \frac{dG(z)}{dz} \right] + \alpha^2 \cdot f_o^2 \cdot G(z) = 0 \qquad (19)$$

Let $G'(z) \equiv dG/dz$, and the boundary conditions are

$$G'(z) = 0 \text{ at } z = -D \qquad (20)$$

$$G'(z) + \frac{N^2(z)}{g} \cdot G(z) = 0 \quad \text{at } z = 0 \qquad (21)$$

The system (19), (20), (21) can be transformed by letting

$$\lambda \equiv \alpha^2 \cdot f_o^2 \tag{22}$$

$$\text{and } G\zeta(z) \equiv G'(z)/N^2(z) \tag{23}$$

The $G\zeta(z)$ modes will later be seen to correspond to vertical displacement of water. The system then becomes:

$$G''\zeta(z) = -\lambda \cdot N^2(z) \cdot G\zeta(z) \tag{24}$$

$$G\zeta(z) = 0 \qquad z = -D \tag{25}$$

$$G\zeta(z) - \frac{1}{g\lambda} G'\zeta(z) = 0 \quad z = 0 \tag{26}$$

Equation (26) makes use of the relation

$$G(z) = -\frac{1}{\lambda} G'\zeta(z) \tag{27}$$

The eigenvalue problem (24), (25), (26), may be solved numerically for any $N^2(z)$ profile, yielding a complete set of basis functions, $G\zeta_i(z)$, each with eigenvalue λ_i. $G_i(z)$ may be derived from $G\zeta_i(z)$ using (27), and we can now express many variables using this combination of horizontal structure and vertical modes. Equations (9)-(12) become:

$$\rho_m(\underline{x},t) = -\frac{\rho_s(z)}{g}\cdot N^2(z)\cdot\sum_{i=1}^{M} G\zeta_i(z)\cdot\phi_i(x,y,t)$$

$$= -\rho_s'(z)\cdot\sum_{i=1}^{M} G\zeta_i(z)\cdot\phi_i(x,y,t) \qquad (28)$$

$$v(\underline{x},t) = \frac{1}{f_o}\cdot\sum_{i=0}^{M} G_i(z)\cdot\phi_{ix}(x,y,t) \qquad (29)$$

$$u(\underline{x},t) = -\frac{1}{f_o}\cdot\sum_{i=0}^{M} G_i(z)\cdot\phi_{iy}(x,y,t) \qquad (30)$$

$$w(\underline{x},t) = \sum_{i=1}^{M} G\zeta_i(z)\cdot\phi_{it}(x,y,t) \qquad (31)$$

The set $\{G\zeta_i(z)\}$ corresponds to vertical displacement of water by the mesoscale motions, while the set $\{G_i(z)\}$ is a basis for the pressure, velocity, and streamfunction. In the transformation from equations (19,20,21) to equations (24,25,26), one solution of the original set became trivial and was discarded. If the free surface boundary condition is exchanged for that of a rigid lid (w=0 at z=0), or if a mixed layer exists at the surface (N=0 at z=0), then the boundary condition (21) becomes:

$$G'(z) = 0 \quad \text{at } z = 0 \qquad (21')$$

The set (19,20,21') has a solution $G_o(z)$ = constant, $\lambda=0$, which is a trivial solution of (24,25,26) and cannot be used in equation (27). This mode, $G_o(z)=B$, is often referred to as the "barotropic" velocity mode, because it is depth independent. Thus, for every $i > 1$, $G_i{}^{\zeta}(z)$ corresponds to some $G_i(z)$, but $G_o(z)$ corresponds to $G_o{}^{\zeta}(z)=0$, so the "velocity" or "streamfunction" modes are summed on i=0 to M, while the displacement modes need only be summed on i=1 to M. This means that the density field provides no information about the amplitude of the "barotropic" velocity mode, which has been a source of painful indeterminacy for generations of oceanographers.

Other quantities of interest may be derived in the same manner. For example, eastward transport through the meridional rectangular region defined horizontally by (x_1,y_1) to (x_1,y_2) and vertically between z_1 and z_2 is

$$U(x_1,y_1,y_2,z_1,z_2,t) = \int_{(x_1,y_1)}^{(x_1,y_2)} \int_{z_1}^{z_2} u(x,t)\cdot dy\cdot dz \qquad (32)$$

$$= \frac{1}{f_o}\cdot \sum_{i=1}^{M} \frac{-1}{\lambda_i} \left(\phi_i(x_1,y_2,t) - \phi_i(x_1,y_1,t) \right)\left(G^{\zeta}{}_i(z_2) - G^{\zeta}{}_i(z_1)\right)$$

$$+ \frac{1}{f_o}\cdot[\phi_o(x_1,y_2,t) - \phi_o(x_1,y_1,t)]\cdot B\cdot(z_2-z_1) \qquad (33)$$

The interrelations greatly simplify the inverse procedure. Instead of estimating ρ, u, v, w, transport, or streamfunction separately, the problem can be divided into estimating ϕ, ϕ_x, ϕ_y, and ϕ_t, greatly reducing the amount of work. Naturally, adopting this framework is most useful if the analytical modes $G_i(z)$ and $G^{\zeta}{}_i(z)$ form an efficient basis, so that only a small number of modes are needed to describe most of the features observed in the ocean. On the other hand, the assumptions involved are no stricter than those normally employed by dynamic oceanography, and should not result in inconsistencies within the inversions. In addition, the modes do not need to be orthogonal to be used in the inversion - the only complication introduced by non-orthogonality comes in computing expected energies.

3.3 NON-DYNAMICAL MODE BASES

It is sometimes desireable to use some other set of modes as a vertical basis in place of the analytical modes. In this case, the analysis described above would still hold, except for the analytical transformation between the velocity modes and the displacement modes. Since an arbitrary set of modes will not be a solution set for the vertical structure equation (19) or (24), equation (27) will no longer apply.

Suppose, for example, that a set of basis functions for the vertical structure of the density field have been obtained: $\{F\rho_i(z)\}$. These may be empirical orthogonal functions (E.O.F.s) derived from data, or may be completely arbitrary, describing layers or some other pre-defined vertical structures.

The density perturbations, $\rho_m(\underline{x},t)$ are still assumed to be in quasi-geostrophic equilibrium with the other fields, and the linearity of the equations makes superposition hold, so let $\eta_i(x,y,t)$ be the horizontal structure of mode i,

$$\rho_m(\underline{x},t) = \sum_{i=1}^{M} F\rho_i(z)\cdot\eta_i(x,y,t) \qquad (34)$$

The density perturbations are produced by the vertical motions of water acting on the adiabatic density gradient, so displacement modes are given by

$$F\zeta_i(z) = (\rho_s'(z))^{-1}\cdot F\rho_i(z) = F\rho_i(z)/(d\rho_s/dz) \qquad (35)$$

or

$$F^{\zeta}{}_i(z) = \frac{-g}{\rho_s(z)N^2(z)} \cdot F^{\rho}{}_i(z) \qquad (36)$$

The two forms (35) and (36) are not necessarily equivalent when numerically calculated because the derivative in (35) must be the local adiabatic gradient of potential density, not just the simple derivative, particularly if ρ is potential density relative to the surface. Calculations of N^2 must also take this derivative properly, in order to avoid false regions of apparent instability, so the form (36) is often easier to implement. In general, whenever vertical derivatives appear, it is important that they locally remove pressure effects, to avoid bias from non-linearities. These considerations are necessary when converting to and from temperature, potential temperature, and sound speed.

Sound speed modes must always be computed from an empirical relation like (35), where $C_s(z)$ is the basic state sound speed:

$$F^{c}{}_i(z) = \left(\frac{dC_s(z)}{dz}\right)_{\text{potential}} \cdot F^{\zeta}{}_i(z) \qquad (37)$$

Similar relations hold for temperature, potential temperature, salinity, and the tracers, whether the set of F(z)'s are analytical, empirical, or arbitrary.

Equation (27) relates displacement modes to velocity modes without resorting to a reference level assumption, because the indeterminacy of barotropic velocity given density measurements showed up as the lack of constraints on $\phi_o(x,y,t)$, the amplitude of $G_o(z)$, the vertically uniform analytical mode of horizontal velocity.

The indeterminacy thus has a clear dynamic meaning as the amplitude of the barotropic mode. Analytical or numerical estimates of energy missed in this way can be made. When non-analytic modes are used, the "reference level" problem is more complex. In order to convert from displacement to velocity, we must use equation (23) and then integrate vertically to find $F_i(z)$, the i^{th} empirical velocity mode.

$$F_i(z) = \int_{z_o}^{z} N^2(z')F\zeta_i(z')dz' + F_i(z_o) \qquad (38)$$

$F_i(z_o)$ is unknown, and corresponds to the "reference level" velocity (with the reference level at z_o). Any set of displacement modes $F\zeta_i(z)$, i = 1 to M, can be used with equation (38) to generate a set of velocity modes $F_i(z)$. $F_o(z)$ is a uniform velocity, as before, but the energy in this mode depends largely on the reference levels z_o picked for each mode.

The description of velocity still has the simple form:

$$u(\underline{x},t) = -\frac{1}{f_o} \cdot \sum_{i=0}^{M} F_i(z) \cdot n_{iy}(x,y,t) \qquad (39)$$

$$v(\underline{x},t) = \frac{1}{f_o} \cdot \sum_{i=0}^{M} F_i(z) \cdot n_{iy}(x,y,t) \qquad (40)$$

The empirical functions $F_i(z)$ can generally be picked to be a more efficient basis for the perturbation field than the analytic function, $G_i(z)$, but they require more prior information than the analytical modes, and suffer from the reference level problem. Using the analytical modes as a basis also allows the use of the equivalent barotropic equation (Flierl, 1978) to add linear or nonlinear dynamics into the models, and eventually, into the inversions. The EOFs, on the other hand, do not provide an efficient "state vector" for the quasi-geostrophic dynamical equations, so the models based on EOFs are practically limited to employ diagnostic constraints only, while models based on analytical modes may use the prognostic equations, such as vorticity conservation.

CHAPTER 4

PROBABILISTIC ESTIMATION

4.1 GENERAL DISCUSSION

Consider a general estimation problem, where N data, $\{d_i: \ i=1,N\}$ are taken, and an estimate of some field $\Psi(\underline{x},t)$ is desired. The data are only useful if they depend on ("sample") Ψ in some way, which may or may not be deterministic. In vector notation, this is written:

$$\underline{d} = \underline{F}(\Psi(\underline{x},t),\underline{x},t) \qquad (1)$$

The problem posed in this chapter is how best to invert this relation (1) in order to obtain the best possible estimate of $\Psi(\underline{x},t)$, given the data $\underline{d} = \{d_i\}$. The full inversion problem for tomography requires this generality, since the data may consist of many types, and the desired output field may not appear explicitly in the forward problem. For example, in the 1981 Tomography experiment, the full data set consists of travel times, travel time differences between rays in an arrival pattern (called "ray differentials"), temperature, pressure, and current records from moored instruments, and CTD stations taken during 3 surveys. The desired output fields also encompass a wide range, including sound speed, velocity, temperature, density, transport, heat content, and perhaps vorticity.

In a standard moored experiment of the past, the instruments directly measured one quantity, such as horizontal velocity, at several points in space. The results were Fourier transformed in time to yield an estimate of the time scales important in the motions, and covariances between instruments were calculated to yield estimates of spatial scales, and, with lagged covariances, propagation velocities. More recently, optimal estimation techniques were employed to yield continuous maps of the quantities measured only at points (Bretherton, et al. 1973), and, much more recently, to yield estimates of a quantity, vorticity, (McWilliams, 1976), (Hua and Owens, 1982) not directly measured.

It is a small (and logical) step to generalize entirely, so that a wide variety of measurements made at different space and time locations could be combined by one, as yet unspecified, estimation procedure, to yield the estimates of desired output quantities at any space and time locations which can be shown to be the "best", given the criteria necessary to define "best". The objective analysis mentioned earlier is thus a special case of one estimation scheme where the criteria for "best" consist of linearity and minimum expected squared difference between the true field and the mapping field, given an assumption of a statistical ensemble.

The ingredients of any estimation method will generally be:

1) A constraint on the estimator, such as linearity in the data.

2) Criteria to define a figure of merit for the estimator, such as the weighted sum of absolute values of the results and/or the residuals. These criteria generally will require choosing the framework in which the calculations take place, such as a choice between deterministic and statistical calculations.

3) A set of assumptions about the various quantities involved in the estimation procedure. These assumptions include the "forward problem," which relates the data taken to the quantities which may affect it; as well as error estimates and models for the unknown fields.

In this chapter, I will consider a number of methods for arriving at estimates of the output fields, and discuss their features in a framework that is not specific to the tomography experiment, but applies generally to problems of inferences from data within a physical framework. Readers primarily interested in the results of the inverses applied to the 1981 tomography experiment may wish to skip to chapter (8) or (9).

4.2 ESTIMATION BASED ON PROBABILITY DISTRIBUTIONS

One very general framework of estimation theory is well discussed in the electrical engineering literature. It uses the concept of information pioneered by Shannon (1948), and many specific estimators are special cases of this approach. One standard text is Van Trees (1968), but the subject has recently been broached in the geophysical literature (Tarantola and Valette, 1982b). The theory is too complex to make it worthwhile to carry it completely through in an example, but a brief discussion is worthwhile as the theory provides an organized background out of which various specific estimators may be derived. I will use the notation of Tarantola and Valette (1982b).

Let $\underline{d}$ = vector of data values, and $\underline{p}$ = vector of parameter values. These may be countably infinite in length, which means they can represent continuous systems, given the discretization due to computers and minimum scales of interest. These vectors are combined into one vector, $\underline{x}$, of length m, where every element of $\underline{x}$ has a probability distribution, $f_i(x_i)$, describing the likelihood with which it can take on any given value. In the case of the data, this probability describes the possible deviation of the true value from the recorded value. Thus, when an experimenter records only a data value, d_o and a standard deviation due to error, σ_o, but no other error moments,

this is consistent with the assumption that the probability distribution function for the true value, d, of the quantity measured is

$$f(d) = (2\cdot\pi\cdot\sigma_o^2)^{-1/2}\cdot\exp[-(d-d_o)^2/2\sigma_o^2] \qquad (2)$$

Before the experiment takes place, there is a joint probability distribution for the model parameters, the a priori distribution, which contains all information about the parameter values independent of the data taken. If these a priori expectations are combined with the data probability density function (p.d.f.'s), then we can write

$$\rho(\underline{x}) = \rho(\underline{p},\underline{d})$$

the m-dimensional a priori p.d.f., which is input to the inverse procedure. The other essential ingredient of this general framework is the relation between the data and the model. This can be expressed within the theory as a joint p.d.f.; $\Theta(\underline{p},\underline{d})$, where $\underline{d}$ and $\underline{p}$ are not independent. If $\underline{d}$ and $\underline{p}$ are independent, so that

$$\theta(\underline{p},\underline{d}) = \theta_p(\underline{p})\cdot\Theta_d(\underline{d}), \qquad (3)$$

then it can be shown that it did no good to collect the particular set of data, $\underline{d}$. In a practical application, the model and data will be related, so the distribution will not be separable.

For example, a deterministic relation between model parameters and data can be written in non-linear functional form:

$$\underline{d} = \underline{G}(\underline{p}) \qquad (4)$$

This can be expressed in probability form:

$$\theta(\underline{p},\underline{d}) = \delta(\underline{d} - \underline{G}(\underline{p}))\cdot\mu(\underline{p})$$

where $\mu(\underline{x})$ is a p.d.f. which reflects the state of null information about $\underline{p}$ and $\delta(\ )$ is the Dirac delta function.

The state of null information, which has been called $\mu(\underline{x})$ after Tarantola, is a concept used to streamline the construction of the conditional probability density functions necessary for the estimation procedure. $\mu(\underline{x})$ is a p.d.f. for the data and model parameters which can be constructed without any knowledge. The simplest example of $\mu(\underline{x})$ would be jointly independent uniform distributions between $\pm\infty$, although other forms may be possible or preferable (see Tarantola and Valette, 1982b).

If the theoretical information, $\theta(\underline{x})$, is independent of the a priori model and data, $\rho(\underline{x})$, then they can be combined simply to obtain the a posteriori state of information:

$$\sigma(\underline{x}) = \rho(\underline{x})\cdot\theta(\underline{x})/\mu(\underline{x}) \quad . \qquad (6)$$

This a posteriori set of p.d.f.'s may then be operated on in a variety of ways to obtain the results desired. For example, the estimated parameter values, $\hat{\underline{p}}$, may be picked such that

$$\sigma_p(\underline{p}) \quad = \quad \text{maximum at } \underline{p} = \hat{\underline{p}},$$

This is the maximum likelihood estimator, and is frequently used, primarily for its simplicity, although the statistically rigorous estimator for an arbitrary p.d.f. would be the center of mass,

$$\hat{\underline{p}} = \langle \underline{p} \rangle = \int \underline{p} \cdot \sigma_p(\underline{p}) d\underline{p} \qquad (8)$$

It is possible to show (Tarantola and Valette, 1982a, this thesis), that when the assumed probability density functions are Gaussian, then the maximum likelihood estimator is linear, and corresponds to the least-square error estimator. In fact, for Gaussian distributions, the maximum likelihood estimate is the same as the expected value, so that it is statistically rigorous. If the distributions are not Gaussian, then the computations may be more difficult, but the formulation still applies, although the maximum likelihood estimator no longer necessarily even has simplicity to recommend it.

This theory can be generalized to allow cases where the constraints and data are not conveniently expressible as probability distributions. The quantity called 'Information', defined by Shannon (1948) is, for a probability distribution function $f(\underline{x})$,

$$I(\underline{x}) = \log[1/f(\underline{x})] \qquad (9)$$

$I(\underline{x})$ represents the amount of information that we gain from an observation of the random process $\underline{X} = \underline{x}$, and the expected value of information defined with e as a base is equal to the entropy, as defined in statistical physics,

$$\langle I \rangle = E = -\int f(\underline{x}) \ln[f(\underline{x})]\, d\underline{x} \qquad (10)$$

Maximum expected information thus corresponds to maximum entropy, and is the state where the probability function is as smooth as is consistent with the constraints of a priori knowledge and the data.

For example, if a random scalar variable x has an unknown p.d.f. f(x), but is known to be non-negative and to have mean μ, then the maximum entropy f(x) is (Papoulis, 1981)

$$f(x) = (1/\mu) \exp(-x/\mu) \qquad (11)$$

On the other hand, if only the mean μ and variance σ^2 of x is known, then the maximum entropy distribution is

$$f(x) = (2\pi\sigma^2)^{-1/2} \cdot \exp[-(x-\mu)^2/2\sigma^2] \qquad (12)$$

(Shannon, 1948). The Gaussian probability assumption is thus somewhat justifiable from a maximum entropy standpoint, given no higher moments or extra constraints.

Unfortunately, there is a present paucity of oceanographic data, precluding accurate statistics, not to mention specification of probability distribution. In the absence of such data, assuming the unknowns to be the result of a Gaussian random process would seem to have some basis, if only as an heuristic consequence of the central limit theorem. Thus, the least-square estimators may be used without committing gross errors by assumption, and they are convenient as well.

4.3 OPTIMAL ESTIMATES FOR GAUSSIAN DISTRIBUTIONS

The probabilistic discussion given above may seem abstract, but it is instructive to apply it to an example which has often been treated by standard inverse methods. Suppose that we wish to estimate an unknown field, $p(\underline{x},t)$, given a data set $\underline{\tilde{d}}^T = (\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N)$ containing random observation error, $\underline{\varepsilon}$, normally distributed with known covariance, so that the true value $\underline{d} = \underline{\tilde{d}} - \underline{\varepsilon}$. Assume, in addition, that $p(\underline{x},t)$ is normally distributed around an independently derived value, $\tilde{p}(\underline{x},t)$. Then we can form the prior probability density $\rho(\underline{\lambda})$,

$$\rho(\underline{\lambda}) = \gamma \cdot \exp[-1/2(\underline{\lambda}-\underline{\tilde{\lambda}})^T \underline{\underline{C}}_a^{-1}(\underline{\lambda}-\underline{\tilde{\lambda}})] \qquad (1)$$

where $\underline{\lambda}^T = \{p(\underline{x},t), \underline{d}^T\}$, and γ is a normalization factor to make $\rho(\underline{\lambda})$ a probability density function, and $\underline{\underline{C}}_a$ reflects the uncertainty of both the model and data,

$$\underline{\underline{C}}_a = \begin{matrix} \alpha & 0\ 0\ 0 \\ 0 & \\ 0 & \langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle \\ 0 & \end{matrix} \qquad (2)$$

$$\underline{\underline{C}}_a^{-1} = \begin{matrix} \alpha^{-1} & 0\ 0\ 0 \\ 0 & \\ 0 & \langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle^{-1} \\ 0 & \end{matrix} \qquad (3)$$

α is the expected variance of $p(\underline{x},t)$ around $\tilde{p}(\underline{x},t)$, and $\underline{\underline{C}}_\varepsilon = \langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle$ is the error covariance matrix.

Note that the error is uncorrelated with the a priori estimate of $p(\underline{x},t)$. If $\alpha \to \infty$, then we have no starting information about the value of $p(\underline{x},t)$.

We also require the existence of a theoretical or statistical relation between $\underline{d}$ and $p(\underline{x},t)$,

$$\theta(\underline{\lambda}) = \gamma' \exp\{-1/2(\underline{\lambda}-\overline{\underline{\lambda}})^T \underline{\underline{C}}_T^{-1}(\underline{\lambda}-\overline{\underline{\lambda}})\} \qquad (4)$$

$\overline{\underline{\lambda}}^T = \{\overline{p}(\underline{x},t), \overline{d}_1, \overline{d}_2, \ldots, \overline{d}_N\}$ = an estimate of the expected value (mean) of $\underline{\lambda}$, and $\underline{\underline{C}}_T$ is the estimated or assumed theoretical or statistical covariance for $\underline{\lambda}$ around $\overline{\underline{\lambda}}$. $\underline{\underline{C}}_T$ can be safely assumed to be invertible in principle, since the problem is underdetermined. A covariance matrix is positive definite, but some of the eigenvalues may be very small, making the matrix numerically singular. This covariance matrix expresses the expected variation of the true value around the estimate of the mean.
If $\overline{\underline{\lambda}}$ is unknown, or poorly known, this ignorance can be expressed heuristically by increasing the variance around $\overline{\underline{\lambda}}$. Bretherton, Davis, and Fandry (1973) used this technique, setting the variance of $p(\underline{x},t)$ around $\overline{p}(\underline{x},t)$, $\langle[p(\underline{x},t)-\overline{p}(\underline{x},t)]^2\rangle$ to ∞ to allow for an unknown mean (Liebelt (1967) discusses this too).

In real applications something is usually known about the mean, so that a finite variance may be used, but the resulting estimator will be biased if the true mean is different from the $\bar{p}(\underline{x},t)$ assumed (Liebelt, 1967).

$$\langle\hat{p}(\underline{x},t)\rangle \neq \langle p(\underline{x},t)\rangle \text{ if } \langle p(\underline{x},t)\rangle \neq \bar{p}(\underline{x},t)$$

The biased estimator tends to remain closer to the mean specified in advance than an unbiased estimator, so if this technique is to be used to produce an estimate of a mean over the entire length of a data time series, it is preferable (for economic as well as statistical reasons) to average the data before using the estimator, and then revise the estimation procedure to estimate the mean by modifying the covariances. On the other hand, the biased estimator will yield a lower variance of $\hat{p}(\underline{x},t)$, the estimate of $p(\underline{x},t)$, than an unbiased estimator, so a resolution/bias trade-off needs to be examined for each specific problem. For the present, I will retain the means in the expressions as if they were known, although it must be understood that their significance can be adjusted by the variance weighting.

The cross-covariances between $p(\underline{x},t)$ and the data, $\underline{d}$, provide the essential information needed to complete the problem. If $p(\underline{x},t)$ and $\underline{d}$ are independent, then the

cross-correlation terms vanish, and $\underline{d}$ does not constrain $p(\underline{x},t)$. This "forward problem" may be expressed analytically or statistically, and will be discussed later, but for now, just assume that we have estimates of the model-data covariance,

$$\underline{C}_{pd} = \langle [p(\underline{x},t)-\bar{p}(\underline{x},t)][\underline{d}-\bar{\underline{d}}]^T \rangle, \tag{5}$$

the model covariance,

$$C_p = \langle [p(\underline{x},t)-\bar{p}(\underline{x},t)][p(\underline{x},t)-\bar{p}(\underline{x},t)] \rangle, \tag{6}$$

and the data-data covariance, which includes expected modelling error, but not measurement error,

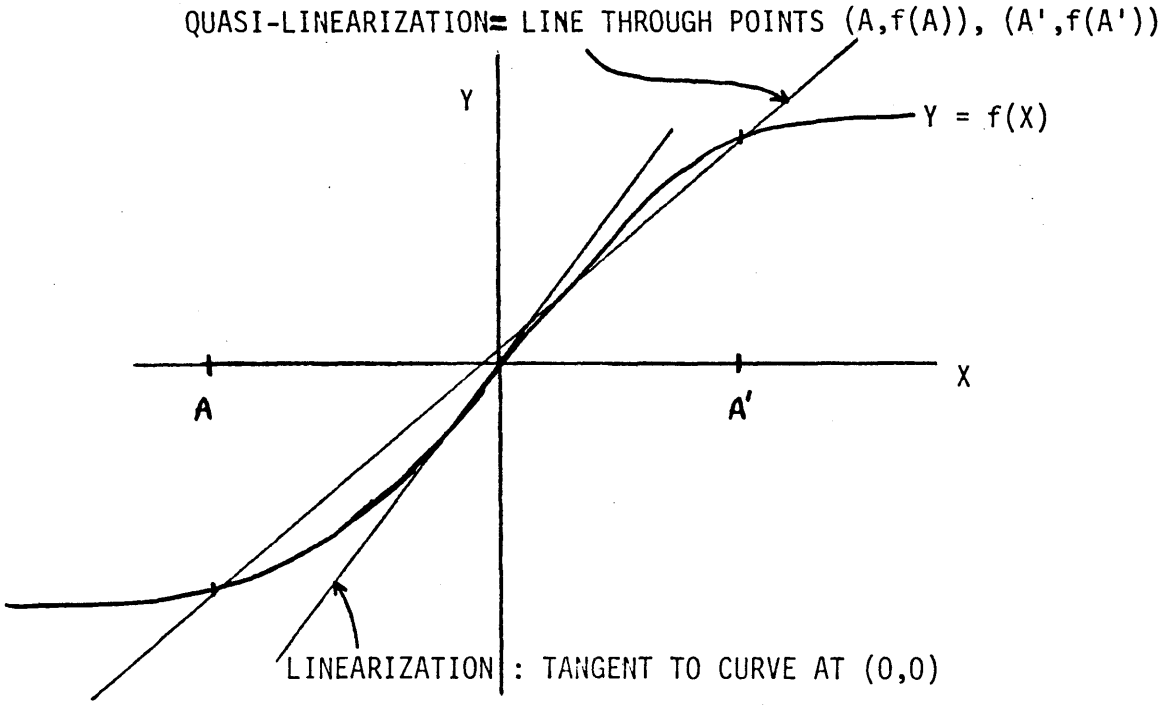
$$\underline{\underline{C}}_d = \langle [\underline{d}-\bar{\underline{d}}][\underline{d}-\bar{\underline{d}}]^T \rangle \tag{7}$$

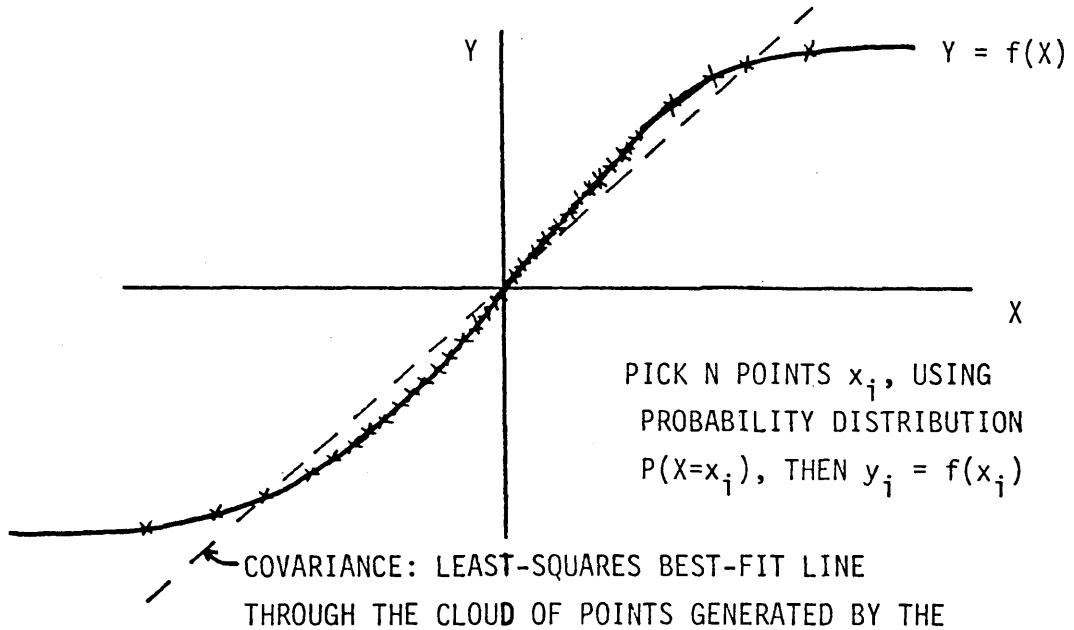
Given these covariance matrices, the total covariance matrix can be written as a partitioned combination of (5),(6), and (7);

$$\underline{\underline{C}}_T = \begin{pmatrix} C_p & \underline{C}_{pd} \\ \underline{C}_{pd}^T & \underline{\underline{C}}_d \end{pmatrix} \tag{8}$$

Note that there have been no explicit assumptions about the linearity or non-linearity of the model-data relation. The covariance form cannot rigorously represent a nonlinear forward problem, but it can express the robust quasi-linearization as used in non-linear control theory. (See Figure 4.1)

FIGURE 4.1 SKETCH TO ILLUSTRATE THE DISTINCTION BETWEEN LINEARIZATION, QUASI-LINEARIZATION, AND COVARIANCE (OR CORRELATION).





If the prior information ($\tilde{\underline{\lambda}}$ and $\underline{\underline{C}}_a$) is independent of the forward problem ($\bar{\underline{\lambda}}$ and $\underline{\underline{C}}_T$), then the posterior probability distribution, $\sigma(\underline{\lambda})$ may be written as the product of the other two distributions:

$$\sigma(\underline{\lambda}) = \gamma'' \cdot \rho(\underline{\lambda}) \cdot \Theta(\underline{\lambda}) \tag{9}$$

where γ'' is another normalization factor.

Using (1) and (4), and letting $\gamma'''=\gamma\cdot\gamma'\cdot\gamma''$, to keep the normalization consistent, we obtain an expression for the a posteriori probability density function for both the data and the unknowns:

$$\sigma(\underline{\lambda}) = \gamma''' \cdot \exp\{-1/2[(\underline{\lambda}-\bar{\underline{\lambda}})^T \underline{\underline{C}}_T^{-1}(\underline{\lambda}-\bar{\underline{\lambda}}) + (\underline{\lambda}-\tilde{\underline{\lambda}})^T \underline{\underline{C}}_a^{-1}(\underline{\lambda}-\tilde{\underline{\lambda}})]\} \tag{10}$$

If $\sigma(\underline{\lambda})$ had the form:

$$\sigma(\underline{\lambda}) \propto \exp[-1/2(\underline{\lambda}-\hat{\underline{\lambda}})^T \hat{\underline{\underline{C}}}^{-1}(\underline{\lambda}-\hat{\underline{\lambda}})] \tag{11}$$

then $\hat{\underline{\lambda}}$ would be the maximum likelihood, minimum variance estimate of $\underline{\lambda}$, and $\hat{\underline{\underline{C}}}$ would be the estimate of its covariance matrix. We can complete the square in (10) to obtain the form (11). Begin by expanding out (10) completely:

$$\sigma(\underline{\lambda}) \propto \exp[-1/2(\underline{\lambda}^T \underline{\underline{C}}_T^{-1}\underline{\lambda} - \underline{\lambda}^T \underline{\underline{C}}_T^{-1}\bar{\underline{\lambda}} - \bar{\underline{\lambda}}^T \underline{\underline{C}}_T^{-1}\underline{\lambda} + \bar{\underline{\lambda}}^T \underline{\underline{C}}_T^{-1}\bar{\underline{\lambda}}$$
$$+ \underline{\lambda}^T \underline{\underline{C}}_a^{-1}\underline{\lambda} - \underline{\lambda}^T \underline{\underline{C}}_a^{-1}\tilde{\underline{\lambda}} - \tilde{\underline{\lambda}}^T \underline{\underline{C}}_a^{-1}\underline{\lambda} + \tilde{\underline{\lambda}}^T \underline{\underline{C}}_a^{-1}\tilde{\underline{\lambda}})] \tag{12}$$

Because the quadratic forms are symmetric:

$$\tilde{\underline{\lambda}}^T \underline{\underline{C}}_a^{-1} \underline{\lambda} = \underline{\lambda}^T \underline{\underline{C}}_a^{-1} \tilde{\underline{\lambda}}$$

and $\tilde{\underline{\lambda}}$ and $\bar{\underline{\lambda}}$ are constants, this can be re-written:

$$\sigma(\underline{\lambda}) \propto \exp(-1/2 \cdot [\underline{\lambda}^T (\underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1}) \underline{\lambda} - 2\underline{\lambda}^T \underline{\underline{C}}_a^{-1} \tilde{\underline{\lambda}} - 2\underline{\lambda}^T \underline{\underline{C}}_T^{-1} \bar{\underline{\lambda}}]) \quad (13)$$

or,

$$\sigma(\underline{\lambda}) \propto \exp(-1/2[\underline{\lambda}^T (\underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1}) \underline{\lambda} - 2\underline{\lambda}^T (\underline{\underline{C}}_a^{-1} \tilde{\underline{\lambda}} + \underline{\underline{C}}_T^{-1} \bar{\underline{\lambda}})]) \quad (14)$$

This can be solved for $\hat{\underline{\lambda}}$ and $\hat{\underline{\underline{C}}}$ using matrix algebra to write (14) in the perfect square form:

$$\hat{\underline{\underline{C}}}^{-1} = \underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1} \quad (15)$$

$$\hat{\underline{\underline{C}}}^{-1} \hat{\underline{\lambda}} = \underline{\underline{C}}_a^{-1} \tilde{\underline{\lambda}} + \underline{\underline{C}}_T^{-1} \bar{\underline{\lambda}} \quad (16)$$

so,

$$\hat{\underline{\lambda}} = (\underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1})^{-1} (\underline{\underline{C}}_a^{-1} \tilde{\underline{\lambda}} + \underline{\underline{C}}_T^{-1} \bar{\underline{\lambda}}) \quad (17)$$

This form could be used for estimation, but it is informative to break the expression down, particularly since $\hat{\underline{\lambda}}$ is primarily an estimate of the true value of the data:

$$\hat{\underline{\lambda}}^T = (\hat{p}(\underline{x},t), \hat{d}_1, \ldots, \hat{d}_N).$$

In addition, as mentioned above, $\underline{\underline{C}}_T$ may have several small eigenvalues, so the inverse may be difficult to obtain numerically. Fortunately, it is possible to modify the expression in (17). Consider the form:

$$\underline{\underline{Q}} = (\underline{\underline{A}}\text{-}1 + \underline{\underline{B}}\text{-}1)\ \underline{\underline{A}}\text{-}1 \tag{18}$$

$\underline{\underline{Q}}$, $\underline{\underline{A}}$, and $\underline{\underline{B}}$ are positive definite (non-singular matrices.

$$\underline{\underline{Q}}^{-1} = \underline{\underline{A}}(\underline{\underline{A}}^{-1}+\underline{\underline{B}}^{-1}) = \underline{\underline{I}} + \underline{\underline{A}}\underline{\underline{B}}^{-1} = (\underline{\underline{B}}+\underline{\underline{A}})\underline{\underline{B}}^{-1} \tag{19}$$

so

$$\underline{\underline{Q}} = \underline{\underline{B}}(\underline{\underline{B}}+\underline{\underline{A}})^{-1} \tag{20}$$

Applying this to (17), we obtain

$$\hat{\underline{\lambda}} = \underline{\underline{C}}_T(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\tilde{\underline{\lambda}} + \underline{\underline{C}}_a(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\bar{\underline{\lambda}} \tag{21}$$

This expression can be simplified further by using the partitioning of $\underline{\underline{C}}_a$ and $\underline{\underline{C}}_T$ as shown above:

$$\underline{\underline{C}}_a = \begin{matrix} \alpha & 0 \\ 0 & \underline{\underline{C}}_\varepsilon \end{matrix} \tag{2}$$

$$\underline{\underline{C}}_T = \begin{matrix} C_p & \underline{C}_{pd} \\ \underline{C}^T_{pd} & \underline{\underline{C}}_d \end{matrix} \tag{8}$$

In order to invert these matrices, we need to take advantage of the partitioning (Liebelt, 1967). If

$$\underline{\underline{Q}} = \begin{matrix} \underline{\underline{A}} & \underline{\underline{B}} \\ \underline{\underline{B}}^T & \underline{\underline{C}} \end{matrix} \qquad (22)$$

then

$$\underline{\underline{Q}}^{-1} = \begin{matrix} (\underline{\underline{A}} - \underline{\underline{B}}\underline{\underline{C}}^{-1}\underline{\underline{B}}^T)^{-1} & -\underline{\underline{A}}^{-1}\underline{\underline{B}}(\underline{\underline{C}} - \underline{\underline{B}}^T\underline{\underline{A}}^{-1}\underline{\underline{B}})^{-1} \\ -\underline{\underline{C}}^{-1}\underline{\underline{B}}^T(\underline{\underline{A}} - \underline{\underline{B}}\underline{\underline{C}}^{-1}\underline{\underline{B}}^T)^{-1} & (\underline{\underline{C}} - \underline{\underline{B}}^T\underline{\underline{A}}^{-1}\underline{\underline{B}})^{-1} \end{matrix} \qquad (23)$$

Using this formula, $(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}$ becomes

$$\begin{pmatrix} \alpha + C_p & \underline{\underline{C}}_{pd} \\ \underline{\underline{C}}^T_{pd} & \underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon \end{pmatrix}^{-1} \equiv \begin{pmatrix} \beta & \underline{\underline{C}}_{pd} \\ \underline{\underline{C}}^T_{pd} & \underline{\underline{C}}_o \end{pmatrix}^{-1} =$$

$$\begin{matrix} (\beta - \underline{\underline{C}}_{pd}\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd})^{-1} & -\beta^{-1}\underline{\underline{C}}_{pd}(\underline{\underline{C}}_o - \underline{\underline{C}}^T_{pd}\beta^{-1}\underline{\underline{C}}_{pd})^{-1} \\ -\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd}(\beta - \underline{\underline{C}}_{pd}\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd})^{-1} & (\underline{\underline{C}}_o - \underline{\underline{C}}^T_{pd}\beta^{-1}\underline{\underline{C}}_{pd})^{-1} \end{matrix} \qquad (24)$$

$$\equiv \begin{matrix} (C_1)^{-1} & -\beta^{-1}\underline{\underline{C}}_{pd}(\underline{\underline{C}}_n)^{-1} \\ -\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd}(C_1)^{-1} & (\underline{\underline{C}}_n)^{-1} \end{matrix} \qquad (25)$$

This formidable expression must be substituted into (21) and multiplied out (see Appendix). To calculate only the a posteriori estimates of the true values of the

unknown field, we need consider only the top row term multiplying the data:

$$\hat{p}(\underline{x},t) = \alpha\beta^{-1}\cdot\underline{\underline{C}}_{pd}\underline{\underline{C}}_n^{-1}(\tilde{\underline{d}} - \bar{\underline{d}}) + [(C_1-\alpha)\bar{p}(\underline{x},t) + \alpha\tilde{p}(\underline{x},t)]C_1^{-1} \quad (26)$$

β, C_o, C_1, and C_n have been implicitly defined above:

$$\beta \equiv \alpha + C_p \quad (27)$$

$$\underline{\underline{C}}_o \equiv \underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon \quad (28)$$

$$C_1 \equiv \beta - \underline{\underline{C}}_{pd}\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd} \quad (29)$$

$$\underline{\underline{C}}_n \equiv \underline{\underline{C}}_o - \underline{\underline{C}}^T_{pd}\beta^{-1}\underline{\underline{C}}_{pd} \quad (30)$$

If $\alpha\to\infty$ (no a priori information about $p(\underline{x},t)$):

$$\hat{p}(\underline{x},t) = \underline{\underline{C}}_{pd}\underline{\underline{C}}_o^{-1}(\tilde{\underline{d}} - \bar{\underline{d}}) + \bar{p}(\underline{x},t) \quad (31)$$

This form will be obtained later using the Gauss-Markov theorem, but the result here proves this form to be the minimum variance non-linear estimator, provided the probability density functions are gaussian.

The optimal estimate of the data values may not seem directly useful, but it is important in calculating the validity of the assumptions built into the inverse. Using the algebra in the Appendix, the a posteriori estimates of the data values can be obtained directly, or the noise in the data can be estimated using equation (21) of the

Appendix. The estimates must then be compared with the prior expectations on which the estimator was built, as a check on consistency within the inverse framework. The estimates of data errors are called "residuals", and should be examined for clues to improper energy levels or missing physics.

4.4 PUTTING ERROR BARS ON THE ESTIMATES

Because the probabilistic estimation method calculates a distribution for the true value, $\underline{\lambda}$, around $\hat{\underline{\lambda}}$, it provides the error covariance, $\hat{\underline{\underline{C}}}$, for the estimate (see Appendix). Again, at present consider only the scalar term describing the variance of $p(\underline{x},t)$ around $\hat{p}(\underline{x},t)$:

$$E_p^2 = \langle [p(\underline{x},t) - \hat{p}(\underline{x},t)]^2 \rangle \tag{32}$$

$$= \alpha \cdot (C_p - \underline{C}_{pd}\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd}) \cdot (C_p + \alpha - \underline{C}_{pd}\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd})^{-1} \tag{33}$$

If $\alpha \to 0$, so that $p(\underline{x},t)$ is known perfectly in advance, then $E_p^2 \to 0$ as well. If $\alpha \to \infty$, so that nothing is known a priori about the true value of $p(\underline{x},t)$, then (33) becomes

$$E_p^2 = C_p - \underline{C}_{pd}\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd} \tag{34}$$

$$= C_p - \underline{C}_{pd}(\underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon)^{-1}\underline{C}^T_{pd} \tag{35}$$

The estimate of expected error is based on the expected variance, so any scaling of C_p will scale E_p^2 in the same way. To better understand the meaning of the $E_p^2(\underline{x},t)$, or "error map", consider an ensemble of oceans, constructed to obey the prior expectations. For each ocean in the ensemble, there is a data set, $\underline{d}$, and the estimator can produce $\hat{p}(\underline{x},t)$ using $\underline{d}$. The squared difference between

this estimate and the true field for this location in this element of the ensemble is $[p(\underline{x},t) - \hat{p}(\underline{x},t)]^2$. If this is calculated for every element of the ensemble and averaged, then $E_p^2(\underline{x},t)$ is obtained.

The error estimates obtained this way include both error due to the error in the data, $\underline{\varepsilon}$, and the so-called "resolution" error, due to inadequate sampling by the data. For example, if $\underline{\underline{C}}_\varepsilon$ is large, so that data error dominates $\underline{\underline{C}}_d$, the signal, then the error tends toward C_p, the expected variance of the model. The same thing happens if $\underline{\underline{C}}_{pd}$, the model-data covariance, goes to zero, for then the data taken contain no information about the model. In inverse theory jargon, resolution refers to the the ability of the total observation system, meaning both the data taking and the inverse, to reproduce any given true state. The observation system acts as a filter, blind to some structures of the true state while amplifying or distorting others. The ideal forward problem-inverse system would produce a $\hat{p}(\underline{x},t)$ equal to the true state, $p(\underline{x},t)$, for all $\underline{x},t$. The inverse system could then be characterized as a δ-function operator:

$$\hat{p}(\underline{x},t) = \int \delta(\underline{x}-\underline{x}',t-t')p(\underline{x}',t')\, d\underline{x}'\, dt' \qquad (36)$$

A practical inverse system will never obtain this ideal, but, for linear forward problem and inverse, the functional form of (36) can still be used, so that

$$\hat{p}(\underline{x},t) = \int K(\underline{x},\underline{x}',t,t')p(\underline{x}',t')\, d\underline{x}'\, dt' \qquad (37)$$

Note that the kernel, $K(\underline{x},\underline{x}',t,t')$ is not homogeneous, in general. If the kernel <u>is</u> homogeneous, so that

$$\hat{p}(\underline{x},t) = \int K(\underline{x}-\underline{x}',t-t')p(\underline{x}',t')\, d\underline{x}'\, dt' \qquad (38)$$

then the inverse system can be represented as a transfer function in spectral space by Fourier transforming in $\underline{x}$ and t to obtain $\underline{k}$ and s:

$$\hat{P}(\underline{k},s) = K(\underline{k},s)\cdot P(\underline{k},s) \qquad (39)$$

This particularly simple form allows the resolution of the system to be expressed using terms from signal processing, specifying the points in spectral space at which the transfer function, $K(\underline{k},s)$, reduces the energy in the true field by half. For example, a system of moorings might be characterized by having a "half power" point at 24 hours and at 50 km., meaning that motions with periodicity of 1 cycle per day are halved in power by the observing system, as are structures with a spatial scale of 50 km. Of course, if the system <u>was</u> characterizable in this way, then the filtering could be reversed by dividing by $K(\underline{k},s)$, provided that $K(\underline{k},s)$ is not zero at any point.

In practical problems, the simple form will not apply, since an efficient inverse procedure will compensate for simple attenuation, and the resolution is limited by non-homogeneous spatial averaging. In the probability estimation framework, the inverse does not lend itself to a form like (39), but it does return an estimate of the variance of the estimated result, $p(\underline{x},t)$.

In the case where no a priori information about the exact value of the unknown field is available ($\alpha \to \infty$), the covariance of the result, $C_{\hat{p}}$, is

$$C_{\hat{p}} = \underline{C}_{pd}(\underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon)^{-1}\underline{C}^T_{pd} \qquad (40)$$

If, instead of mapping to only one point in the volume, the estimator is constructed to map to many points, the entire discussion above carries over, but with p as a vector instead of a scalar. The covariance functions are still continuous in $\underline{x}$ and t, but equation (40) becomes

$$\underline{\underline{C}}_{\hat{p}} = \underline{\underline{C}}_{pd}(\underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon)^{-1}\underline{\underline{C}}^T_{pd} \qquad (40')$$

We have thus produced an estimate of the covariance of the estimated field for the set of points that were mapped. This will presumably be broader than the covariance, $\underline{\underline{C}}_p$, assumed originally, and the broadening could reasonably be used to define an approximate but simple figure of merit for the inverse. If the two covariances are both averaged,

so that they are homogeneous in space and time lags, then they both could be transformed, and the transfer function representation, (39) could be used to define "resolution lengths". This is quite an involved procedure, particularly when the result is of questionable value, so the expected error map and test cases will be used instead.

One other important feature of the probabilistic inverse framework is that it provides a means for checking the validity of the a priori assumptions made in constructing the inverse. Once $\hat{\underline{\lambda}}$ has been obtained (Eq. (21), or see appendix), it may be substituted into the prior probability density $\rho(\underline{\lambda})$, Eq. (1):

$$\rho(\hat{\underline{\lambda}}) = \gamma \exp[-1/2(\hat{\underline{\lambda}}-\tilde{\underline{\lambda}})^T \underline{\underline{C}}_a^{-1}(\hat{\underline{\lambda}}-\tilde{\underline{\lambda}})] \qquad (41)$$

We thus have a quantitative check of consistency between the model and the data. Eq. (41) is most effective in quantifying how well the data fit the forward problem and error covariance matrices specified for the inverse, particularly if there is no a priori value for $p(\underline{x},t)$. In a practical procedure, the estimated $\hat{p}(\underline{x},t)$ can be checked against the expected variances specified as part of the model-data relations. This quantifies the often informal examination of residuals that occurs in applications, but does not provide or justify a specific technique for revising the initial model in response to a misfit in the initial inverse calculation. For information on adaptive techniques, see Bretherton and McWilliams, (1980).

CHAPTER 5

INVERSE TECHNIQUES = PROBABILISTIC ESTIMATION

5.1 THE STOCHASTIC INVERSE (GAUSS-MARKOV THEOREM)

The fundamental assumption made in constructing the stochastic inverse is that of a statistical space in which both the data, $\underline{d}$, and the unknown field $p(\underline{x},t)$, are random variables. Note that the data are represented as a set of N discrete values, while the desired field is a continuous function of 4 variables. The estimation problem is that of estimating $p(\underline{x},t)$ for all $\underline{x},t$, but the method of solution we will use simplifies this global problem to that of estimating $p(\underline{x},t)$ point by point. Consider an ensemble average, $\langle\rangle$, defined on the space of random variables consisting of $\underline{d}$ and $p(\underline{x}_o,t_o)$ (the value of the unknown field at a given point.) The linear least square error estimator, $\hat{p}$, must then satisfy the following condition:

(1) Linearity:

$$\hat{p}(\underline{x}_o,t_o) = \sum a_i(\underline{x}_o,t_o)(d_i - \bar{d}_i) + \bar{p}(\underline{x}_o,t_o) \qquad (1)$$

where $\bar{\underline{d}}$, $\bar{p}(\underline{x}_o,t_o)$ are estimates of the means.

(2) Minimum squared error:

$$E^2 = \langle(p(\underline{x}_o,t_o) - \tilde{p}(\underline{x}_o,t_o))^2\rangle = \text{minimum.} \qquad (2)$$

The weights, $a_i(\underline{x}_o,t_o)$ are chosen to satisfy (2).

This procedure is elementary, and appears in many texts such as Aki and Richards, (1980), but a brief exposition will be given here for completeness. Write $p'(\underline{x}_o,t_o) = p(\underline{x}_o,t_o) - \bar{p}(\underline{x}_o,t_o)$, and $\underline{d}' = \underline{d} - \bar{\underline{d}}$, where $\underline{d}'$, $p'(\underline{x}_o,t_o)$ are perturbations around the estimated means. The condition (2) can be written as an extremum principle:

$$\frac{\partial E^2}{\partial a_i} = 0 \qquad i=1 \text{ to } N \qquad (3)$$

Substituting in the form of the estimator from (1):

$$\frac{\partial}{\partial a_i}\langle (p'(\underline{x}_o,t_o) - \sum a_i d'_i)^2 \rangle = 0 \qquad (4)$$

Taking the derivative,

$$2\langle (p'(\underline{x}_o,t_o) - \sum a_j d'_j) \cdot d'_i \rangle = 0 \qquad (5)$$

or

$$\sum a_j \langle d'_j d'_i \rangle = \langle p'(\underline{x}_o,t_o) d'_i \rangle \qquad (6)$$

This is a set of N equations in N unknowns, so

$$a_i(\underline{x}_o,t_o) = \sum \langle p'(\underline{x}_o,t_o) d'_j \rangle \, (\langle \underline{d}' \underline{d}'^T \rangle^{-1})_{ji} \qquad (7)$$

In vector form:

$$\underline{a}^T(\underline{x}_o,t_o) = \langle p'(\underline{x}_o,t_o) \underline{d}'^T \rangle (\langle \underline{d}' \underline{d}'^T \rangle^{-1}) \qquad (8)$$

so that

$$\hat{p}(\underline{x}_o,t_o) = \bar{p}(\underline{x}_o,t_o) + \langle p'(\underline{x}_o,t_o) \underline{d}'^T \rangle \langle \underline{d}' \underline{d}'^T \rangle^{-1} \, \underline{d}' \qquad (9)$$

If we wish to estimate $p(\underline{x},t)$ at more than one space-time location, then we only need to add a row of $\underline{a}^T(\underline{x}_i,t_i)$ for each new point $(\underline{x}_i,t_i)$ at which an estimate is required.

$$\hat{\underline{p}}' = \begin{matrix} p'(\underline{x}_1,t_1) \\ \vdots \\ p'(\underline{x}_M,t_M) \end{matrix} = \begin{matrix} \underline{a}^T(\underline{x}_1,t_1) \\ \vdots \\ \underline{a}^T(\underline{x}_M,t_M) \end{matrix} \langle \underline{d}'\underline{d}'^T \rangle^{-1} \underline{d}' \qquad (10)$$

$$= \underline{\underline{A}}\underline{d}' = \underline{\underline{A}}(\underline{d} - \overline{\underline{d}})$$

The complete estimation operator can then be written as:

$$\underline{\underline{A}} = \langle \underline{p}'\underline{d}'^T \rangle \, (\langle \underline{d}'\underline{d}'^T \rangle)^{-1} \qquad (11)$$

This result is commonly called the Gauss-Markov theorem.

Noise has not been explicitly mentioned in this derivation, but is implicitly included as part of the data. The expected errors for this estimator (11) are easy to calculate by substituting (8) into (2):

$$E^2 = \langle p'p' \rangle - \langle p'\underline{d}'^T \rangle \cdot (\langle \underline{d}'\underline{d}'^T \rangle^{-1}) \cdot \langle \underline{d}'p' \rangle \qquad (12)$$

For estimates at more than one point in space-time, the noise estimate can be converted using the vector notation introduced above. The single point variance generalizes to a total estimation error covariance matrix, $\underline{\underline{C}}_E$:

$$\underline{\underline{C}}_E = \langle \underline{p}'\underline{p}' \rangle - \langle \underline{p}'\underline{d}'^T \rangle \cdot (\langle \underline{d}'\underline{d}'^T \rangle^{-1}) \cdot \langle \underline{d}'\underline{p}' \rangle \qquad (13)$$

The error estimates, E^2 or $\underline{\underline{C}}_E$, contain variance due both to data error and incomplete resolution of the unknown field by the data-inverse system. For some purposes, it is interesting to separate out the error due only to data, although the error estimate made this way is not really statistically rigorous. If the data error is $\underline{\varepsilon}$, with covariance $\underline{\underline{C}}_\varepsilon$, then the covariance of the solution due only to the noise in the data is:

$$\underline{\underline{C}}_N = \underline{\underline{A}} \cdot \underline{\underline{C}}_\varepsilon \cdot \underline{\underline{A}}^T \tag{14}$$

For most applications, only the diagonal elements of $\underline{\underline{C}}_N$ or $\underline{\underline{C}}_E$ are usually of interest.

The least-square estimator can also be used to do spectral estimation. Since $\underline{p}' = \underline{\underline{A}}\underline{d}'$, estimating the unknown at several points, the covariance for the unknown can be estimated as

$$\langle \underline{p}'\underline{p}'^T \rangle = \langle \underline{\underline{A}}\underline{d}'(\underline{\underline{A}}\underline{d}')^T \rangle = \underline{\underline{A}} \langle \underline{d}'\underline{d}'^T \rangle \underline{\underline{A}}^T \tag{15}$$

Where $\langle \underline{d}'\underline{d}'^T \rangle$ is the observed data-data covariance computed throughout the experiment. The covariance matrix will usually consist of an irregular distribution of space and time lags, corresponding to all the separations between mapping points, and is not necessarily isotropic or

stationary. This quick and dirty estimate of the model covariance can be compared to the a priori assumptions, or can be averaged by lags into a stationary (covariance is a function only of lag) form, interpolated to a regular, 4-dimensional grid, and Fourier transformed to obtain a rough approximation to the 4-dimensional spectrum of the unknown field. Multi-dimensional, "beam-forming" algorithms could perhaps also be applied, to avoid the interpolation step, but it might be simpler just to map to a dense, regular grid.

In the special case where the inverse operator is time independent, it is easier to compute a frequency spectrum, point-by-point, for the unknown field. The obvious approach would be to convert the time series of data into a time series of estimates, and transform the new time series. If frequency bin averaging is to be used, then it is more efficient to take advantage of the linearity of the estimation operator and the Fourier transform by commuting the operations, and compute the spectrum of the data first. If the time series of data is $\underline{d}'(t)$, with the Fourier transform operator denoted as $F(\cdot)$, so that the Fourier transform of the data time series is $\underline{D}(s) = F(\underline{d}'(t))$, then the transform of the unknown field is $\underline{P}(s) = F(\underline{p}'(t))$, and the two are related by

$$\underline{P}(s) = F(\underline{p}'(t)) = F(\underline{\underline{A}}\underline{d}'(t)) = \underline{\underline{A}}F(\underline{d}'(t)) = \underline{\underline{A}}\underline{D}(s). \quad (16)$$

The power spectrum for the unknown field is then

$$\hat{\underline{P}}(s)^{*}\hat{\underline{P}}(s)^{T} = \underline{\underline{A}}\underline{D}(s)^{*}\underline{D}(s)^{T}\underline{\underline{A}}^{T} \qquad (17)$$

where * is the complex conjugate.

5.2 COMPARISON OF INVERSE METHODS

At first, it may seem odd that the Gauss-Markov theorem, which says nothing about probability distribution functions, gives the same result as the information-theoretical derivation of Chapter 4 for the case where there is no a priori information about the specific value of the unknown field. Liebelt (1967) and others have called Gauss-Markov estimation "distribution-independent" because it makes no explicit assumptions about the forms of the probability distributions for the unknowns. One only requires the first and second moment matrices to produce a minimum-variance estimator, although it is not explicitly guaranteed to be the optimal non-linear estimator.

In fact, the two problems can be seen to be equivalent if we recall (from Chapter 3) that the gaussian distribution is the smoothest (maximum entropy) distribution that satisfies the constraints of having a given mean and variance. When only mean and variance are given, as in the Gauss-Markov theorem, the state of information corresponds to that of a given Gaussian probability density. The Gauss-Markov estimator/"stochastic inverse" is the minimum variance, maximum likelihood estimator out of the set of all estimators, both linear and non-linear, which require a priori estimates of only the first and second moments. The two derivations may thus be reconciled, although the probabilistic derivation is somewhat more general.

Note that the changes to the output of the estimator affect only the rows of the model-data covariance matrix. The data-data covariance matrix is fixed by the data available in the experiment, and therefore does not change when a new output is desired. When any particular field or distribution of mapping points is desired, one needs only to compute the appropriate model-data covariance matrix and then multiply it by the inverse of the data-data covariance matrix, which has been computed once and saved.

The estimator is continuous, capable of producing estimates at all $\underline{x}$,t, and it is general within the linearity constraint on the form of the estimator, because it only uses statistical data. No mention has been made of error levels or of an explicit relationship between d_i and $p(\underline{x},t)$, linear or otherwise. The framework within which the result was derived assumes the availability of ensemble averages, but in a given application, limited assumptions and model physics may be used to construct the necessary covariance matrices. In these cases, the stochastic inverse can be shown to be equivalent to other traditional inverse forms.

To show how the various methods compare, the estimator in the form of equation (11) will be used. This inverse can estimate the unknown field at arbitrarily many points in the space, preserving the continuity of $p(\underline{x},t)$ in $\hat{p}(\underline{x},t)$.

Regardless of the degree of nonlinearity of the relation between d_i and p, for small perturbations it may be linearized around the "mean" state $\underline{d}' = p' = 0$ (the mean has been removed earlier, and quotation marks are used because approximations may have been made.) Let $\underline{\underline{G}}$ be an N x ∞ matrix representing the N linear functionals relating $\underline{d}$ to $\underline{p}$: then the ith "row" of $\underline{\underline{G}}$ is a linear functional operator,

$$\int g_i(\underline{x},t)(\cdot)d\underline{x}dt \ , \qquad (18)$$

since each datum, d'_i, is given by:

$$d'_i = \int g_i(\underline{x},t)p'(\underline{x},t)d\underline{x}dt + \varepsilon_i \qquad (19)$$

Equation (19) can be written more compactly by using an operator form, representing $p'(\underline{x},t)$ as a vector, with an infinite number of components:

$$\underline{d}' = \underline{\underline{G}}\ \underline{p}' + \underline{\varepsilon} \qquad (20)$$

$\underline{\varepsilon}$ is a random error vector containing errors due to both model errors and observation errors. The second modelling step needed is to specify a continuous function for the covariance of the unknown field, $\langle p'(\underline{x}_1,t_1)p'(\underline{x}_2,t_2)\rangle$, which can be represented as a matrix in the form we have adopted: $\underline{\underline{C}}_p = \langle \underline{p}'\underline{p}'^T \rangle$. The final modelling step is as important as the previous two, and consists of specifying the error covariance matrix: $\underline{\underline{C}}_\varepsilon = \langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle$.

Now, substitute the statement of the forward problem, (20), into the estimation framework, eq. (11):

$$\underline{\underline{A}} = \langle \underline{p}'(\underline{\underline{G}}\underline{p}' + \underline{\varepsilon})^T \rangle \langle (\underline{\underline{G}}\underline{p}' + \underline{\varepsilon})(\underline{\underline{G}}\underline{p}' + \underline{\varepsilon})^T \rangle^{-1} \quad (21)$$

$$= (\langle \underline{p}'\underline{p}'^T\underline{\underline{G}}^* \rangle + \langle \underline{p}'\underline{\varepsilon}^T \rangle)(\langle \underline{\underline{G}}\underline{p}'\underline{p}'^T\underline{\underline{G}}^* \rangle + \langle \underline{\underline{G}}\underline{p}'\underline{\varepsilon}^T \rangle + \langle \underline{\varepsilon}\underline{p}'^T\underline{\underline{G}}^* \rangle + \langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle)^{-1} \quad (22)$$

where $\underline{\underline{G}}^*$ denotes the adjoint of the linear functional operator (see Tarantola and Valette, 1982a). If the structure of the variable part of the field, $\underline{p}'$, is uncorrelated with the noise (an assumption violated if some of the model error is due to linearization or missing linear physics), then $\langle \underline{p}'\underline{\varepsilon}^T \rangle = 0 = \langle \underline{\varepsilon}\underline{p}^T \rangle$, and, since $\underline{\underline{G}}$ is an operator, not a random variable, it may be taken outside the ensemble averages, and (22) becomes

$$\underline{\underline{A}} = \langle \underline{p}'\underline{p}'^T \rangle \underline{\underline{G}}^* \, [\underline{\underline{G}} \langle \underline{p}'\underline{p}'^T \rangle \underline{\underline{G}}^* + \langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle]^{-1} \quad (23)$$

This is the form in which the inverse is applied to practical problems, and is identical to the form of "total inversion" (Tarantola and Valette 1982a). Suppose $\underline{p}'$ and $\underline{\underline{G}}$ are made finite dimensional by a truncated decomposition in M orthogonal functions, $h_j(\underline{x},t)$;

$$\underline{p}'_j = \int h_j(\underline{x},t) p'(\underline{x},t) \, d\underline{x} \, dt \quad j=1,..,M \quad (24)$$

Then, the operator, $\underline{\underline{G}}$ becomes a simple matrix as well:

$$(\underline{\underline{G}})_{ij} = \int h_j(\underline{x},t)g_i(\underline{x},t)d\underline{x}dt \quad i=1 \text{ to } N, \; j=1 \text{ to } M. \quad (25)$$

If $\langle \underline{p}'\underline{p}'^T \rangle \equiv \underline{\underline{W}}$ and $\langle \underline{\varepsilon}\underline{\varepsilon}^T \rangle \equiv \underline{\underline{E}}$, then (23) becomes

$$\underline{\underline{A}} = \underline{\underline{W}}\,\underline{\underline{G}}^T(\underline{\underline{G}}\,\underline{\underline{W}}\,\underline{\underline{G}}^T + \underline{\underline{E}})^{-1} \quad (26)$$

which is the standard geophysical inverse with weighting (Aki and Richards, 1980).

If these forms are all retained, but some manipulations are performed involving a strange-looking form of the identity matrix, $\underline{\underline{I}} \equiv \underline{\underline{E}}^{1/2}\cdot\underline{\underline{E}}^{-1/2}$, ($(Q)^{1/2}$ is defined so $(Q)^{1/2}(Q)^{1/2} = Q$), then (26) becomes:

$$\underline{\underline{A}} = \underline{\underline{W}}\,\underline{\underline{G}}^T(\underline{\underline{E}}^{1/2}\cdot\underline{\underline{E}}^{-1/2}[\underline{\underline{G}}\,\underline{\underline{W}}\,\underline{\underline{G}}^T]\underline{\underline{E}}^{-1/2}\cdot\underline{\underline{E}}^{1/2} + \underline{\underline{E}})^{-1} \quad (27)$$

or,

$$\underline{\underline{A}} = \underline{\underline{W}}\,\underline{\underline{G}}^T(\underline{\underline{E}}^{1/2}\cdot[\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\,\underline{\underline{W}}\,\underline{\underline{G}}^T\underline{\underline{E}}^{-1/2} + \underline{\underline{I}}]\cdot\underline{\underline{E}}^{1/2})^{-1} \quad (28)$$

Because the matrix to be inverted is non-singular, a true inverse exists, and the factors of $\underline{\underline{E}}^{1/2}$ can be pulled outside the inverse:

$$\underline{\underline{A}} = \underline{\underline{W}}\,\underline{\underline{G}}^T\underline{\underline{E}}^{-1/2}\cdot[(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\,\underline{\underline{W}}\,\underline{\underline{G}}^T\underline{\underline{E}}^{-1/2} + \underline{\underline{I}})^{-1}]\cdot\underline{\underline{E}}^{-1/2} \quad (29)$$

Equation (29) is identical to (26), but can be thought of as corresponding to a case where the forward problem has been weighted by the inverse square root of the error covariance matrix:

$$\underline{\underline{E}}^{-1/2}\underline{d}' = \underline{\underline{E}}^{-1/2}\underline{\underline{G}} \cdot \underline{p}' + \underline{\underline{E}}^{-1/2} \cdot \underline{\varepsilon} \qquad (30)$$

As mentioned above, if the matrix to be inverted is nonsingular then this transformation is a vector identity and cannot affect the estimator, but the form (29) is well known in the literature as the "tapered least-squares" estimator. The eigenvectors of $(\underline{\underline{E}}^{-1/2}\underline{\underline{GWG}}^{T}\underline{\underline{E}}^{-1/2} + \underline{\underline{I}})$ are the same as the eigenvectors of $(\underline{\underline{E}}^{-1/2}\underline{\underline{GWG}}^{T}\underline{\underline{E}}^{-1/2})$, and the eigenvalues differ only by the additive 1 due to the presence of the identity matrix. $\underline{\underline{GWG}}^{T}$ is the estimated data covariance matrix based on the linearized forward problem, $\underline{\underline{G}}$, and the estimated covariance matrix for the unknowns ($\underline{\underline{W}}$). This matrix is non-negative definite, but may have small (or zero) eigenvalues.

In most practical cases, the process of observation will introduce errors into the data, and adding the covariance of these errors, $\underline{\underline{E}}$, to the ideal, model-derived data-data covariance stabilizes the singularities to the extent required by the level of errors in the data. In some applications where the covariance matrix justification may

not be convenient the addition of a scalar multiple of the identity matrix is an ad hoc way to obtain a stable inverse that retains the same eigenvectors as the original (singular) matrix. Because this procedure "tapers" the singularity by adding "noise" to the diagonal to reduce the amplification of noise by the reciprocals of the small eigenvalues, it is called "tapered least-squares". This technique can only be justified in terms of least-squares methods if the matrices are weighted so as to have the form (29).

The Singular Value Decomposition (SVD) is a method for inverting non-square matrices (Lanczos, 1961). It is only applicable to cases where both the data and the unknown are discrete vectors. For concreteness, consider the following weighted linear forward problem,

$$\underline{\underline{E}}^{-1/2}\underline{d}' = (\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})(\underline{\underline{W}}^{-1/2}\cdot\underline{p}') + \underline{\underline{E}}^{-1/2}\cdot\underline{\varepsilon} \qquad (31)$$

Where the symbols are as defined above.

$\underline{\underline{E}}$ is (NxN) square, and $\underline{\underline{W}}$ is (MxM) square, and are the data measurement error and model covariances, respectively. $(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})$ is (NxM), and does not posess an inverse in the standard sense.

A practical inverse can be constructed, following Lanczos, by recognizing $(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})$ as a linear transformation between the model space and the data space, and solving the coupled eigenvalue problems for the bases of the two spaces:

$$(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})\cdot\underline{v}_i = \lambda_i \cdot \underline{u}_i \tag{32}$$

$$(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})^T\cdot\underline{u}_i = \lambda_i \cdot \underline{v}_i \tag{33}$$

Let $(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})$ be called $\underline{\underline{G}}'$, and the sets of eigenvectors be called $\underline{\underline{U}} = \{\underline{u}_i\}$, (NxN), and $\underline{\underline{V}} = \{\underline{v}_i\}$, (MxM), with $\underline{\underline{\Lambda}}$ the associated (NxM) matrix with the eigenvalues on its diagonal:

$$\underline{\underline{G}}'\underline{\underline{V}} = \underline{\underline{U}}\underline{\underline{\Lambda}} \tag{34}$$

$$\underline{\underline{G}}'^T\underline{\underline{U}} = \underline{\underline{V}}\underline{\underline{\Lambda}}^T \tag{35}$$

(Lanczos (1961) gives a full discussion of the analysis here.) These eigenvalues are usually obtained as the positive square roots (no loss of generality) of the singular values, λ_i^2, obtained from solving the simple eigenvalue problem for the square matrix;

$$(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})\cdot(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})^T = \underline{\underline{G}}'\underline{\underline{G}}'^T \ \text{(NxN)} \tag{36}$$

or

$$(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2})^T\cdot(\underline{\underline{E}}^{-1/2}\underline{\underline{G}}\ \underline{\underline{W}}^{1/2}) = \underline{\underline{G}}'^T\underline{\underline{G}}' \quad (MxM) \tag{37}$$

solving whichever problem is smaller. If the problem is underdetermined (M>N), then (36) is used, so that we solve

$$\underline{\underline{G}}'\underline{\underline{G}}'^T\cdot\underline{u}_i = \lambda_i^{\,2}\cdot\underline{u}_i \tag{38}$$

or

$$\underline{\underline{G}}'\underline{\underline{G}}'^T\cdot\underline{\underline{U}} = \underline{\underline{U}}\cdot\underline{\underline{\Lambda}}^T\underline{\underline{\Lambda}} \tag{39}$$

This problem has N eigenvectors, $\underline{\underline{U}}$, (a complete set), but some of the associated eigenvalues will be zero. The decomposition of $\underline{\underline{G}}'$ into eigenvectors and eigenvalues is

$$\underline{\underline{G}}' = \underline{\underline{U}}\cdot\underline{\underline{\Lambda}}\cdot\underline{\underline{V}}^T \tag{40}$$

This suggests that a "pseudo inverse" (Lanczos, (1961)) could be defined as

$$(\underline{\underline{G}}')_N^{-1} \equiv \underline{\underline{V}}\cdot(\underline{\underline{\Lambda}}^T)^{-1}\cdot\underline{\underline{U}}^T \tag{41}$$

($(\underline{\underline{\Lambda}}^T)^{-1}$ is an (MxN) matrix with $1/\lambda_i$ as the i^{th} diagonal element, i=1 to N)

since

$$(\underline{\underline{G}}')_N^{-1}\cdot\underline{\underline{G}}' = \underline{\underline{V}}\cdot(\underline{\underline{\Lambda}}^T)^{-1}\cdot\underline{\underline{U}}^T\cdot\underline{\underline{U}}\cdot\underline{\underline{\Lambda}}\cdot\underline{\underline{V}}^T \tag{42}$$

$$= \underline{\underline{V}}\cdot(\underline{\underline{\Lambda}}^T)^{-1}\cdot\underline{\underline{\Lambda}}\cdot\underline{\underline{V}}^T \tag{43}$$

$$= \sum_{i=1}^{N}\underline{v}_i(1/\lambda_i)\lambda_i\ \underline{v}_i^T \tag{44}$$

Unfortunately, the factor of $(1/\lambda_i)$ can be troublesome if $(\underline{\underline{G}}')_N^{-1}$ is to be applied to data. The inverse can be stabilized by removing the negligible eigenvalues, leaving R significantly non-zero eigenvalues, $\{\lambda_i: \ i=1,R\}$. Then $\underline{\underline{G}}'$ can be written in terms of these "activated" eigenvectors and eigenvalues only:

$$\underline{\underline{G}}'_r = \underline{\underline{U}}_r \cdot \underline{\underline{\Lambda}}_r \cdot \underline{\underline{V}}_r^T \simeq \underline{\underline{G}}' \tag{45}$$

$\underline{\underline{\Lambda}}_r$ is (RxR) with the non-zero eigenvalues on its diagonal. $\underline{\underline{U}}_r$ is (NxR) and $\underline{\underline{V}}_r$ is (MxR), and they contain the associated "activated" eigenvectors and are the basis sets for the range and domain, respectively, of the transformation $\underline{\underline{G}}'$. The pseudo inverse of $\underline{\underline{G}}'$ can then be written as

$$(\underline{\underline{G}}')_r^{-1} \equiv \underline{\underline{V}}_r \cdot \underline{\underline{\Lambda}}_r^{-1} \cdot \underline{\underline{U}}_r^T \tag{46}$$

$$= \underline{\underline{G}}'^T \cdot (\underline{\underline{G}}' \underline{\underline{G}}'^T)_r^{-1} \tag{47}$$

$$= (\underline{\underline{E}}^{-1/2} \underline{\underline{G}} \ \underline{\underline{W}}^{1/2})^T \cdot (\underline{\underline{E}}^{-1/2} \underline{\underline{G}} \underline{\underline{W}} \underline{\underline{G}}^T \underline{\underline{E}}^{-1/2})_r^{-1} \tag{48}$$

The pseudo inverse solution to the weighted forward problem is then:

$$\underline{\underline{W}}^{-1/2} \hat{\underline{p}}' = (\underline{\underline{E}}^{-1/2} \underline{\underline{G}} \ \underline{\underline{W}}^{1/2})^T \cdot (\underline{\underline{E}}^{-1/2} \underline{\underline{G}} \underline{\underline{W}} \underline{\underline{G}}^T \underline{\underline{E}}^{-1/2})_r^{-1} \cdot (E^{-1/2} \underline{d}') \tag{49}$$

or

$$\hat{\underline{p}}' = \underline{\underline{W}}^{1/2} \cdot (\underline{\underline{E}}^{-1/2} \underline{\underline{G}}\ \underline{\underline{W}}^{1/2})^T \cdot (\underline{\underline{E}}^{-1/2} \underline{\underline{G}} \underline{\underline{W}} \underline{\underline{G}}^T \underline{\underline{E}}^{-1/2})_r^{-1} \cdot (E^{-1/2} \underline{d}') \quad (50)$$

The singular value decomposition enables matrix inversion by ignoring the unstable eigenvalues. The matrix will have the same eigenvectors as the tapered least-squares inverse, provided the weighted forms in (29) and (31) are used. Recall that weighting the forward problem has no effect on the estimator when the noise covariances are included to make the matrix non-singular. Weighting is necessary when noise is not added, for otherwise, when the pseudo inverse is computed using only the R largest eigenvalues, the size of each row is important, and a change of units may change the estimator. The weighting using the error covariance matrix begs the question of why to weight at all--why not add the covariance in directly and save the trouble (and computer time) of computing the eigenvalues and eigenvectors explicitly?

The principle reason for using a truncated set of eigenvalues instead of tapering is that it yields an unbiased estimator for components of the model along the eigenvectors which are preserved in the inverse. This is discussed in Zlotnicki, Parsons, and Wunsch (1982), and will be briefly summarized here. Recall the SVD form of the forward and inverse problems:

$$\underline{\underline{E}}^{-1/2} \underline{d}' = (\underline{\underline{E}}^{-1/2} \underline{\underline{G}}\ \underline{\underline{W}}^{1/2})(\underline{\underline{W}}^{-1/2} \cdot \underline{p}') + \underline{\underline{E}}^{-1/2} \cdot \underline{\varepsilon} \quad (51)$$

or

$$\underline{d}' = \underline{\underline{G}}'\underline{p}' + \underline{\varepsilon} \tag{52}$$

and

$$\hat{\underline{p}}' = (\underline{\underline{G}}')_r^{-1}\underline{d}' \tag{53}$$

$$= \underline{\underline{V}}_r \cdot \underline{\underline{\Lambda}}_r^{-1} \cdot \underline{\underline{U}}_r^T \underline{d}' \tag{54}$$

$$= \sum_{i=1}^{R} \underline{v}_i (1/\lambda_i) \underline{u}_i^T \underline{d}' \tag{55}$$

If we then substitute in the forward problem (51) to put $\underline{d}'$ in terms of $\underline{p}'$, we obtain

$$\hat{\underline{p}}' = \sum_{i=1}^{R} \underline{v}_i (1/\lambda_i) \underline{u}_i^T \cdot (\underline{\underline{U}} \cdot \underline{\underline{\Lambda}} \cdot \underline{\underline{V}}^T) \underline{p}' \tag{56}$$

$$\hat{\underline{p}}' = \sum_{i=1}^{R} \underline{v}_i (1/\lambda_i) \lambda_i \underline{v}_i^T \cdot \underline{p}' \tag{57}$$

$$\hat{\underline{p}}' = \sum_{i=1}^{R} \underline{v}_i \cdot \underline{v}_i^T \cdot \underline{p}' \tag{58}$$

Thus, if $\underline{p}'$ is a linear combination of the R basis vectors, $\underline{V}_r$, then $\langle \hat{\underline{p}}' \rangle = \langle \underline{p}' \rangle$ and the estimator is unbiased.

Suppose that we examine the same form when errors have been added before inversion. Under the imposed weighting, the error covariance is the identity, so the tapered form of the estimator is:

$$\hat{\underline{p}}' = \sum_{i=1}^{N} \underline{v}_i (1/[\lambda_i + 1]) \lambda_i \underline{v}_i^T \cdot \underline{p}' \qquad (59)$$

Now $\langle \hat{\underline{p}}' \rangle \neq \langle \underline{p}' \rangle$ for all $\underline{p}'$. The bias of the probabilistic estimator results (in this simple form) from the noise "tapering" of the eigenvalues in the ideal data-data covariance matrix. The choice of which estimator to use seems to be at least partly dependent on the psychology of the investigator; for a more detailed (and philosophical) discussion see Zlotnicki (1983). The inversions to be presented in this thesis use the biased but minimum variance estimator.

If the model is instead left as a continuous field, $p'(\underline{x},t)$, and the covariance function is assumed to be a Dirac delta function, $\delta(\underline{x}_1-\underline{x}_2, t_1-t_2)$, then this corresponds to imposing no a priori constraints on the variation of $p(\underline{x},t)$, and the Backus-Gilbert (1967) result is reproduced (Tarantola and Valette (1982a)). The Backus-Gilbert formalism requires sophisticated mathematical analysis beyond the matrix algebra presented above, and will not be described here. Eisler, New, and Calderone (1983) have discussed this method of inversion in detail as applied specifically to ocean acoustic tomography.

A main feature of this method is that it produces an unbiased estimator. This is heuristically consistent with the earlier analysis, since the $\delta()$ covariance function for the unknown has infinite energy, the limiting case of uncertainty in the mean value. In practical terms, allowing the expected energy in the unknown field to go to ∞ produces infinite signal to noise ratios, negating the biasing by the eigenvalue tapering. The statistical implications for an estimator generated by assuming (incorrectly) an infinite signal to noise ratio are that the error must be controlled in another way, like the truncation in the SVD inverse.

Given certain assumptions, the stochastic inversion framework can thus be compared to more familiar forms. The simplifications in form allowed by truncation/discretization assumptions such as (24) restrict the generality of the stochastic inverse or the "total inverse" of (23), but each simplification can speed computations. Projecting $p(\underline{x},t)$ on a finite set of basis functions may sometimes be necessary from an economic standpoint, particularly when the kernels, $g_i(\underline{x},t)$ are small-scale and complicated, or when non-linearities force frequent recomputation of the inverse operator.

5.3 NON-LINEARITY AND ITERATION

The pure stochastic inverse as written in (9), (10), or (11) was derived on a basis of statistics, without regard to the order of the systems generating the $p(\underline{x},t)$ or d_i. If, as was done for tomography, the covariances are calculated from a model for $\langle p'(\underline{x}_1,t_1)p'(\underline{x}_2,t_2)\rangle$ and from a functional expression for the forward problem, the functional must be linear to obtain the simple form in (23). For many applications, the functionals, $g_i(\underline{x},t)$, linearized around a reference state $p_o(\underline{x},t)$, $\underline{d}_o$, may be valid only for small perturbations. For compactness, let us return to the "vector" notation for $p(\underline{x},t)$. If the estimated perturbation, $\hat{\underline{p}}'$ is large, then the functionals must be recomputed around the new state

$$\underline{p}_1 = \underline{p}_o + \hat{\underline{p}}' \tag{60}$$

The obvious solution would be to re-linearize around the estimated state $\underline{p}_1$:

$$\underline{d} = \underline{G}(\underline{p}) \cong \underline{G}(\underline{p}_1) + \frac{\partial \underline{G}}{\partial \underline{p}} \cdot (\underline{p} - \underline{p}_1) \tag{61}$$

$$= \underline{G}(\underline{p}_1) + \underline{\underline{A}}_1 \cdot (\underline{p} - \underline{p}_1) \tag{62}$$

The inversion would then have the form:

$$\underline{p}_2 = \underline{p}_1 + \underline{\underline{A}}_1^{-1} \cdot (\underline{d} - \underline{G}(\underline{p}_1)) \tag{63}$$

where $\underline{\underline{A}}_1^{-1}$ is the inverse of the "matrix" of partial derivatives which represents the linearized operator.

This type of iteration has several problems when one considers the form of the stochastic inversion. The fundamental assumptions are that we have some information about the first and second moments of $\underline{p}$ and $\underline{d}$. If the reference state is shifted as a result of iteration, then these assumptions are no longer applicable. Even if one argues that they were poor to begin with, the new estimator will require re-computation of the covariance function, as well as the matrices.

To avoid these problems, it is desirable to keep the original reference state and covariance functions, while re-linearizing the forward problem around a new state closer to the true state:

$$\underline{d} = \underline{G}(\underline{p}) = \underline{G}(\underline{p}_o) + \underline{\underline{A}}_o \cdot (\underline{p} - \underline{p}_o) \quad \text{(original)} \tag{64}$$

$$= \underline{G}(\underline{p}_k) + \underline{\underline{A}}_k \cdot (\underline{p} - \underline{p}_k) \quad \text{(kth iter.)} \tag{65}$$

$$= \underline{G}(\underline{p}_k) + \underline{\underline{A}}_k \cdot [(\underline{p} - \underline{p}_o) + (\underline{p}_o - \underline{p}_k)] \tag{66}$$

The forward problem can be re-written to reflect variations around the original reference state, as required by the statistics:

$$\underline{d} - \underline{G}(\underline{p}_k) + \underline{\underline{A}}_k(\underline{p}_k - \underline{p}_o) = \underline{\underline{A}}_k (\underline{p} - \underline{p}_o) \tag{67}$$

and the inversion:

$$\underline{p}_{k+1} = \underline{p}_o + \underline{\underline{A}}_k^{-1} [\underline{d} - \underline{G}(\underline{p}_k) + \underline{\underline{A}}_k(\underline{p}_k - \underline{p}_o)] \tag{68}$$

where $\underline{\underline{A}}_k^{-1}$ is the inverse operator for the matrix of partial derivatives at the k-th iteration.

Tarantola and Valette (1982a) discuss this iteration technique, calling it "fixed-point iteration", but they do not mention the statistical reason for retaining the original reference state, or the importance of the fixed point for consistency in the covariance functions and with any dynamic model. These latter are the primary reasons for using the fixed point iteration in the tomographic framework. Note that the success of iteration depends on the relative weakness of the non-linearities in the forward problem. If the linearization produces a result of opposite sign to the true value, then iteration cannot be expected to converge. For the acoustics, the linearization is generally robust: even if a strong ring or the wall of the Gulf Stream changes the sound speed by amounts far outside the boundaries of the linearization, the observed travel times will have the correct sign.

5.4 ITERATION SPECIFIC TO THE APPLICATION TO TOMOGRAPHY

To fix ideas, it is useful to consider fixed-point iteration as applied to the tomographic inverse problem, assuming only travel time data. Let $C_o(\underline{x},t)$ be the reference state, $C(\underline{x},t)$ be the true state, and $C'(\underline{x},t)$ the difference (perturbations relative to $C_o(\underline{x},t)$). The forward problem, linearized around the reference state, is:

$$d_i = \int_{\Gamma_{oi}} \frac{ds}{C_o(\underline{x}(s),t)} - \int_{\Gamma_{oi}} \frac{C'(\underline{x}(s),t)\, ds}{C_o(\underline{x}(s),t)^2} \tag{69}$$

Γ_{oi} is the path of the ith ray in the $C_o(\underline{x},t)$ state. The true ray path, propagating in the $C(\underline{x},t)$ sound speed field, will be called Γ_i, and will generally differ from the unperturbed ray path, Γ_{oi}.

The linearized functionals for the acoustic ray inverse problem can be written in operator form, for ease of comparison with the discussion above, replacing $\underline{p}$ by $\underline{C}$.

$$d_i = G_i(\underline{C}_o) + \frac{\partial G_i}{\partial \underline{C}}(\underline{C} - \underline{C}_o) \tag{70}$$

$$\underline{d} = \underline{G}(\underline{C}_o) + \underline{\underline{A}}_o(\underline{C} - \underline{C}_o) \tag{71}$$

so, inverting as before,

$$\underline{C}_1 = \underline{C}_o + \underline{\underline{A}}_o^{-1}\,[\underline{d} - \underline{G}(\underline{C}_o)] \tag{72}$$

The subscript "o" denotes that the ray paths used in the inverse were traced in the unperturbed $C_o(\underline{x},t)$ state.

($C_i(\underline{x},t)$ has been written as $\underline{C}_i$, but may be continuous.) Once $\underline{C}_1$ has been obtained, the fixed point re-linearization is carried out as before:

$$\underline{d} = \underline{G}(\underline{C}_1) + \underline{\underline{A}}_1(\underline{C} - \underline{C}_1) \tag{73}$$

$$d_i = \int_{\Gamma_{1i}} \frac{ds}{C_1(\underline{x}(s),t)} - \int_{\Gamma_{1i}} \frac{[C(\underline{x}(s),t) - C_1(\underline{x}(s),t)]}{C_1(\underline{x}(s),t)^2} ds \tag{74}$$

This must again be re-arranged to have the form of fixed-point iteration:

$$d_i - \int_{\Gamma_{1i}} \frac{ds}{C_1(\underline{x}(s),t)} + \int_{\Gamma_{1i}} \frac{[C_o(\underline{x}(s),t) - C_1(\underline{x}(s),t)]}{C_1(\underline{x}(s),t)^2} ds \tag{75}$$

$$= - \int_{\Gamma_{1i}} \frac{[C(\underline{x}(s),t) - C_o(\underline{x}(s),t)]}{C_1(\underline{x}(s),t)^2} ds \tag{76}$$

The left hand side can be simplified using the expansion as originally used in the linearization:

$$d_i - \int_{\Gamma_{1i}} \frac{ds}{C_o(\underline{x}(s),t)} \tag{77}$$

$$= - \int_{\Gamma_{1i}} \frac{[C(\underline{x}(s),t) - C_o(\underline{x}(s),t)]}{C_1(\underline{x}(s),t)^2} ds \tag{78}$$

Thus, for the acoustics, the fixed-point inverse problem is stated as:

$$di - \int_{\Gamma_{1i}} \frac{ds}{C_o(\underline{x}(s),t)} = - \int_{\Gamma_{1i}} \frac{[C(x(s),t) - C_o(x(s),t)]ds}{C_1(\underline{x}(s),t)^2} \tag{79}$$

Each sucessive iteration changes the data fed into the inverse only if the ray path changes;

$$\underline{C}_2 = \underline{C}_o + \underline{\underline{A}}_1^{-1} \begin{bmatrix} d_1 - \int_{\Gamma_{11}} \frac{ds}{C_o(\underline{x}(s),t)} \\ d_i - \int_{\Gamma_{1i}} \frac{ds}{C_o(\underline{x}(s),t)} \\ d_N - \int_{\Gamma_{1N}} \frac{ds}{C_o(\underline{x}(s),t)} \end{bmatrix} \tag{80}$$

Both the data fed into the inverse and the inverse operator, $\underline{\underline{A}}_1^{-1}$, are calculated for the modified ray paths. $\underline{\underline{A}}_1^{-1}$ inverts the perturbed operator,

$$- \int_{\Gamma_{1i}} \frac{[C(x(s),t) - C_o(x(s),t)]ds}{C_1(\underline{x}(s),t)^2} \tag{81}$$

although the statistical assumptions are referred to the original reference state.

CHAPTER 6

THE STOCHASTIC INVERSE APPLIED TO THE OCEANIC MESOSCALE

6.1 ADOPTING THE VERTICAL MODE BASIS

Given the results of quasi-geostrophic theory (Chapter 3), one wishes to construct the inverse framework to take advantage of any simplifications suggested analytically. By building a body of theory into the inverse, constraints such as non-divergence and geostrophic balance are applied during construction of the inverse operator, reducing indeterminacy and increasing resolution. For the mesoscale tomography experiment, the unknown fields were required to have the forms of solutions to the linearized quasi-geostrophic equations. This structure permits both the parameterization of vertical structure using modes instead of layers, and the calculation of velocities as part of the inverse procedure without any direct velocity measurement, although the indeterminacy of reference level velocity remains (and is explicit in the equations for the velocity associated with the 0^{th} mode). Because of the flexibility and generality of the stochastic inverse framework, I will first treat the application of quasigeostrophic theory to the stochastic inverse, from which the step to other inverse methods should be clear.

The major simplification obtained from the linear quasigeostrophic theory is the separation between the vertical and horizontal variation. The vertical structure equation for streamfunction, $\Psi(\underline{x},t)$, can be solved independently of the horizontal evolution equation, yielding solutions of the form:

$$\Psi(\underline{x},t) = \sum_{i=0}^{n} \phi_i(x,y,t) \cdot G_i(z) \qquad (1)$$

Chapter 3 describes the conversion from one set of vertical basis functions to another, so that, for example, displacement can be written as

$$\zeta(\underline{x},t) = \sum_{i=1}^{n} \phi_i(x,y,t) \cdot G_i^{\zeta}(z) \qquad (2)$$

$G_i^{\zeta}(z)$ and $G_i(z)$ are related analytically as shown in Chapter 3.

This procedure may be extended to tracer-like quantities, such as T, S, sound speed, or oxygen, which do not play direct roles in the evolution equations. The extension is based on distinguishing between perturbations induced by the vertical motion of water due to the mesoscale fluctuations and those which result from the presence and interleaving of different water masses.

Let primed variables denote perturbation quantities, while barred quantities denote practical estimates of true (ensemble) means. The true salinity field, $S(\underline{x},t)$, can then be expressed as:

$$S(\underline{x},t) = \overline{S}(\underline{x},t) + S'(\underline{x},t) \tag{3}$$

$$\langle S'(\underline{x},t)\rangle = 0. \tag{4}$$

The fundamental averaged quantities are T, θ_r, and S. $\Theta_r = \Theta(T,S,p,p_r)$ = potential temperature referenced to p_r, from which several important quantities may be derived.

$$\sigma_r = \sigma(\theta_r,S,p_r) \qquad \text{potential density anomaly (referenced to } p_r) \tag{5}$$

$$C = C(T,S,p) \qquad \text{sound speed} \tag{6}$$

$$N = N(T,S,p) \qquad \text{bouyancy frequency} \tag{7}$$

Potential density is the significant quantity for the dynamics, and its "barred" state represents the basic state around which the dynamical equation were linearized. The rest density profile is determined from the averaged temperature and salinity fields:

$$\overline{\sigma}_r = \sigma(\theta(\overline{T},\overline{S},p_s,p_r),\overline{S},p_r) \tag{8}$$

For any other tracers, simple averages may be computed.

Given these reference states, the perturbations due to the dynamical evolution of the field may be calculated:

$$S'(\underline{x},t) = [\sum_{i=1}^{n} \phi_i(x,y,t)\cdot G_i^{\zeta}(z)]\cdot S_z + R_S(\underline{x},t) \qquad (9)$$

$$= \sum_{i=1}^{n} \phi_i(x,y,t)\cdot G_i^{S}(z) + R_S(\underline{x},t) \qquad (10)$$

$G_i^S(z) \equiv G_i^{\zeta}(z)\cdot S_z$ are the modes of salinity variation due to the mesoscale fluctuations, and $R_S(\underline{x}',t)$ is the residual salinity anomaly not fundamentally connected with the dynamics. The analysis here assumes that the displacements (and perturbation) are small enough to justify the linearization used throughout. Similarly, the potential temperature variation may be written:

$$\Theta'(\underline{x},t) = [\sum_{i=1}^{n} \phi_i(x,y,t)\cdot G_i^{\theta}(z)] + R_\theta(\underline{x},t) \qquad (11)$$

$R_\theta(\underline{x},t)$ is the potential temperature perturbation independent of the dynamics, and

$$G_i^{\theta}(z) \equiv G_i^{\zeta}(z)\cdot(\Theta)_z \qquad (12)$$

are the potential temperature modes resulting from the displacement field. The vertical derivatives of potential temperature, (or in-situ temperature, density, or sound speed) must be calculated locally, assuming adiabatic motions.

Similar relations hold for sound speed and passive tracers, while σ has no residuals by definition. The residuals may be divided into vertical and horizontal modes of variation, using EOF analysis, for example, so that

$$R_S(\underline{x},t) = \sum_{i=1}^{k} \psi_i(x,y,t)\cdot A_i^S(z) + \sum_{i=1}^{k} \zeta(\underline{x},t)\cdot\psi_i(x,y,t)\cdot\frac{dA_i^S}{dz}(z) \tag{13}$$

$$R_\theta(\underline{x},t) = \sum_{i=1}^{k} \psi_i(x,y,t)\cdot A_i^\theta(z) + \sum_{i=1}^{k} \zeta(\underline{x},t)\cdot\psi_i(x,y,t)\cdot\frac{dA_i^\theta}{dz}(z) \tag{14}$$

and so forth. The "tracer modes", A(z), ψ(x,y,t), evolve with the physics of passive advection/diffusion, at least partially independent of the mesoscale evolution.

The $\{G_i(z)\}$ and $\{A_i(z)\}$ form a basis for the vertical structure of each quantity, and observations indicate that this basis is an efficient representation of the observed structure. For potential density anomaly computed from the 65 casts of the first CTD survey of the tomography experiment (D. Behringer) the first, second and third flat-bottom modes fit 85% of the variance below the upper 200 meters. Only a few vertical modes are usually needed to account for most of the variation over the >5 km depth range, a simplification over the number of layers required for a similarly realistic description.

6.2 CONSTRUCTING COVARIANCES USING QUASI-GEOSTROPHY

The covariance calculations are similarly simplified by this decomposition into vertical modes. Let the displacement anomaly, $\zeta'(\underline{x},t)$ ($\zeta \equiv 0$), be represented by the basis of dynamically-derived vertical functions described above;

$$\zeta'(\underline{x},t) = \sum_{i=1}^{n} \phi_i(x,y,t) \cdot G_i^{\zeta}(z) \tag{15}$$

Then the covariance, $\langle \zeta'(\underline{x}_1,t_1)\zeta'(\underline{x}_2,t_2) \rangle$ is given by

$$\langle \zeta'(\underline{x}_1,t_1)\zeta'(\underline{x}_2,t_2) \rangle$$

$$= \sum \sum \langle \phi_i(x_1,y_1,t_1)\phi_j(x_2,y_2,t_2) \rangle \cdot G_i^{\zeta}(z_1) \cdot G_j^{\zeta}(z_2) \tag{16}$$

since the vertical modes are not random variables and may be taken outside of the ensemble average. This expression (16) may be further simplified if the horizontal structure functions are assumed to be uncorrelated between modes:

$$\langle \phi_i(x_1,y_1,t_1)\phi_j(x_2,y_2,t_2) \rangle =$$

$$\delta_{ij} \cdot \langle \phi_i(x_1,y_1,t_1)\phi_i(x_2,y_2,t_2) \rangle \tag{17}$$

This assumption is consistent with linear dynamics, but is also useful in the general case, since robust correlations between modes are not yet known accurately enough to use as data.

Given assumption (17), (16) becomes

$$\langle \zeta'(\underline{x}_1,t_1)\zeta'(\underline{x}_2,t_2)\rangle$$

$$= \sum \langle \phi_i(x_1,y_1,t_1)\phi_i(x_2,y_2,t_2)\rangle \cdot G_i^{\zeta}(z_1)\cdot G_i^{\zeta}(z_2) \qquad (18)$$

It is often useful to normalize the vertical and horizontal structures so that the expected variance for the i^{th} mode is expressed by a scalar, γ_i. Under this simple transformation, introduced purely for flexibility later in the inverse procedure, (18) becomes:

$$\langle \zeta'(\underline{x}_1,t_1)\zeta'(\underline{x}_2,t_2)\rangle =$$

$$\sum \gamma_i \cdot H_i(x_1,y_1,t_1,x_2,y_2,t_2)\cdot G_i^{\zeta}(z_1)\cdot G_i^{\zeta}(z_2) \qquad (19)$$

The functions H_i are not necessarily stationary or isotropic, so that energy gradients within the region are allowed, and γ_i merely sets the overall energy level expected for mode i.

By Mercer's theorem, (Van Trees, 1968), a symmetric function, such as the covariance, may be expanded as a product, so

$$H_i(x_1,y_1,t_1,x_2,y_2,t_2) = \sum_{j=1}^{m} \alpha_{ij}\cdot F_{ij}(x_1,y_1,t_1)\cdot F_{ij}(x_2,y_2,t_2) \qquad (20)$$

If the covariance is derived directly from data, then one possible set of F_{ij}'s is the set of empirical orthogonal functions, where α_{ij} is the jth eigenvalue of H_i, and $F_{ij}(x,y,t)$ is the corresponding eigenvector. This expansion converts the stochastic inverse back to a weighted deterministic linear inverse, by supplying a finite set of basis functions. The expansion (20) directly expresses the trade-off between the deterministic and the stochastic inversions. If (n•m) is allowed to go to ∞, then the continuity of the solution is recovered, but if the expansion is well-defined and truncates for finite (n•m), then a deterministic inversion using the expansion

$$\zeta'(\underline{x},t) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} \cdot F_{ij}(x,y,t) \right) \cdot G_i^{\zeta}(z) \qquad (21)$$

is possible, and may be preferable for reasons of computational efficiency. If (n•m) is too large for economic summation of the series or if the basis functions F_{ij} are not easily definable in advance, then the stochastic inverse is more useful because the detailed physical structure of the horizontal variation does not need to be rigidly specified in the model. It is usually possible to specify vertical structures a priori for the mesoscale. This has been done, in order to streamline processing, for all the inversions to be discussed below.

6.3 ESTIMATION

The simplification and efficiency gained by the use of the modal basis becomes clear if the form of the stochastic inverse operator is calculated. Because the set of modes describes the vertical structure, only their amplitude need be calculated by the inverse. We no longer need to estimate $\zeta'(\underline{x},t)$, $\sigma'(\underline{x},t)$, $C'(\underline{x},t)$, and other quantities separately. Instead, calculate $\hat{\phi}_i(x,y,t)$ once, and then construct the desired fields by multiplying by the appropriate vertical mode functions.

$$\hat{\phi}_i(x,y,t) = \langle \phi_i(x,y,t)\underline{d}^T \rangle \cdot (\langle \underline{d}\underline{d}^T \rangle^{-1}) \cdot \underline{d} \qquad (22)$$

This formula (22) does not require the vertical modes to be orthogonal. Non-orthogonal basis sets complicate the calculation of expected energies because the projections on specific modes become ambiguous.

Once the set of $\hat{\phi}_i(x,y,t)$ has been obtained, the fundamental structures have been established, so all related quantities may be calculated by summing the appropriate expansion.

$$\hat{\zeta}'(\underline{x},t) = \sum^{n} \hat{\phi}_i(x,y,t)\cdot G_i^{\zeta}(z) \qquad (23)$$

$$\hat{\sigma}'(\underline{x},t) = \sum^{n} \hat{\phi}_i(x,y,t)\cdot G_i^{\sigma}(z) \qquad (24)$$

$$\hat{C}'(\underline{x},t) = \sum^{n} \hat{\phi}_i(x,y,t)\cdot G_i^{c}(z) + \sum \hat{\psi}_j(x,y,t)\cdot A_j^{c}(z) \qquad (25)$$

...and so on.

(If no measurements which constrain ψ are available, then it is set to 0). If $\underline{u}(\underline{x},t)$ is desired, then one must estimate

$$\widehat{\frac{\partial \underline{\phi}_i}{\partial y}}(x,y,t) = \langle \frac{\partial \phi_i}{\partial y}(x,y,t)\underline{d}^T\rangle \cdot (\langle \underline{d}\underline{d}^T\rangle^{-1})\cdot \underline{d} \qquad (26)$$

This only requires re-computation of the model-data covariance matrix:

$$\langle \frac{\partial \phi_i}{\partial y}(x,y,t)\underline{d}^T\rangle$$

The data-data covariance matrix (and its inverse) change only if a different data set is used.

6.4 USING ANALYTICAL RELATIONS BETWEEN THE COVARIANCES

The vertical modes corresponding to the various physically interesting quantities may be calculated from one another, and equations (22) and (26) suggest similar properties for the horizontal covariances. Let $\phi_i(x,y,t)$, the horizontal structure of the ith streamfunction mode, be the fundamental quantity for which the covariance is specified. This is consistent with the form of the quasigeostrophic theory, where a streamfunction is used as the basis from which the other fields of interest may be derived. Denote the covariance of the horizontal structure of the ith displacement mode with itself by

$$\langle \phi_i(x_1,y_1,t_1)\phi_i(x_2,y_2,t_2)\rangle = \gamma_i \cdot H_i(x_1,y_1,t_1,x_2,y_2,t_2) \quad (27)$$

The normalized covariance H_i, has not been assumed homogeneous or isotropic. The covariance of the horizontal structure of $u(\underline{x},t)$ with the horizontal structure of displacement is then given by

$$\langle \frac{\partial \phi_i}{\partial y_1}(x_1,y_1,t_1)\phi_i(x_2,y_2,t_2)\rangle = \gamma_i \cdot \frac{\partial H_i}{\partial y_1}(x_1,y_1,t_1,x_2,y_2,t_2) \quad (28)$$

This covariance, in conjunction with the linear functionals supplied by the forward problem, is used to calculate the model-data covariance matrix in (26) above. Note that once a function, H, has been chosen for the displacement/streamfunction horizontal structure, the covariances of related fields may be obtained by operating on H.

In general, suppose we are interested in

$$F_j(\underline{x},t) = L_j[\sum_{i=1}^{n} \phi_i(x,y,t) \cdot G_i^{\zeta}(z)] \tag{29}$$

$F_j(\underline{x},t)$ is a linear function of the basic (displacement) field, so that it commutes with summation and averaging. Then

$$\langle F_j(\underline{x}_1,t_1)F_k(\underline{x}_2,t_2)\rangle =$$

$$\langle L_j[\sum_{i=1}^{n} \phi_i(x_1,y_1,t_1) \cdot G_i^{\zeta}(z_1)] \cdot L_k[\sum_{m=1}^{n} \phi_m(x_2,y_2,t_2) \cdot G_m^{\zeta}(z_2)]\rangle \tag{30}$$

$\langle\ \rangle$ is a linear operation, and L_j and L_k are unaffected by the averaging. In addition, L_j operates only on the first $(\underline{x}_1,t_1)$ coordinate system, while L_k operates only on the $(\underline{x}_2,t_2)$ system, so the operators may be taken outside the ensemble average.

$$\langle F_j(\underline{x}_1,t_1)F_k(\underline{x}_2,t_2)\rangle =$$

$$L_j(L_k[\sum \langle \phi_i(\underline{x}_1,y_1,t_1)\phi_i(x_2,y_2,t_2)\rangle \cdot G_i^{\zeta}(z_1) \cdot G_i^{\zeta}(z_2)]) \tag{31}$$

$$\langle F_j(\underline{x}_1,t_1)F_k(\underline{x}_2,t_2)\rangle =$$

$$L_j(L_k[\sum \gamma_i \cdot H_i(x_1,y_1,t_1,x_2,y_2,t_2) \cdot G_i^{\zeta}(z_1) \cdot G_i^{\zeta}(z_2)]) \tag{32}$$

$$\langle F_j(\underline{x}_1,t_1)F_k(\underline{x}_2,t_2)\rangle =$$

$$\sum \gamma_i \cdot L_j(L_k[H_i(x_1,y_1,t_1,x_2,y_2,t_2) \cdot G_i^{\zeta}(z_1) \cdot G_i^{\zeta}(z_2)]) \tag{33}$$

This is a general result, and encompasses the case where the operators produce the data:

$$F_j(\underline{x}_1,t_1) = d_j,\ F_k(\underline{x}_2,t_2) = d_k \tag{34}$$

In this case, L_j, and L_k represent linear functionals as derived in the forward problem. For example, suppose that

$$F_j(\underline{x}_1,t_1) = T_l'(t_1) + T_m'(t_1) \tag{35}$$

$T_l'(t_1)$ is the travel time anomaly for the l^{th} ray (arbitrary indexing) at time t_1, and the m^{th} ray has the same path but travels in the opposite direction from ray l. Suppose as well that

$$F_k(\underline{x}_2,t_2) = T_q'(t_2) + T_r'(t_2) \tag{36}$$

which has similar structure. Then (33) is a representation of the j, kth element of the data-data covariance matrix Q:

$$(Q)_{jk} =$$

$$2\cdot\sum \int_{\Gamma_l}\int_{\Gamma_q} \{ \frac{\langle \phi_i[(x,y,t)(s_1)]\phi_i[(x,y,t)(s_2)]\rangle}{C(\underline{x}(s_1),t)^2\, C(\underline{x}(s_2),t)^2}\cdot G_i^c(z(s_1))\cdot G_i^c(z(s_2))\cdot ds_1 ds_2 \} \tag{37}$$

$$= 2\sum\gamma_i \int_{\Gamma_l}\int_{\Gamma_q} \frac{H_i[(x,y,t)(s_1),(x,y,t)(s_2)]}{C(\underline{x}(s_1),t)^2\, C(\underline{x}(s_2),t)^2} G_i^c(z(s_1))G_i^c(z(s_2))ds_1 ds_2 \tag{38}$$

Similarly, if $d_j(t) = T'(\underline{x}_j,t)$, the temperature anomaly at $(\underline{x}_j,t)$, and $d_k = u(\underline{x}_k,t)$, the eastward velocity anomaly at $(\underline{x}_k,t)$, then the corresponding element of the data-data covariance matrix is:

$$(\underline{\underline{Q}})_{jk} = \sum_i \gamma_i \cdot \frac{\partial [H_i(x_j,y_j,t,x_k,y_k,t)]}{\partial y_k} \cdot G_i^T(z_j) \cdot G_i^u(z_k) \quad (39)$$

These forms suggest that generality in mode weighting may be obtained easily by retaining the sum over vertical modes, so that

$$\underline{\underline{Q}} = \sum_i \gamma_i \cdot \underline{\underline{Q}}_i \quad (40)$$

$\underline{\underline{Q}}_i$ is the data-data covariance matrix calculated for just one mode. The assumption that there is no correlation between modes has been necessary for the simplification used in this chapter, but that assumption represents a state of restricted information relative to the state where the correlation coefficients between the modes are known. If reliable correlations between modes did exist, these could be incorporated into this framework by adding the cross-terms. In general, to allow maximum generality, it is worthwhile to keep separate matrices for distinct modes or different physics, because the expense of evaluating multiple integrals over ray paths, such as (38), can be major. The matrices may then be linearly combined with coefficients proportional to expected energies, to produce a data-data or model-data covariance matrix for a given estimation attempt without re-computing.

6.5 CONSTRAINTS

When one wishes to apply constraints as part of the estimation framework, each constraint should merely be treated as another datum, with weighting appropriate to its degree of certainty. If, for example, conservation of mass in a box, Γ, with boundary $\partial\Gamma$, is to be enforced, then one can write the constraint as a forward problem for the datum, d:

$$0 = d = \int_{\partial\Gamma} \rho \cdot \underline{u} \cdot \underline{n} \, ds \quad \pm \quad \varepsilon \tag{41}$$

$\underline{u}$ is the velocity vector, ρ is density, ε is the error limit, $\underline{n}$ is the unit normal to the surface of the box, and ds is an element of area of the boundary, $\partial\Gamma$, of the box, Γ.

The integral has the standard form of a datum, and must be linearized to be used in the estimation framework presented in this thesis. Recall from chapter 3 that the basic state, to which the inverses are referenced, has no velocities, and density $\rho = \overline{\rho}$. Equation (41) can then be linearized:

$$0 = d = \int_{\partial\Gamma} \overline{\rho} \cdot \underline{u} \cdot \underline{n} \, ds \quad \pm \quad \varepsilon \tag{42}$$

This constraint (42) asserts 0 mass creation or destruction within the box, within uncertainty ε, as a linear functional of the unknown velocity field. As another example, a no-flow condition could be enforced on the bottom, B, again with normal $\underline{n}$ and area ds:

$$0 = d = \int_B \underline{u}\cdot\underline{n}\, ds \quad \pm \quad \varepsilon \tag{43}$$

Note that no linearization is needed for this type of constraint. Given the constraint in the form (43), the model-data and data-data covariance matrices can be contructed by applying the functionals to the basic covariance functions. Suppose, for example, that the bottom boundary condition, (43), is to be used as a datum. The diagonal element of the data-data covariance matrix is the autocovariance of the datum, d:

$$\langle dd\rangle = \iint_{BB} \langle(\underline{u}(\underline{x}_1,t_1)\cdot\underline{n}(s_1))(\underline{u}(x_2,t_2)\cdot\underline{n}(s_2))ds_1ds_2\rangle \quad \pm \quad \langle\varepsilon\varepsilon\rangle$$

$$= \iint_{BB} \langle(\underline{u}(\underline{x}_1,t_1)^T\underline{n}(s_1))(\underline{u}(x_2,t_2)^T\underline{n}(s_2))\rangle ds_1ds_2 \quad \pm \quad \langle\varepsilon\varepsilon\rangle$$

$$= \iint_{BB} \underline{n}(s_1)^T\langle\underline{u}(\underline{x}_1,t_1)\underline{u}(x_2,t_2)^T\rangle\underline{n}(s_2))ds_1ds_2 \quad \pm \quad \langle\varepsilon\varepsilon\rangle \tag{44}$$

The 3x3 matrix of covariance functions can be evaluated by calculating the covariances as outlined above, using the quasi-geostrophic operators. The estimator would

attempt to satisfy (43) to within ε, using the probabilistic weighting. The residuals would then provide a quantitative consistency check on the constraint, just as they do for other data.

The integral constraints are perhaps the most obvious, but differential constraints can be used as well, again treating the constraint as a datum with some a priori error limit. One could apply the basic thermal wind balances from Chapter 3, but these are trivially satisfied because the covariance functions have been defined to be consistent with quasi-geostrophic structure, and thus satisfy the diagnostic relations identically. For example, consider the non-divergence condition,

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{45}$$

If this condition is imposed as a constraint, then one can write (45) as a datum for each point within the volume of interest:

$$0 = d = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \pm \varepsilon \tag{46}$$

The operations in (46) are linear, so the elements of the data-data covariance can be calculated using the procedure outlined above.

In order to fix ideas, consider one element of the data-data covariance, $\langle dD \rangle$. Suppose, for simplicity, that the other datum, D, is a measurement of streamfunction, by some miracle, so that

$$\langle dD \rangle = \frac{\partial \langle u(\underline{x}_1,t_1)\Psi(\underline{x}_2,t_2)\rangle}{\partial x_1} + \frac{\partial \langle v(\underline{x}_1,t_1)\Psi(\underline{x}_2,t_2)\rangle}{\partial y_1} \qquad (47)$$

$$= \sum \gamma_i \cdot \Big(-\frac{\partial}{\partial x_1} \langle \frac{\partial \phi_i}{\partial y_1}(x_1,y_1,t_1)\phi_i(x_2,y_2,t_2)\rangle G_i(z_1)\cdot G_i(z_2) + \frac{\partial}{\partial y_1} \langle \frac{\partial \phi_i}{\partial x_1}(x_1,y_1,t_1)\phi_i(x_2,y_2,t_2)\rangle G_i(z_1)\cdot G_i(z_2) \Big)$$

$$= 0 \qquad (48)$$

The other elements in this row/column of the data covariance vanish as well, as do the corresponding elements of the model-data covariance.

The diagnostic relations from the quasi-geostrophic approximation were imposed on the covariances because they are generally thought to hold nearly everywhere in the ocean, at least to lowest order. If (45) was to be explicit, with finite error, then an infinite number of "data" could be constructed, one for each point in the volume. The covariance functions for velocity, density, streamfunction, and so on would be independent, so the cross-covariances would be zero, but the relative energy levels and scales would still be adjusted to fit expectations, and would thus resemble the auto-covariances calculated using the quasi-geostrophic framework.

Applying the diagnostic constraints to the model means that they are specified without uncertainty, but they are applied to all points in the volume without over-complicating the estimator. The choice of which constraints to use in the model, and which to apply explicitly in the construction of the estimator must be based on a trade-off between these two considerations. If the uncertainty of the constraint is non-negligable for the purposes of the mapping, then it should be applied explicitly. For integral constraints, such as (43) above, this is convenient, but for a differential constraint, such as conservation of potential vorticity, one may choose either to build it into the model and add an appropriate amount of error to the covariances, to write explicit equations for a spaced set of points in the volume, or to use an integrated version of the constraint on blocks within the volume.

Perhaps the most important advantage of specifying the constraint as an additional datum is the consistency check that the residuals provide. When the model is built to conform to a set of a priori constraints, errors in the constraints will be distributed over all data, and may be difficult to dignose. When the constraint provides a datum, the misfit of that datum with the other data and constraints clearly and quantitatively evaluates the consistency and effectiveness of the constraint.

The inverse procedure combines consistency checks with constraints in a natural way. As one adds constraints to the model, one reduces the indeterminacy of the unknown field, thus reducing the number of degrees of freedom available to fit the data. The addition of constraints therefore both (1) reduces the effective noise power by restricting the "bandwidth" of the signal to which the estimator is sensitive, and (2) reduces the ability of the original model to fit the data, possibly driving up the residuals.

The addition of the constraints increases the resolution of the estimator, but if the fit to the data declines badly, so that the estimates of the data errors (residuals) are larger than allowed in advance, then the validity of the constraint (or the prior estimate of the noise level) must be re-examined. "Residual watching" has been an art, but it can be quantified under the formalism of the probabilistic inverse. In any case, the inverse techniques enable one to simultaneously check the validity of a conjecture and benefit from the increased information available if the conjecture was true.

The fact that the model proposed for the ocean variations incorporates the quasi-geostrophic diagnostic relations greatly increases the resolving power of the tomographic system. Consider the case where the analytical modes are used as a basis. The density data or the

acoustic data can then resolve all but the 0^{th} mode, so the indeterminacy of velocity reduces to indeterminacy of one mode amplitude. This should be relatively easy to estimate, using reciprocal travel times or a few current meters, particularly given the large scales expected for the barotropic mode.

This enhanced resolving power has been questioned on the basis that it is blind to contradictions in the basic assumptions. This is untrue, because the residuals from the estimators give a direct and quantitative measure of how well the model accounts for the data. The choice of whether to test or incorporate theoretical results must be made on a scientific basis. If, for example, the problem of acoustic propagation was not well-understood, then the data from the 1981 experiment could only be used to check consistency with the predictions of the theory, by comparing the ray arrivals measured at the receiver with those predicted by the theory, given a hydrographic survey of the area. By assuming that the acoustics are known, we can instead map the hydrography independently. In the same vein, it is to our advantage to incorporate any theoretical results which are not under test. Given that dynamic height maps have been used for many years, the inversion procedure presented above should be no more controversial, particularly since it does not assume a reference level.

CHAPTER 7

CLOCK ERRORS, MOORING MOTION, AND ANCHOR POSITION

7.1 INTRODUCTION

Ocean tomography as realized in the 1981 experiment depended on autonomous sources and receivers moored at mid-depth in a 300 x 300 km. array. Each instrument had an independent clock, and could sway in any direction as the mooring leaned in response to currents. Both mooring lean and clock drift can produce measured travel time changes which swamp the 40 msec. expected from mesoscale variation, so it is imperative that x,y,z offsets of a mooring from its assumed position and offsets of the instrument's clock from the true time can either be removed directly or compensated for. The 1981 tomography experiment was designed with systems to measure these errors so that they could be removed when the acoustic data was processed. The WHOI mooring tracking system was used with each acoustic mooring, recording position to within a few meters, and the frequency shifts of the quartz oscillators used as clocks were logged daily (see Chapter 1).

These correction systems were not invulnerable to failure, and mooring motion corrections were not available at least part of the time on all instruments, while two instruments were completely without mooring motion corrections. The corrections were subject to errors, as well.

For example, the clock drift measurements showed large, transient shifts (R. Spindel, personal communication) when the moorings were deployed, and the treatment of these transients is not necessarily obvious.

It is important to note that the mooring motion corrections only supply shifts with respect to an unknown reference position. Adding the uncertainty of LORAN navigation in the area in which the moorings were set to the possible horizontal motion of the mooring while it sinks during deployment means that the position estimates provided by the ship navigation at the time of setting were only good to about ± 2 km. in both the x and y directions. The depths of the instruments were also uncertain, due to possible errors in the lengths of the cables used to construct the moorings and in the bottom depths. Pressure recorders can lessen this uncertainty if they are available, but the instrument depths used in the 1981 experiment were uncertain to within 2 to 200 meters. Errors in position, if uncorrected, would prevent the use of numerical travel times as a reference state, because the differences between the observed travel times and the numerical travel times would be dominated by the position differences between those used for the ray trace and those which actually occurred.

Peter Worcester (1977) had to deal with the problems of uncorrected mooring motion when he transmitted between independently drifting and heaving ships, and portions of the discussion below follow his lead. Robert Spindel, at WHOI, is responsible for most of the procedures for tracking the moorings, calibrating the clocks, and applying the recorded corrections to the data.

Finally, even when mooring motion corrections are available, they are lacking in two respects: 1) The instrument moves vertically as well as horizontally, and these vertical shifts can distort the ray arrival pattern, even invalidating the ray identification if the mooring shifts by an extreme amount. 2) The simple corrections, $\Delta T = \Delta R/C$, described above for the horizontal position shifts are not completely accurate descriptions of the effects of changing instrument position on travel time, and the differences can easily be order 4 msec.

7.2 DAY-DIFFERENTIAL AND RAY-DIFFERENTIAL TRAVEL TIMES

A relatively simple solution to the problem of unknown mooring reference position is to abandon the numerical travel times, (and thus, the a priori reference state) and look instead at the travel time changes between day pairs during the experiment. If the ocean structure was known for one or more days of the experiment, as a result of a CTD survey, for example, then all differences could be taken relative to this day. Perturbations inferred from the travel time differences could then be added to the known state of the ocean on the reference day to produce an estimate of the total ocean structure. This type of travel time information will be called "day-differential", and was the type of data used to construct the maps shown in the preliminary discussion published shortly after the experiment (The Ocean Tomography Group, 1982). In a longer experiment, the travel times could be averaged over the length of the deployment, and the differences with respect to this mean travel time would produce perturbations relative to the mean ocean, provided the experiment was sufficiently long to adequately estimate the mean. Day-differentials have several good features: they are immune to all constant shifts in time base for each source-receiver pair, not just those arising from the unknown reference positions, and also minimize the effects of errors due to mis-identification of rays.

Day differentials do not solve the problem of uncorrected mooring motion, and in fact exacerbate it, because the errors on the two days add together. The need for a survey of the ocean to use as a reference state is problematic, since on the one hand, one of the goals of mesoscale tomography is to provide an alternative to expensive and slow ship surveys, and, on the other hand, the survey requires a finite amount of time, about three weeks in the case of the tomography experiment, so that the picture of the ocean obtained by the CTD is somewhat incompatible with the tomographic picture obtained in 200 seconds. It is possible to partially correct for this time problem by applying mesoscale dynamics to the CTD field, using Rossby wave propagation to estimate a snapshot of the ocean, although this approach requires many extra assumptions with unpredictable errors. In any case, day differentials throw away the absolute travel time information which is available from the tomography instrumentation, and, since the set of mooring motion corrections is incomplete, are only useful for about 5 days of the experiment.

The horizontal mooring motions can be partially removed from the inverse by referencing the travel times for each source receiver pair to one of the rays in the pattern. Thus, if there were 5 resolved arrivals for a given source-receiver pair, one of the arrivals would be chosen as a reference and subtracted from the other 4, yielding 4 "ray-differential"

travel times which contain only distortions of the arrival pattern (Worcester, 1977). Horizontal mooring motions just displace the arrival pattern, to lowest order, so ray-differentials provide a certain amount of immunity to uncorrected horizontal mooring motion. If only ray differentials were used, the day differentals would not be necessary, since the pattern distortions could be referenced to the numerical arrivals.

Unfortunately, the expected variations of the ray differentials are very small, order 10 msec RMS for the mesocale tomography experiment, so that the error levels become very critical. An error in ray identification will, when the ray differentials are calculated, swamp the ocean variation. Shifts in instrument depth strongly distort the pattern of ray arrivals, and can be important sources of ray differential variance. Horizontal mooring position shifts do distort the pattern weakly, and this source of error can be order 5 msec if the ray pattern contains rays with widely differing angles. The random measurement noise is doubled for ray differentials, as a result of the subtraction, so that if the random errors are 5 msec or larger the ray differentials will exert little influence on the maps.

When inversion calculations were attempted for the 1981 tomography data using ray differentials, the various noise sources were found to render pure ray differentials nearly useless. The Ocean Tomography Group paper used day differential travel time data for all paths, and used both day and ray differentials for two instruments, S4 and R5, for which mooring motion corrections were unavailable, and the results were still limited to the few days where nearly all instruments had complete corrections.

7.3 THE STRUCTURE OF MOORING MOTION AND TRAVEL TIME "NOISE"

Because of the limitations of "differential" travel times, a more sophisticated approach to combatting the noise from clock drift and mooring motion is required. The key concept is that these sources of variance in travel time are not white, but have identifiable physics and finite cross covariances. The ocean variations have characteristic patterns of effects on the acoustic travel times. These are calculated when constructing the data-data covariance matrix for the sound speed perturbations, $\underline{\underline{Q}}_c$. The eigenvectors of $\underline{\underline{Q}}_c$ are the expected modes of variation of the data vector, $\underline{d}$, due to the evolution of the mesoscale features, and the associated eigenvalues are the expected powers of these modes.

In the same way, the measurement noise has a particular covariance structure, the clock shifts another, the mooring motion another, and so on. The measurement noise in travel time determination is due to oceanic noise and the finite bandwidth of the transmissions. These errors are random and uncorrelated between paths, so the covariance function for this physics is a δ-function, and this parameterization cannot be improved on. Source or receiver clock shifts, on the other hand, have exactly the same effect on each ray in a given source receiver arrival pattern.

The source clock shift will be the same for all rays which leave that source, and the receiver clock shift will likewise be constant for all rays which hit the same receiver. Clock errors can thus be parameterized in terms of only one number (time dependent) for each mooring, and the effect on a given ray will depend only on which source-receiver pair it belongs to. For a ray k from source i to receiver j, the contribution to the measured travel time from clock errors ε_i and ε_j will be

$$\Delta T_k = \varepsilon_j - \varepsilon_i \qquad (1)$$

The clock shifts, ε_i, are independent between instruments, so the cross covariances of this state vector representation should be zero, and no further parameterization is necessary. Instead of a white noise variance added to all data, the clock noise can be expressed by $N_m = N_s + N_r$ parameters, reducing the effect of unknown clock error. The correlations of the clock shifts between rays allows this parameterization and the resultant gain in resolution over the white noise assumption.

The mooring motion noise is also correlated, as can be seen by examining its physical basis. In a perturbation framework, the travel time anomaly due to mooring motion or

anchor position offset for ray k can be written as a linear function of the x,y, and z shifts of both moorings, $\underline{\Delta x_i}$, $\underline{\Delta x_j}$:

$$\Delta T_k = \frac{\partial T_k}{\partial x_i}\cdot\Delta x_i + \frac{\partial T_k}{\partial y_i}\cdot\Delta y_i + \frac{\partial T_k}{\partial z_i}\cdot\Delta z_i + \frac{\partial T_k}{\partial x_j}\cdot\Delta x_j + \frac{\partial T_k}{\partial y_j}\cdot\Delta y_j + \frac{\partial T_k}{\partial z_j}\cdot\Delta z_j \qquad (2)$$

The partial derivatives in (2) can be estimated by ray tracing for different coordinates, but a simple perturbation approach allows analytical calculation of these quantities. First, decompose the horizontal terms into two parts: the dependence of travel time on horizontal range, R_k; and the dependence of horizontal range on the individual x or y coordinate:

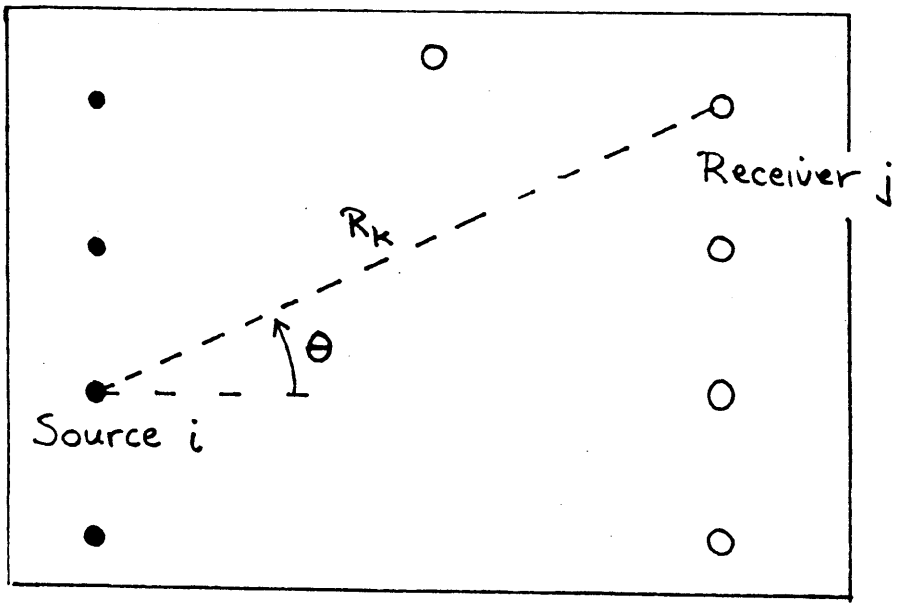
$$\frac{\partial T_k}{\partial x_i}\cdot\Delta x_i = \frac{\partial T_k}{\partial R_k}\cdot\frac{\partial R_k}{\partial x_i}\cdot\Delta x_i \qquad (3a)$$

$$\frac{\partial T_k}{\partial y_i}\cdot\Delta y_i = \frac{\partial T_k}{\partial R_k}\cdot\frac{\partial R_k}{\partial y_i}\cdot\Delta y_i \qquad (3b)$$

$$\frac{\partial T_k}{\partial x_j}\cdot\Delta x_j = \frac{\partial T_k}{\partial R_k}\cdot\frac{\partial R_k}{\partial x_j}\cdot\Delta x_j \qquad (3c)$$

$$\frac{\partial T_k}{\partial y_j}\cdot\Delta y_j = \frac{\partial T_k}{\partial R_k}\cdot\frac{\partial R_k}{\partial y_j}\cdot\Delta y_j \qquad (3d)$$

The $\partial R/\partial x,y$ terms can be calculated from the simple geometry, see figure (7.1), while the $\partial T/\partial R$ and $\partial T/\partial z$ terms can be approximated by assuming that the ray has a finite width, with the phase fronts normal to the ray path, so that the extra travel time resulting from the perturbed instrument



$$R_k = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

$$\frac{\partial R_k}{\partial x_i} = \frac{(x_i - x_j)}{R_k} = -\cos\theta$$

$$\frac{\partial R_k}{\partial y_i} = \frac{(y_i - y_j)}{R_k} = -\sin\theta$$

$$\frac{\partial R_k}{\partial x_j} = \cos\theta \qquad \frac{\partial R_k}{\partial y_j} = \sin\theta$$

FIGURE 7.1 SKETCH OF HORIZONTAL GEOMETRY AND DEPENDENCE OF SOURCE-RECEIVER RANGE ON SOURCE AND RECEIVER (X,Y) COORDINATES.

position will be the time it takes for the phase front to reach it (C. Spofford, personal communication). If the ray path is assumed to be locally straight, at an angle θ_1 to the horizontal, (positive for an upward-heading ray), and the local sound speed is C_1, (see Figures 7.2 and 7.3), then, at the receiver:

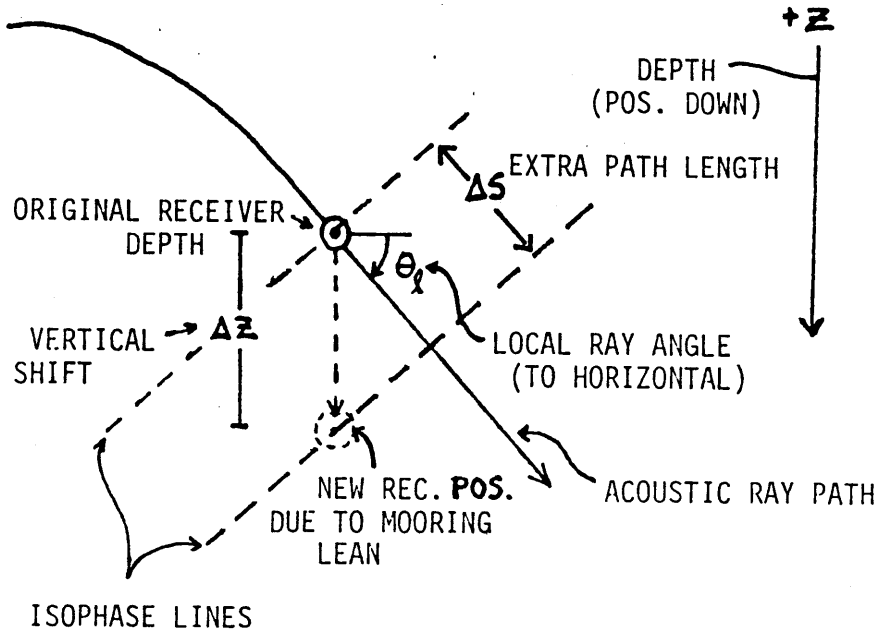
$$\frac{\partial T}{\partial z} = -\frac{\sin\theta_1}{C_1} \qquad (4)$$

$$\frac{\partial T}{\partial R} = \frac{\cos\theta_1}{C_1} = P \quad \text{(see Chapter 2.)} \qquad (5)$$

These are calculated at both source and receiver locations, and (4) has opposite sign at the source. The partial of travel time with respect to horizontal range is P, the ray parameter, so it is conserved along a given ray if the range dependence can be neglected. This means that the simple approximation that travel time is a function only of horizontal separation is correct, but that P, and not C_1, is the constant of proportionality. The travel time changes for vertical position offsets are different for source and receiver because $\sin\theta_1/C_1$ is not conserved.

Note that these expressions require the rays to be identified, so that the angles at both source and reciever are known. The converse is also true, however, as the mooring moves, the behavior of each peak in the arrival pattern will depend on the angle with which it arrives. In this way, mooring motion allows a single receiving instrument to be used

VERTICAL SHIFT IN RECEIVER POSITION

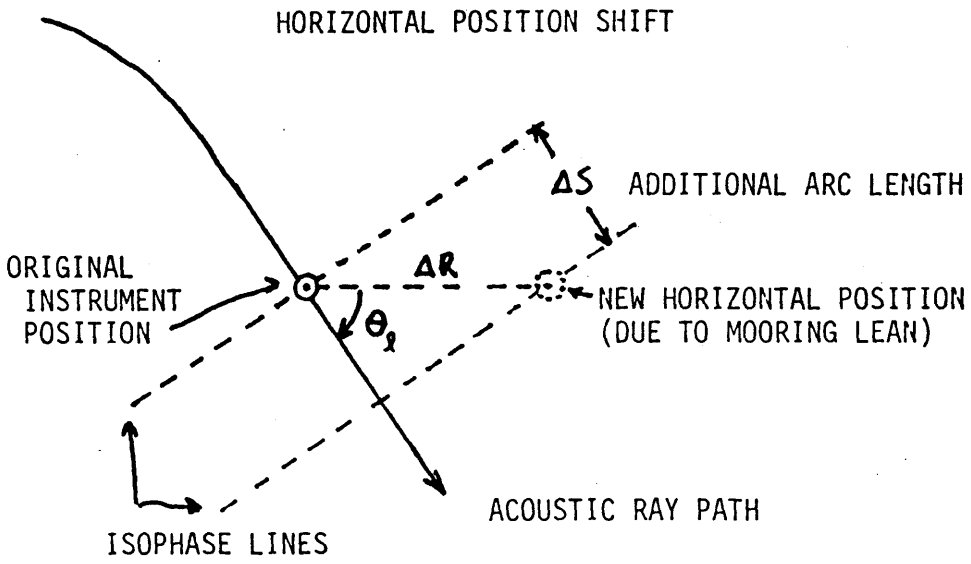


ΔS = ADDITIONAL ARCLENGTH ALONG RAY = $-\Delta z \sin\theta_\ell$

ΔT = ADDITIONAL TRAVEL TIME = $\frac{\Delta S}{c_\ell} = \frac{-\Delta z \sin\theta_\ell}{c_\ell}$

$$\frac{\Delta T}{\Delta z} = -\frac{\sin\theta_\ell}{c_\ell}$$

FIGURE 7.2 SKETCH OF HOW A CHANGE IN SOURCE OR RECEIVER DEPTH CHANGES RAY ARC LENGTH (AND THUS, TRAVEL TIME).



$$\Delta S = \Delta R \cos\theta_l$$

$$\Delta T = \frac{\Delta R \cos\theta_l}{C_l}$$

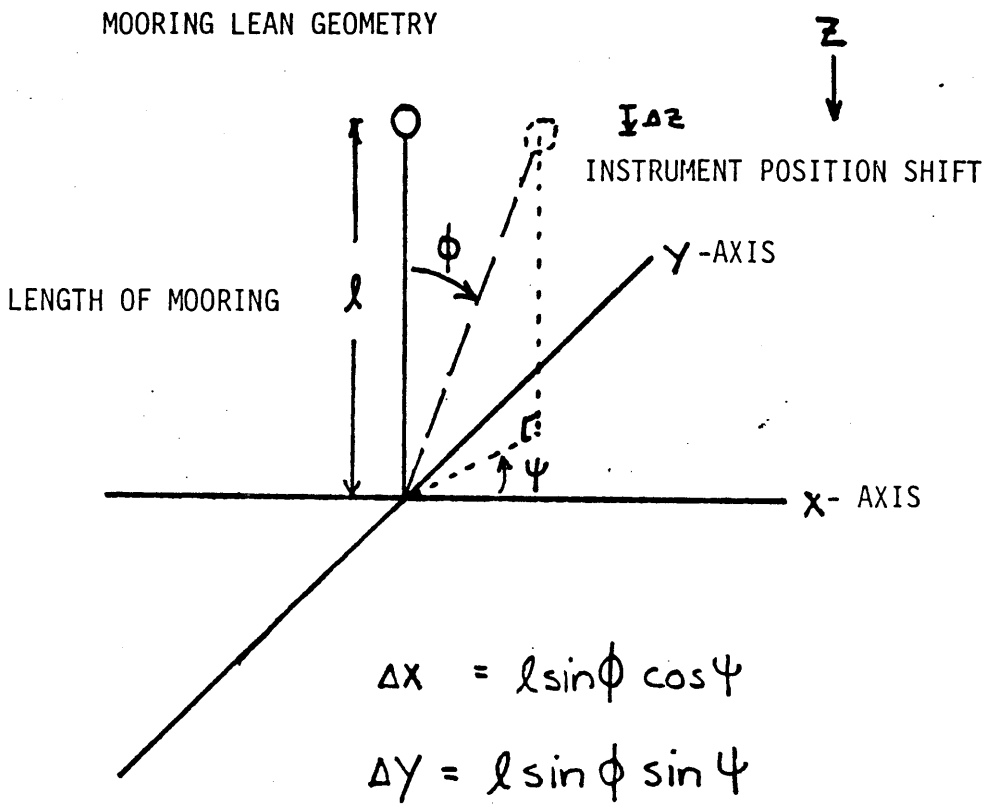
$$\frac{\Delta T}{\Delta R} = \frac{\cos\theta_l}{C_l}$$

FIGURE 7.3 SKETCH OF HOW A CHANGE IN SOURCE-RECEIVER RANGE CHANGES RAY ARC LENGTH (AND TRAVEL TIME.)

as a beam-former, adding angular information useful in ray identification. Vertical motion is most effective at distinguishing between angle, because of the $\sin(\theta_1)$ dependence, but horizontal motions can contribute, provided that the noise level is small enough.

Parameterization reduces the mooring motion errors to 3 unknowns per mooring. These are presumed to be independent, although, if the moorings were rigid, there would be only two unknowns per mooring, lean angle and lean direction, (Figure 7.4), so the number of parameters could be reduced. Unfortunately, the moorings were by no means rigid, but significant correlation between horizontal displacement and depth exists. For maximum generality and simplicity, I will leave the expression for mooring motion travel time in the form (2). Expected correlations between the parameters could be calculated using numerical mooring models, and then input into the inversions.

FIGURE 7.4 INSTRUMENT POSITION CHANGES AS A RESULT OF MOORING LEAN.



$$\Delta x = \ell \sin\phi \cos\psi$$

$$\Delta y = \ell \sin\phi \sin\psi$$

$$\Delta r = \ell \sin\phi$$

$$\Delta z \cong \frac{\ell \sin^2\phi}{2}$$

ϕ = LEAN ANGLE

ψ = LEAN DIRECTION

7.4 INCLUDING INSTRUMENT OFFSETS IN THE ESTIMATION PROCEDURE

Once the mooring motion and clock error dependences have been calculated for each ray, a data-data covariance matrix can be constructed. Let $\underline{\underline{M}}$ be the matrix of partial derivatives converting mooring motion and clock error to travel time for each ray, and $\underline{\Delta S}$ be the vector of x,y,z and time offsets for all the moorings, so that, if $\underline{\Delta T}$ is the vector of travel time anomalies,

$$\underline{\Delta T} = \underline{\underline{M}} \cdot \underline{\Delta S} \qquad (6)$$

By assumption, each element of $\underline{\Delta S}$ is independent of the others, so the covariance matrix for $\underline{\Delta S}$, $\underline{\underline{C}}_S = \langle \underline{\Delta S} \underline{\Delta S}^T \rangle$, will be diagonal, with each diagonal element reflecting the expected variance of that component on the day under consideration. These expected errors are estimated on the basis of the quality of the corrections available on that day, and change day to day. This covariance matrix for the mooring shifts can be used as the column weighting in a singular value inversion. If the stochastic inverse is used, then $\underline{\underline{C}}_S$ is needed to construct the data-data covariance matrix for the mooring shifts;

$$\underline{\underline{Q}}_m = \underline{\underline{M}} \cdot \underline{\underline{C}}_S \cdot \underline{\underline{M}}^T \qquad (7)$$

The total covariance matrix for the travel time has 3 components: variation due to the ocean sound speed changes, $\underline{\underline{Q}}_c$, the mooring shifts, $\underline{\underline{Q}}_m$, and the remaining random error which is uncorrelated between rays, $\underline{\underline{C}}_\varepsilon$ (diagonal);

$$\underline{\underline{Q}} = \underline{\underline{Q}}_c + \underline{\underline{Q}}_m + \underline{\underline{C}}_\varepsilon \tag{8}$$

Since the mooring shifts are now included in the inversion in parameterized form, they can be estimated by constructing the complete stochastic inverse operator;

$$\hat{\underline{\Delta S}} = \underline{\underline{C}}_S \cdot \underline{\underline{M}}^T \cdot (\underline{\underline{Q}}^{-1}) \cdot \underline{d}' \tag{9}$$

$$\hat{C}'(\underline{x},t) = \langle C'(\underline{x},t) \cdot \underline{d}^T \rangle \cdot (\underline{\underline{Q}}^{-1}) \cdot \underline{d}' \tag{10}$$

Some of the data used in the inverse may not be travel times, but, in any case, each row of $\underline{\underline{M}}$ will express the dependence of that datum on the mooring shifts. For example, a pressure measurement on one of the moorings would provide constraints on the motion of that mooring. In fact, the records obtained from the mooring tracking transponders could be used directly as data in the inverse, short-circuiting any need for seperate calculations in advance. In the limit, the mooring lean angle and direction would be the unknowns, and the motion of the water as observed by the acoustics and the current meters would have to be consistent with the mooring

motions. These perhaps complex interconnections could be exploited to increase resolution, since the indeterminacy would be reduced by each addition of physical relations, but at some point, the resolution gain would not be worth the extra effort required to add the extra physics to the inverse.

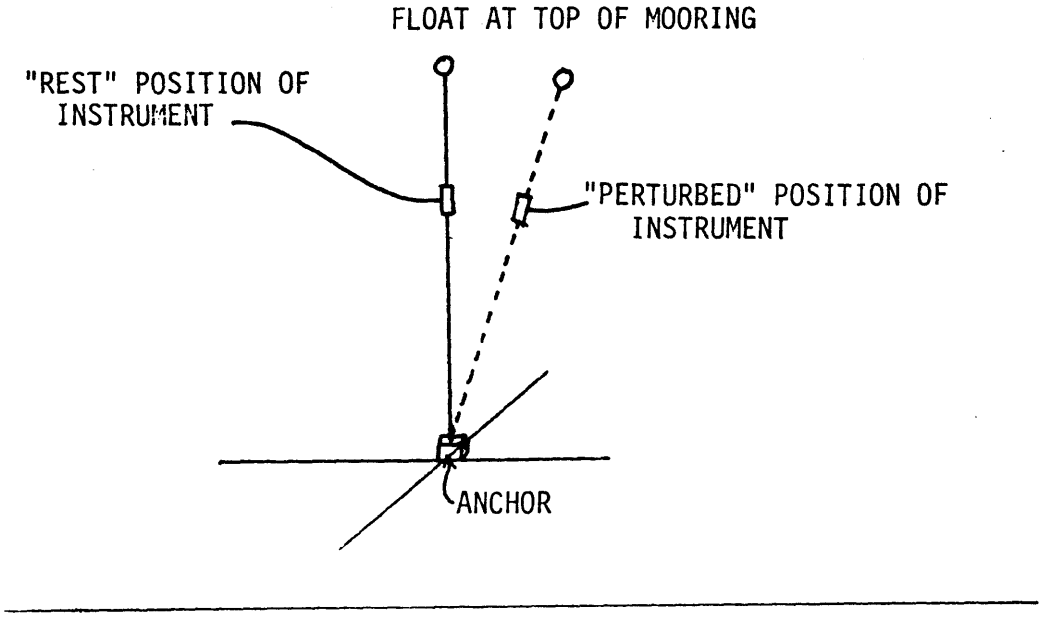
This point of diminishing returns determined the decision to leave mooring motion as $\Delta x, \Delta y$, and Δz instead of lean, because of the non-linearity of the dependence of $\Delta x, \Delta y$, and Δz on the angular displacements (see Figure 7.4). The cartesian coordinates also make the system more robust, in that it is not necessary to assume that the mooring leans as a rigid rod.

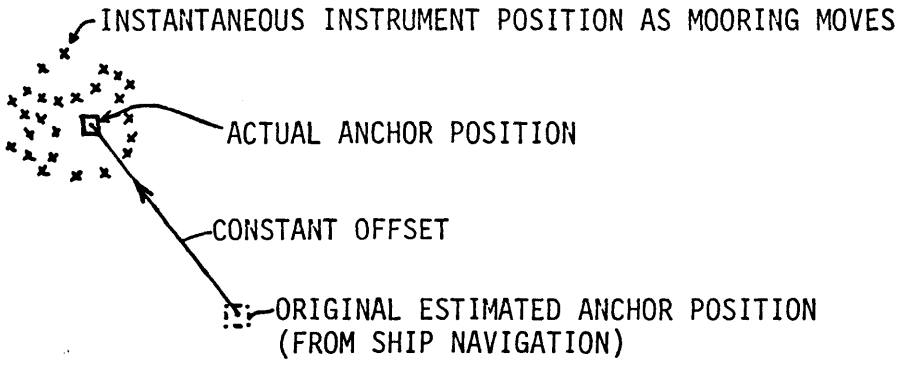
Retaining three degrees of freedom is necessary to treat non-moored applications of tomography. For example, it allows one to consider outfitting SOFAR floats with the more sophisticated transmitters, and using tomographic techniques instead of the simple position calculations now used. At the very least, one could expect to gain accuracy in the position fix, and perhaps some simple information about the location of the wall of the Gulf Stream. The "ultimate" inversions for the 1981 experiment may include the more efficient parameterization of mooring motion.

When the inversions produce position estimates as well as ocean maps, it becomes much easier to address the problem of mooring reference position. Given that the mooring anchor locations are uncertain to within about 2 kilometers, the travel time anomalies (with respect to rays traced numerically) due to anchor position dominate the observed anomalies, but must be constant throughout the experiment, (see Figure 7.5), so that the anchor positions may be estimated to within about 50 meters by averaging position estimates.

The inversions could then proceed with the variance due to the remaining uncertainty in anchor position added to the mooring motion variance, so that the inversions would be completely independent of any ocean survey. If on the other hand, the goal is not to compare the acoustics against the ocean survey, but to obtain the best estimate of the ocean given all data, then the CTD data can just be included as part of the data for the inverse, increasing the resolution of both the ocean and the mooring anchor positions. Because the anchor positions are constant, resolution can be improved by parameterizing the inverse both in terms of the constant anchor positions, with large variances, and the mooring motions, with generally smaller variances, but changing day to day. This separation will also be part of the "ultimate" inverse, but has not been carried out here.

FIGURE 7.5 SKETCH OF REST DEPTH OF INSTRUMENT AND OF MOORING MOTION COMPARED TO ANCHOR POSITION UNCERTAINTY.





If the covariances used in the inverse include time, then the mooring anchor positions should be parameterized as constants, with perfect coherence over all time separations, while the offsets due to mooring motion would have coherences which decay on a time scale of a few hours to days.

Finally, it is now easy to see how to treat the case where absolute travel times are not available. In this case, there is an additional (constant) unknown for each source-receiver pair, which would be estimated using data throughout the experiment. The case where the sources and receivers are suspended from ships is also tractable now, even without high-accuracy navigation, since the tomographic system can have useful resolution in the absence of accurate position information. The engineering trade-offs for large-scale tomography can also be more flexible, since the need for periodic clock checks, mooring tracking, or ocean surveys may be eliminated by sufficient travel time precision and enough source-receiver pairs.

7.5 DISCUSSION

If the mooring motion and anchor position offsets are lumped together, then there are 4 undetermined parameters per mooring. For N_S sources and N_R receivers (= N_m instruments), instrument offsets would then constitute $N_m \cdot 4$ unknown parameters. Of these, it is easy to see that a uniform clock shift among all instruments does not affect the data. Likewise, a uniform translation (in x or y), or a solid body rotation of the array cannot affect travel time. There are thus $(N_m-1)\cdot 4$ parameters which affect the data, but in a given case, degeneracy may reduce the number further. If the rays of a single source-receiver pair do not give range information (a worst-case assumption), and k vertical modes can describe the ocean, then there are $k \cdot N_S \cdot N_R$ independent pieces of information which may be gathered for the inverse problem for the ocean. This means that we should expect that about $(N_m-1)\cdot 4 + k \cdot N_S \cdot N_R$ independent rays could be used. For a 4 source, 5 receiver array in a region where the ocean appears to have energy in only 3 modes, we expect that about 92 rays would be independent in a noise-free experiment.

When white measurement noise is present, all rays add at least a small amount of independent information (about the noise), but the resolution of the ocean will degrade, even when more rays are added. If, for example, the noise variance

is greater than the travel time variance due to the 3rd mode, then there are really only 2 resolved modes, so about 72 rays would be expected to be independent. In a practical case, more than this minimum number of rays would be required, because some rays would be only weakly independent, but this calculation gives a good rule of thumb. If one calculates the expected variance due to horizontal feature position for a single source-receiver pair (range information), then one can estimate the error level at which the range information becomes accessible. This would allow, for example, a back-of-the-envelope evaluation of the possibility of 2-dimensional vertical (x-z) slice reconstruction from a single source-receiver pair.

For the mesoscale geometry and present equipment, about 8 to 10 arrivals are distinguishable at the receivers. If we conservatively estimate 5 independent rays per source-receiver pair, then an array of N_S sources and N_R receivers would produce $5 \cdot N_S \cdot N_R$ data, as opposed to $(N_R + N_S - 1) \cdot 4$ mooring offset parameters, in the worst case. It is clear that, as the number of instruments grows, lack of position information becomes very easy to compensate for, even with an inefficient parameterization. On the other hand, as the range of the transmissions grows, the number of rays per source-receiver pair grows as well. Once again, undetermined offsets become less of a problem, provided the precision of the system is sufficient to distinguish the available arrivals.

CHAPTER 8

DATA TREATMENT IN THE 1981 EXPERIMENT

8.1 DATA RETURN

In this chapter, I will describe the complete data processing procedures for the 1981 ocean acoustic tomography experiment, from the instrument processing to inversion procedures. For additional details about the experiment, see Chapters 1, 7, or the description in the paper by the Ocean Tomography Group (1982).

The 1981 ocean acoustic tomography experiment used 4 acoustic sources and 5 receivers, arranged in an array as shown in Figure (1.4). The array was centered on about 26 N, 70 W, nearly coinciding with the region where the MODE experiment was carried out (MODE Group, 1976). During the course of the experiment, 3 CTD and bottle hydrographic surveys were made by NOAA ships in the region, and several AXBT flights were made by the Navy, in order to have traditional measurements in the region for comparison with the tomography results.

The moorings were deployed in February 1981, with an expected duration of 4 months, and the three hydrographic surveys were spaced through this interval. Unfortunately, battery problems shut down most of the Woods Hole receivers by about day 120, so the full array was operating for only about 70 days, although the SIO receivers recorded data out to day 172 (see table (8.1)).

TABLE 8.1 MOORING DATA RETURN

FROM A MEMO FROM R.SPINDEL 9/28/81

A: CLOCKS

MOORING	START DAY	TOTAL DAYS
S1	21	219
S2	61	178
S3	36	203
S4	30	208
R1	47	66
R2	46	150
R3	43	134
R4	43	155
R5	48	135

B: MULTIPATH DATA

MOORING	START DAY	TOTAL DAYS
R1	49	120
R2	46	69
R3	49	87
R4	46	63
R5	49	120

NOTE: R2,3,4 FAILED EARLY DUE TO BATTERY PROBLEMS

TABLE 8.1 CONTINUED

C: MOORING MOTION

MOORING	START DAY	TOTAL DAYS
S1	32	160
S2	34	185
S3	35	185
S4	36	NONE
R1	47	175
R2	46	185
R3	45	185
R4	38	185
R5	48	FRAGMENTARY

(NOTE: SOME OF THE RECORDS ABOVE CONTAIN GAPS)

The sources and receivers were equipped with the Woods Hole mooring tracking systems, as mentioned above, supplying data on mooring motion for many of the instruments during much of the experiment. This great time variability in the quality of the data requires that the inverse framework and the data reduction programs must be flexible enough to handle data with gaps and inhomogeneities.

The acoustics operated one day in three, transmitting each hour for 24 hours and then shutting down for 48. The WHOI receivers recorded each transmission, but the SIO receivers listened only every other hour. To avoid interference and reverberation, the sources transmitted at 15-minute intervals, with source 1 transmitting on the hour, source 2 on the quarter hour, and so on. The sources transmitted on a carrier of 224 Hz with a bandwidth of 20 Hz, sending 24 repetitions of a 127-digit phase-coded shift register sequence. The complete sequence lasts for 7.9375 seconds. The receivers were set to turn on at a specific amount of time after each source began to transmit, and recorded for long enough to receive 22 repetitions of the code. The receiver turn-on delay was calculated on the basis of the planned mooring locations so that the receivers would ideally record the middle 22 transmissions of the code. As a result, 8 seconds of variation in either direction, due to uncertain mooring positions, was allowed.

The receivers recorded 2 samples per digit (= 254), and the 22 repetitions of the code "wrapped around", so that sample 255 was added to the sample 1 already in bin 1, and the 22 transmissions of the code were summed. This worked to increase signal to noise ratio within the stringent power limitations. The wrap around means that the first bin of the receiver corresponds to a travel time equal to the receiver delay, plus or minus 7.9375 seconds. This indeterminacy does not cause any ambiguity in absolute travel time because 8 seconds of travel time means about 12 km. of range, and the mooring locations were known to within ±2 km.

The averaged received code was correlated with a stored record of the code as transmitted, a process called phase-matched filtering (Birdsall, 1976), which produced a set of correlation peaks (Figure 8.1). The largest peaks each correspond to the arrival of a distinct acoustic ray, or, in some cases, a set of rays whose travel times are seperated by less than the resolution width of the system. Some of the receivers stored these 254 complex numbers directly, while the others stored only the 11 highest peaks. The length of each digit is 62.5 msec, so the system can resolve peaks separated by more than 62.5 msec.

FIGURE 8.1 A ARRIVALS FOR SOURCE 2 - RECEIVER 2: FROM THE 1ST TRANSMISSION ON DAY 100.

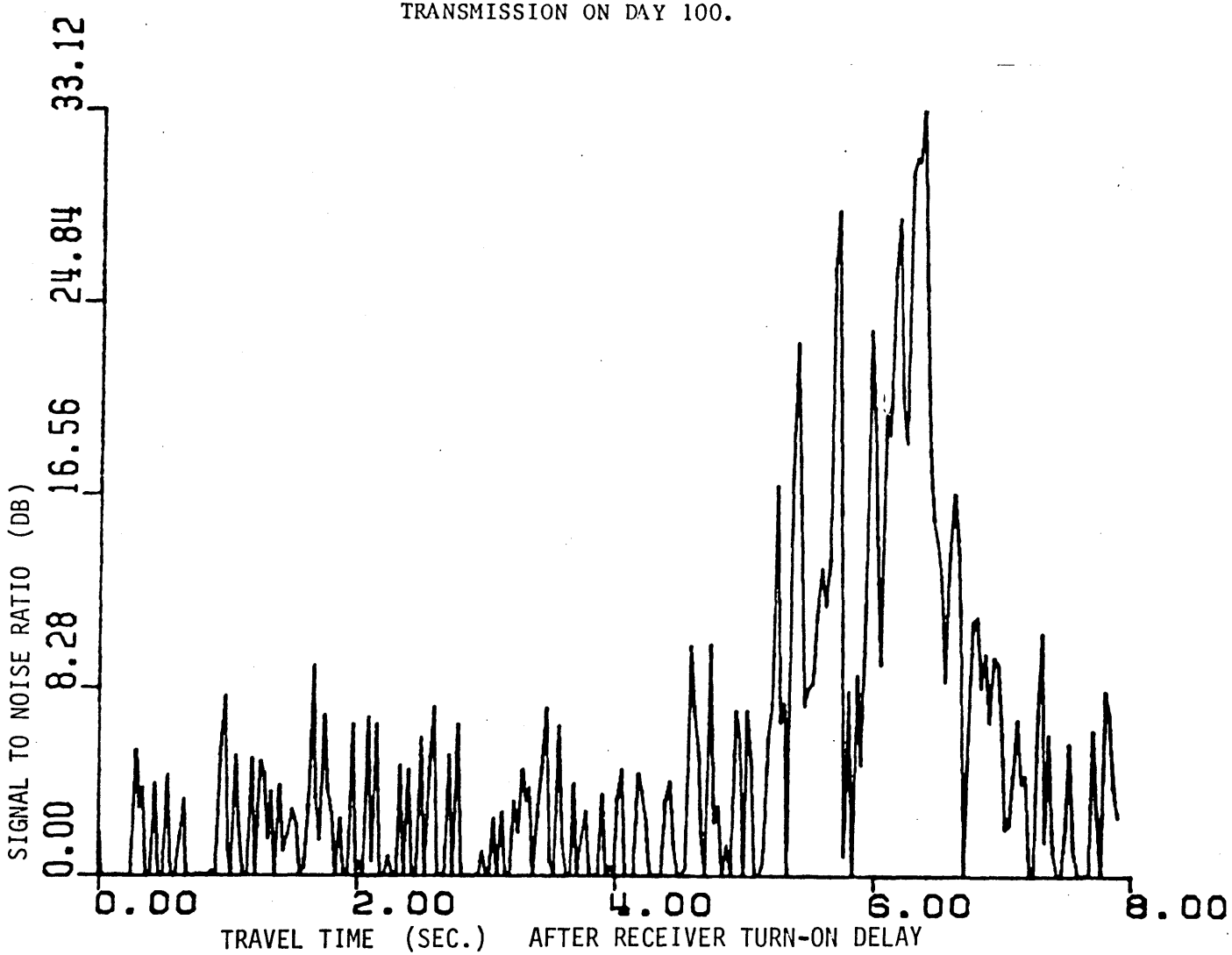


FIGURE 8.1 B ARRIVALS FOR SOURCE 2 - RECEIVER 2: FROM THE 2ND TRANSMISSION ON DAY 100.

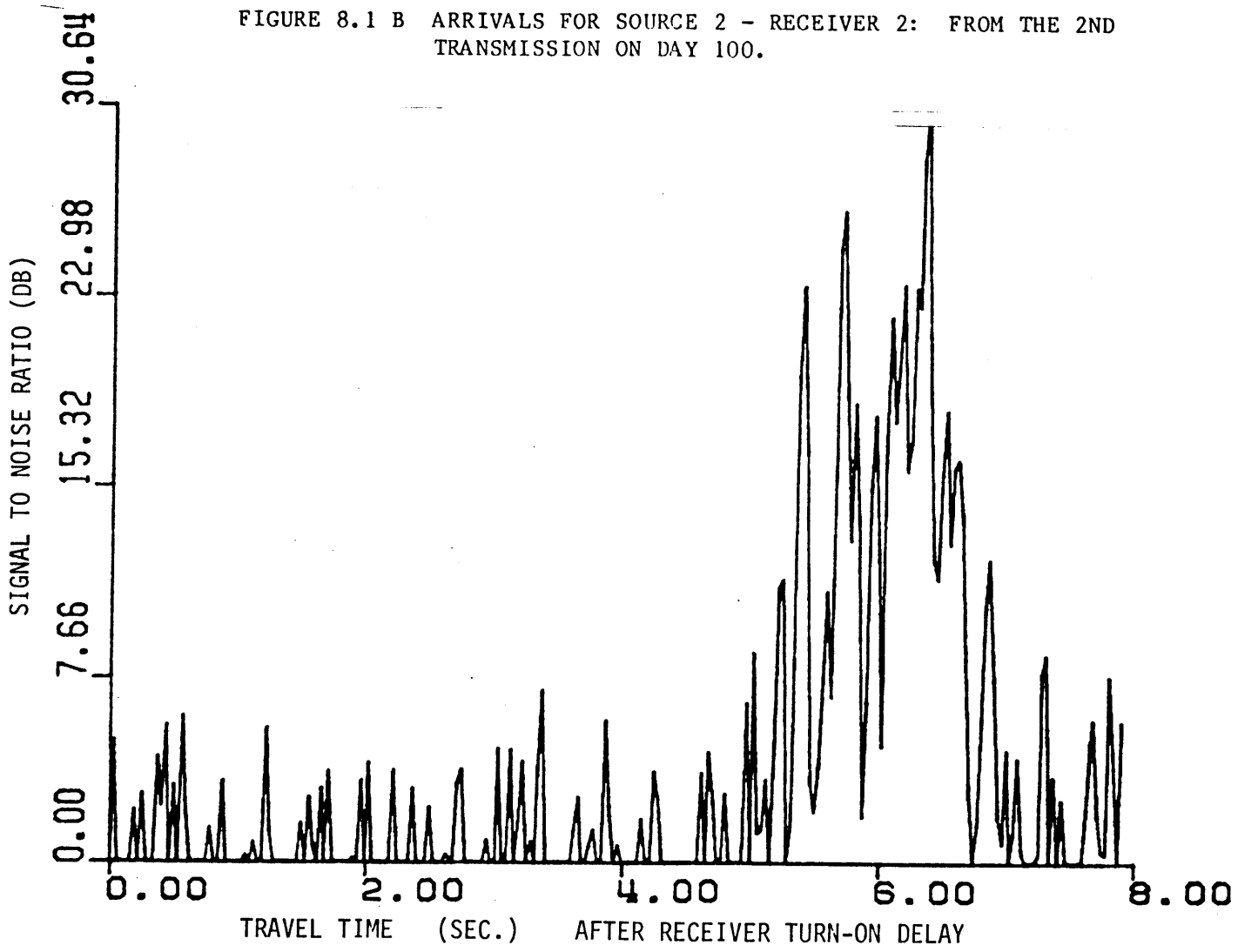


FIGURE 8.1 C ARRIVALS FOR SOURCE 2 - RECEIVER 2: FROM THE 3RD TRANSMISSION ON DAY 100.

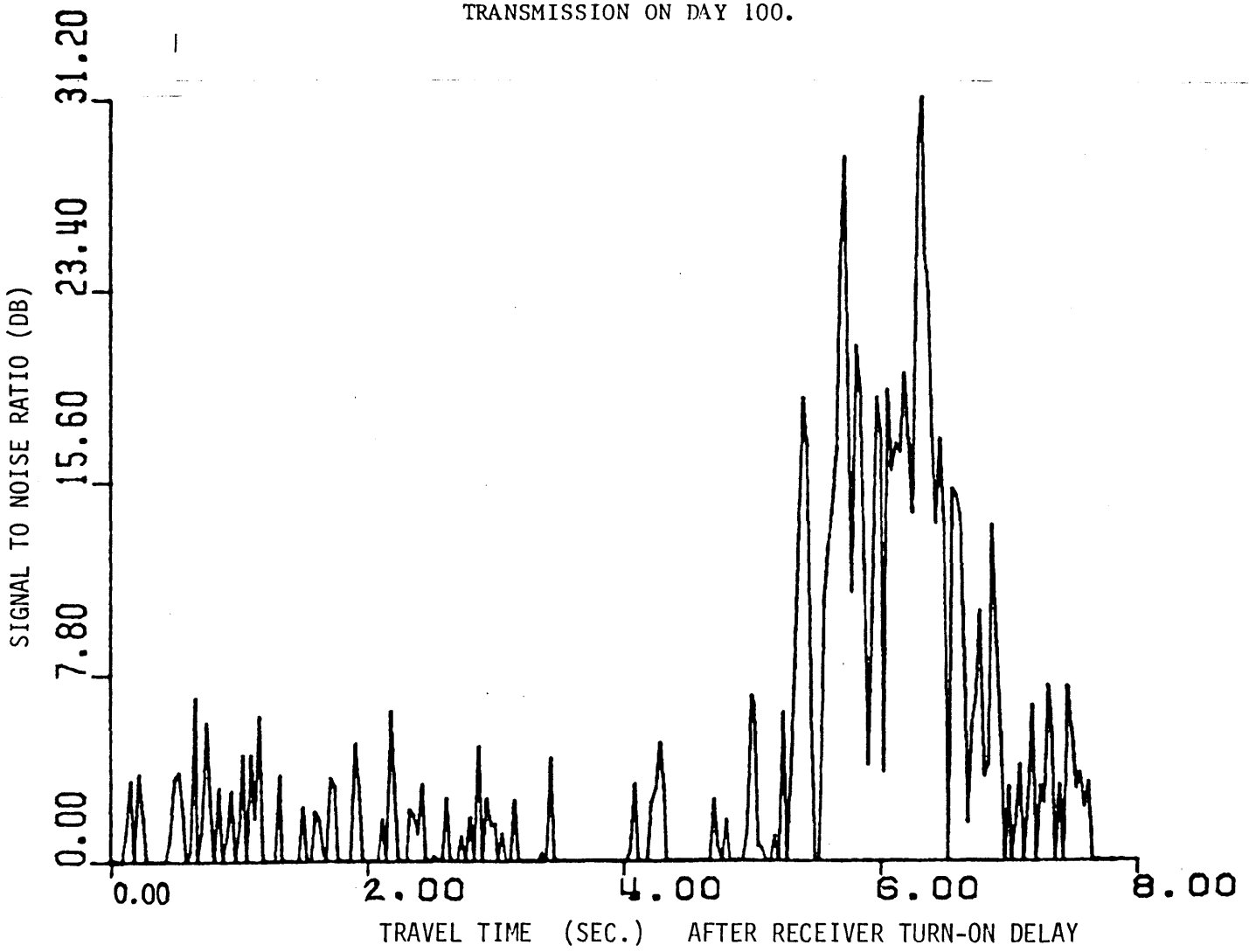


FIGURE 8.1 D ARRIVALS FOR SOURCE 2 - RECEIVER 2: FROM THE 4TH TRANSMISSION ON DAY 100.

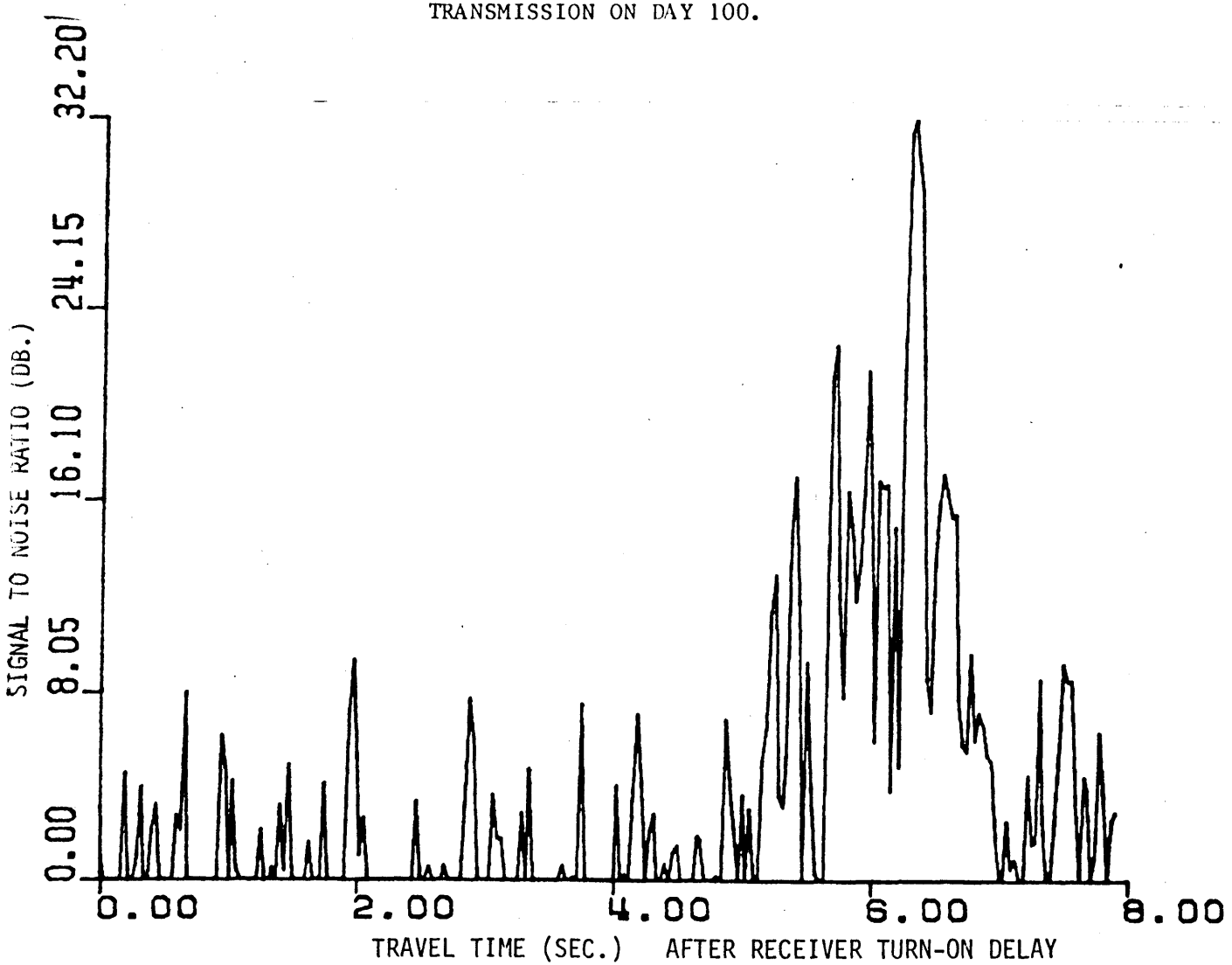
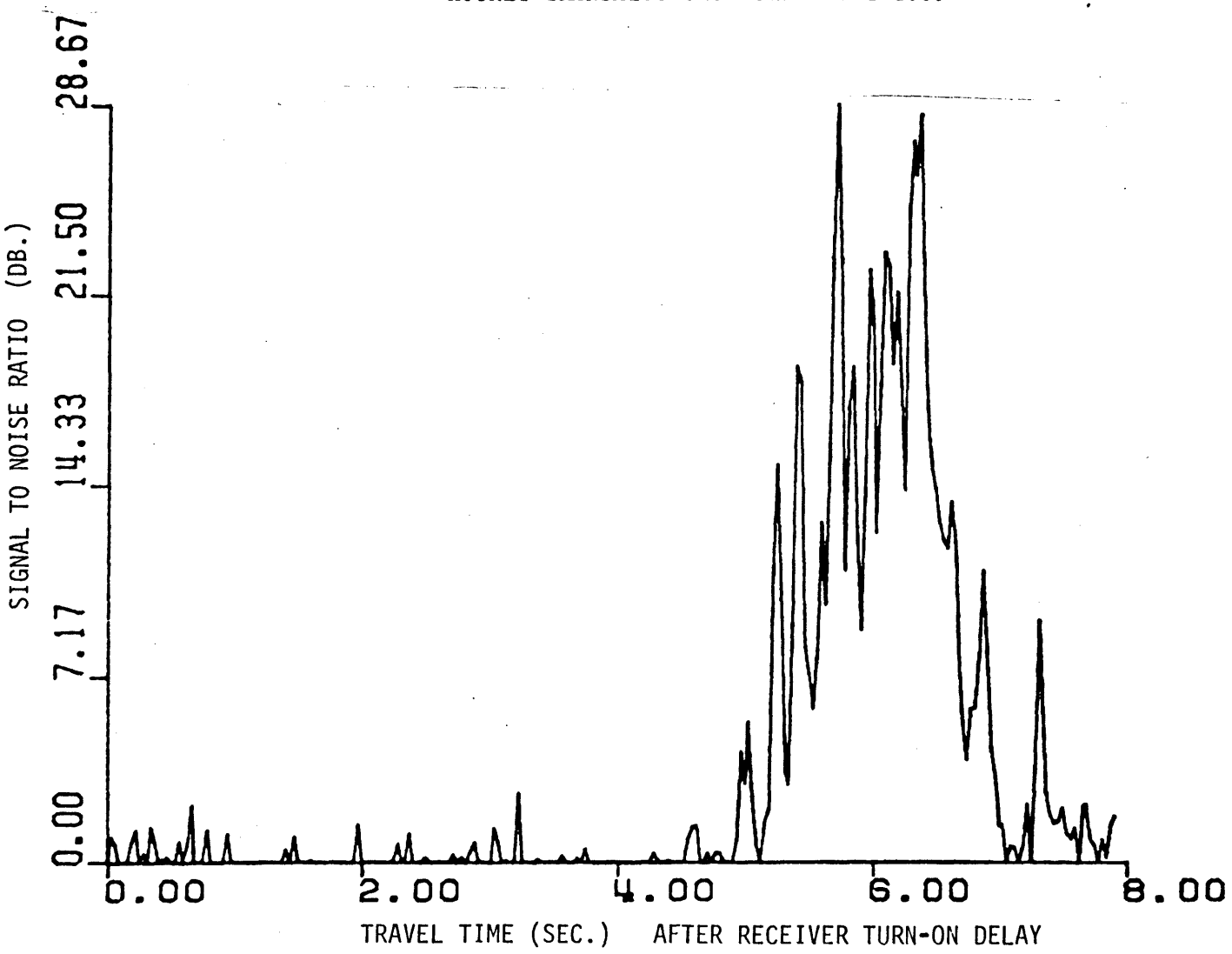


FIGURE 8.1 E ARRIVALS FOR SOURCE 2 - RECEIVER 2: AVERAGED OVER 24 HOURLY TRANSMISSIONS DURING DAY 100.



Neglecting the effects of micro-multipaths, the rms uncertainty for resolved peaks is less than 2 msec. To maintain this precision, the 254 points must be interpolated by at least 16 times, using band limited interpolation. During the preliminary data processing for the experiment, cubic splines were used to interpolate by 16 times. This reduced the sample spacing to 1.95 msec, limiting quantization errors to the level of the precision.

Each hour, each receiver stores 4 sets of correlation peaks, one for each source. Each set of peaks will be called an "arrival pattern". Figures (8.1 A-D) show the changes in these arrival patterns over 4 sucessive hours. The hourly returns show significant variations in amplitude, at least partly as a result of the micro-multipath interference described above. The arrival times in the pattern also change in response to the internal waves and tidal currents as well as the mesoscale field. Although the inverse problem could in principle include both internal waves and tides, it is easier, at least for the purposes of this thesis, to average the arrival patterns over a day to eliminate much of the rapid variation. The simplest way to perform the average is to add up all returns for a given day, producing a smoother pattern (Figure 8.1 (E)) which makes it somewhat easier to pick out arrival peaks.

8.2 PEAK FINDING AND TRACKING

The next step in processing is "peak finding", in which the peaks of the interpolated arrival pattern are located and stored. Peak location (arrival time) and signal to noise ratio are saved for all peaks above a cut-off signal to noise ratio which is set in order to screen out most of the peaks due to acoustic noise. The signal to noise ratio is saved because the uncertainty of the peak time depends on the S/N ratio. The sets of stored peaks form a time series, one for each source-receiver pair, which can be displayed to show the evolution of the acoustic ray arrival times over the course of the experiment (Figures (1.6) or (8.2)). The continuity of the pattern of distinct ray arrivals is clear over the entire experiment in this figure.

The arrival patterns in figure (8.2 A) have been corrected for mooring motion and clock drift by using the measurements made by the acoustic mooring tracking and the rubidium-referenced measurements of the frequency shifts of the quartz oscillators in each instrument. In the case of clock drift, the arrival pattern for each source-receiver pair was shifted in the wrap-around 7.9375 second window to compensate for the clock errors of the two instruments involved. The mooring motion corrections were made by computing the changes to the horizontal range between the

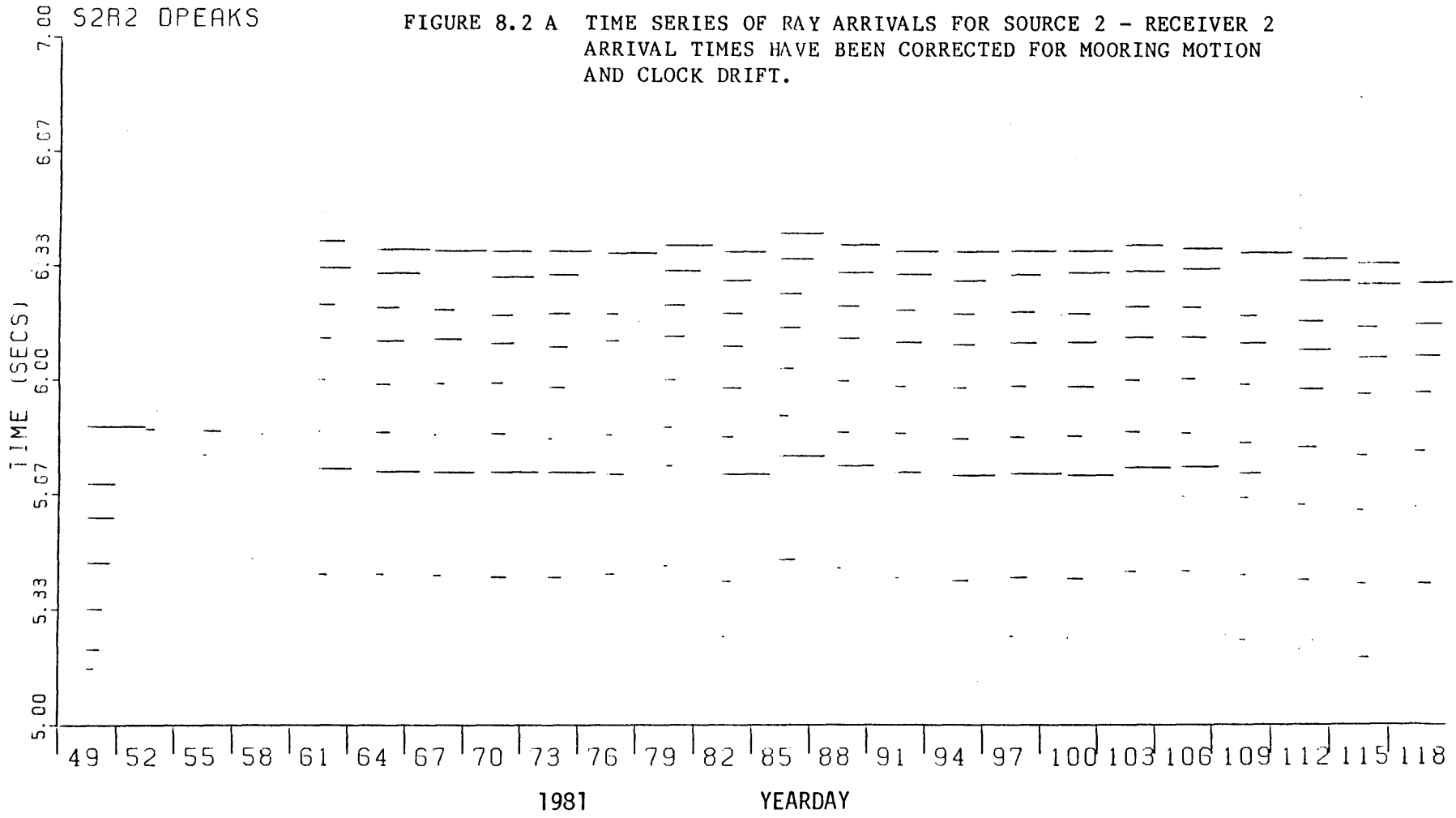
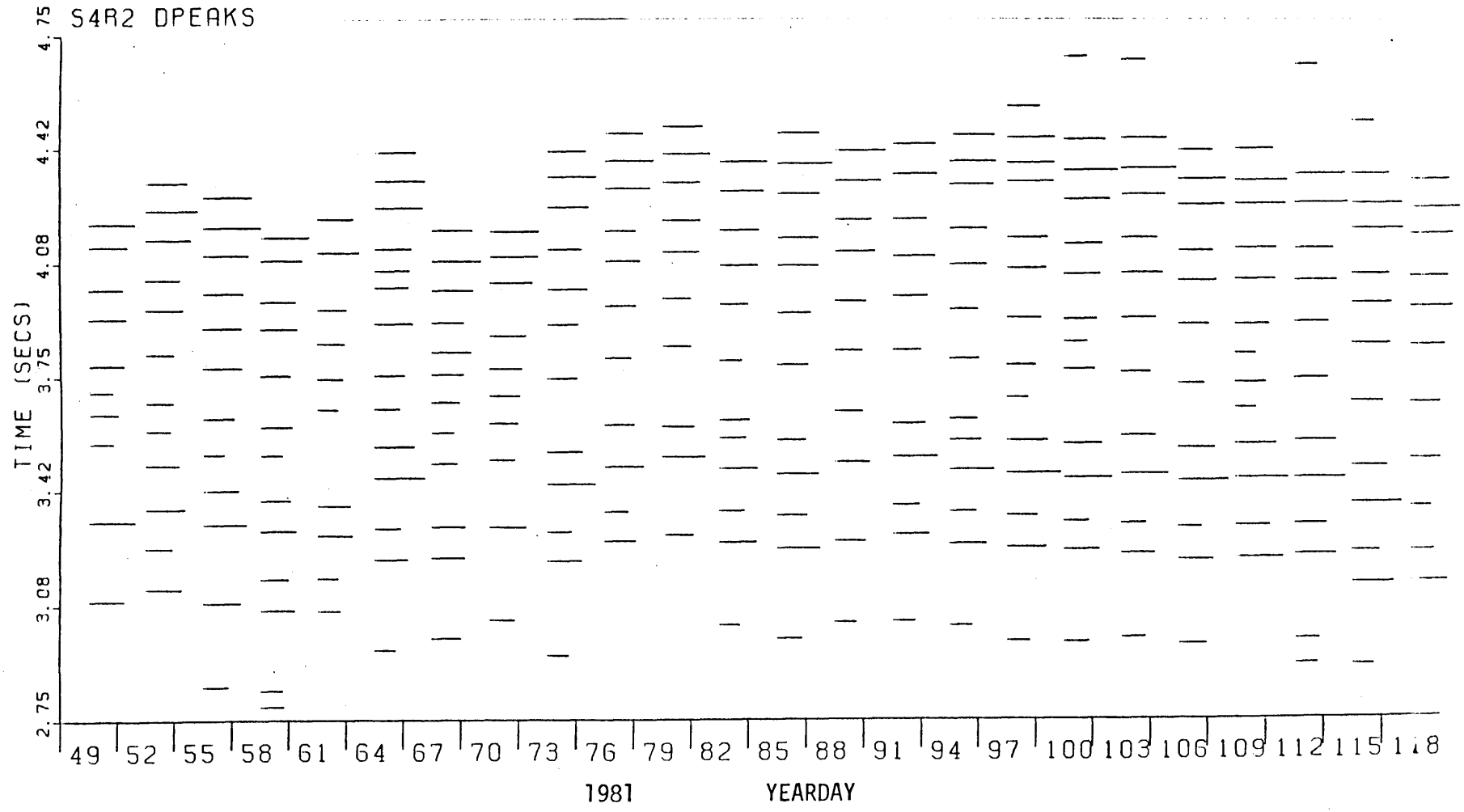


FIGURE 8.2 A TIME SERIES OF RAY ARRIVALS FOR SOURCE 2 - RECEIVER 2 ARRIVAL TIMES HAVE BEEN CORRECTED FOR MOORING MOTION AND CLOCK DRIFT.

FIGURE 8.2 B TIME SERIES OF RAY ARRIVALS FOR SOURCE 4 - RECEIVER 2
ARRIVAL TIMES HAVE NOT BEEN CORRECTED FOR MOORING
MOTION, BUT CLOCK DRIFT HAS BEEN REMOVED.



two instruments due to their motion and dividing by an averaged sound speed to obtain a travel time correction which was also used to shift the return pattern in the window. The effect of the corrections is clear if an uncorrected time series (Figure (8.2 B)) is examined. Note that the continuity of the arrival pattern is conserved, in spite of the large travel time changes due primarily to the motion of the mooring.

The next step in the data reduction attempts to quantify the continuity of the arrival pattern. Each important peak in the pattern is selected and tracked over the entire time series, producing a time series of arrival times associated with that particular peak. The process of peak tracking is nearly completely dependent on the robustness of the arrival pattern as the criterion for following a particular peak as the pattern moves around in response to the ocean. Figure (8.3) shows the results of the tracking step for two time series, the corrected series from figure (8.2 A) and the uncorrected peaks from figure (8.2 B). With many of the intermittent and noisy peaks removed, the pattern becomes easier to follow, even without mooring motion corrections.

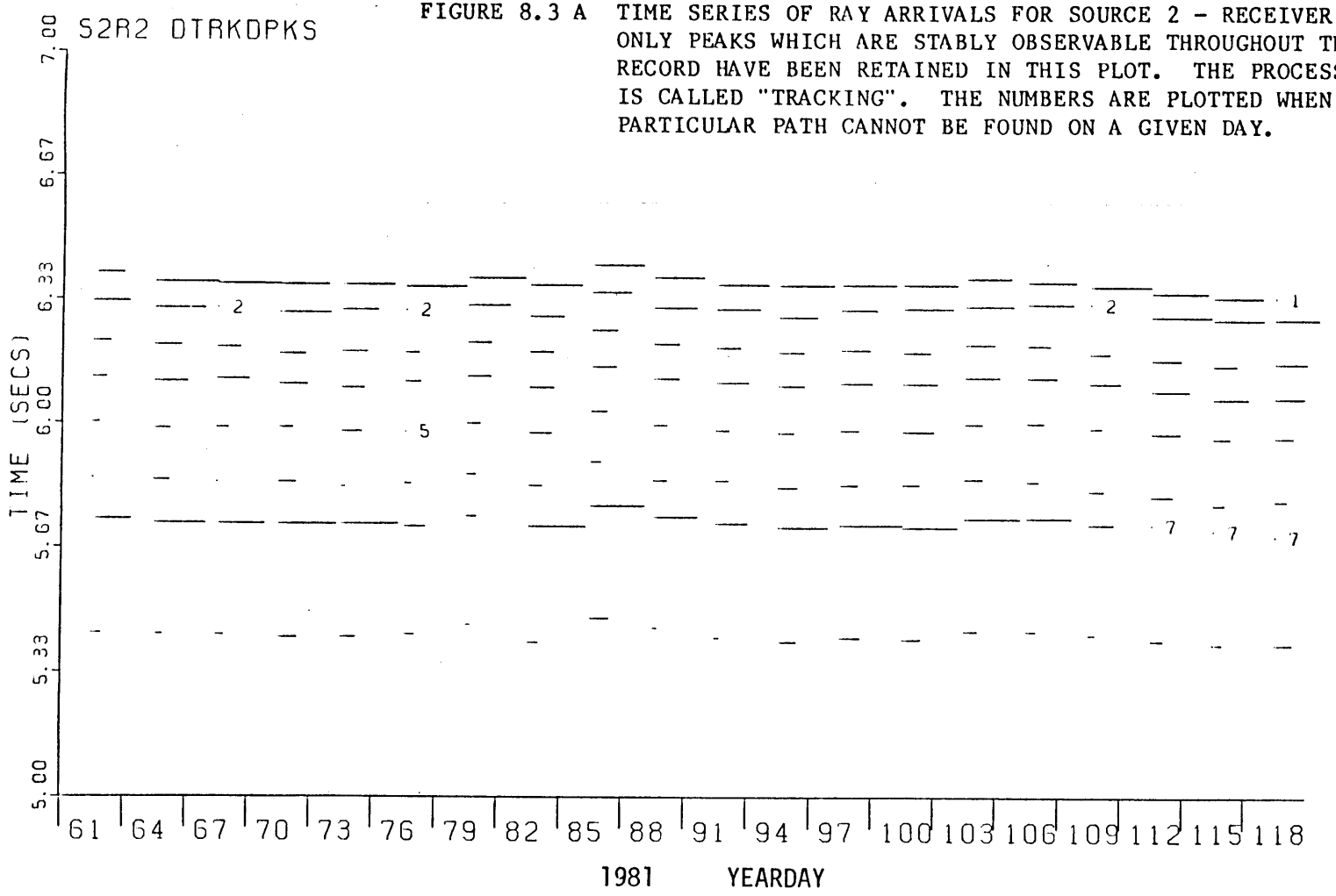
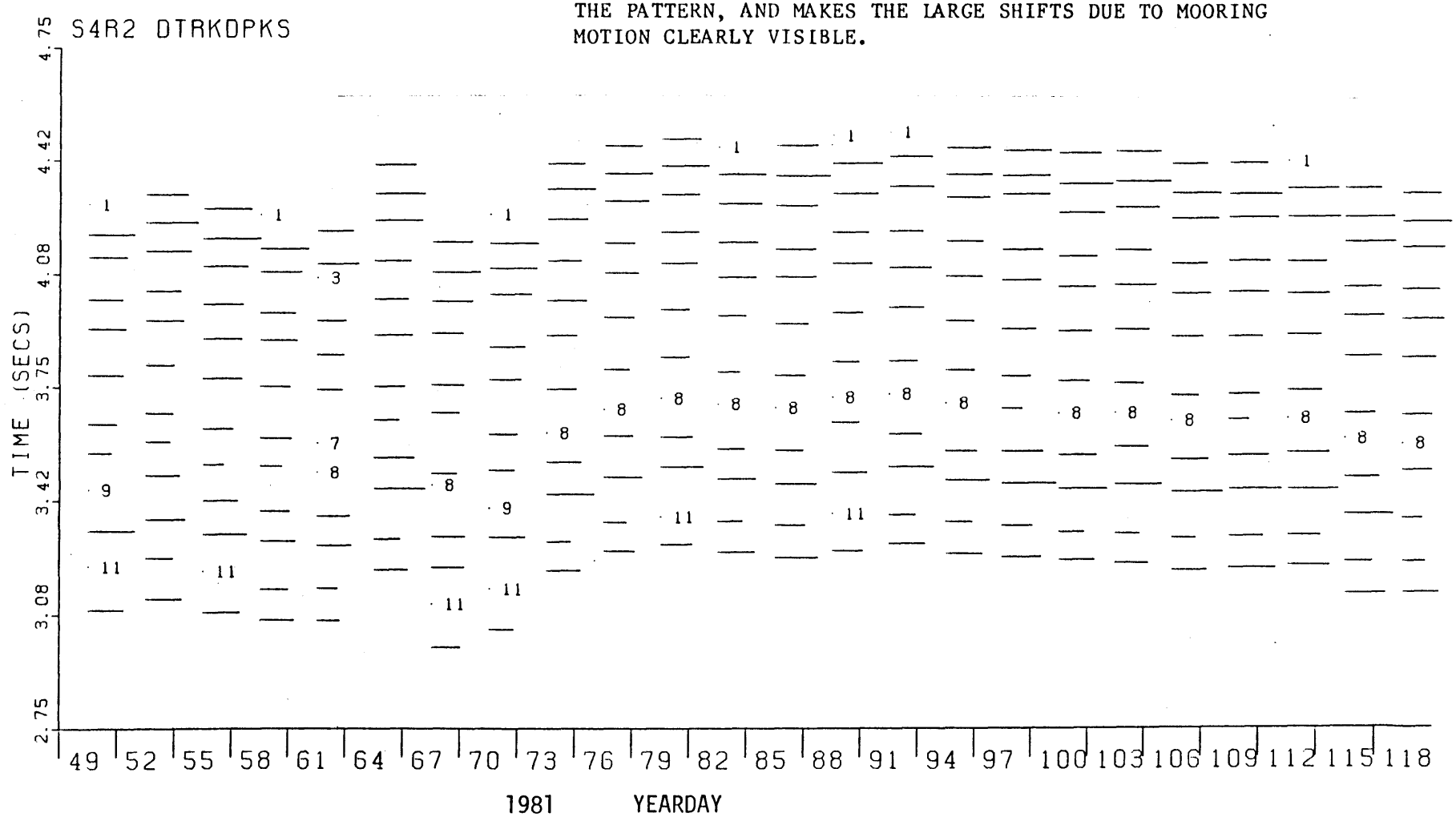


FIGURE 8.3 A TIME SERIES OF RAY ARRIVALS FOR SOURCE 2 - RECEIVER 2. ONLY PEAKS WHICH ARE STABLY OBSERVABLE THROUGHOUT THE RECORD HAVE BEEN RETAINED IN THIS PLOT. THE PROCESS IS CALLED "TRACKING". THE NUMBERS ARE PLOTTED WHEN A PARTICULAR PATH CANNOT BE FOUND ON A GIVEN DAY.

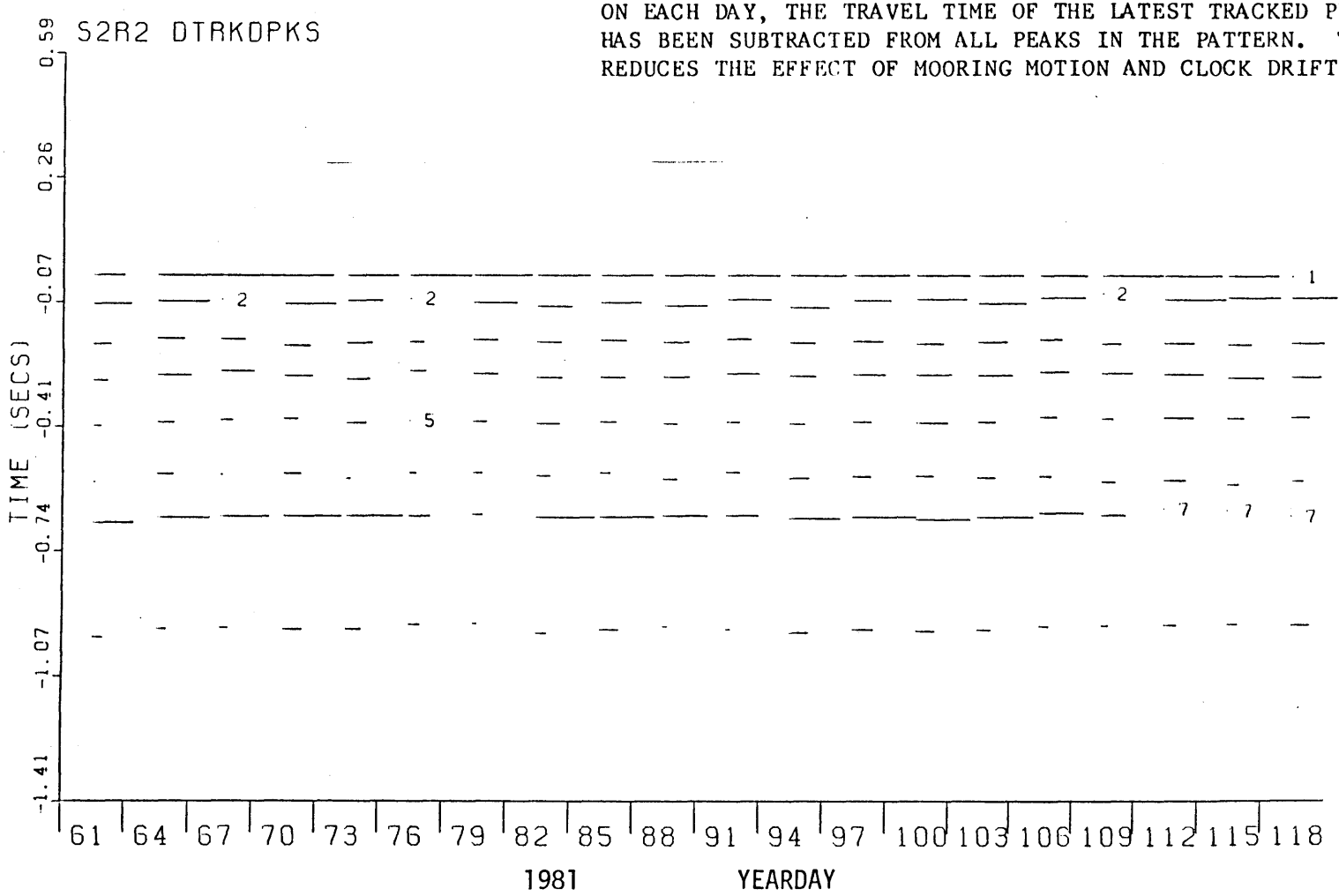
FIGURE 8.3 B TIME SERIES OF TRACKED, UNCORRECTED RAY ARRIVALS FOR S4-R2. NOTE HOW THE REMOVAL OF NOISY PEAKS EMPHASIZES THE PATTERN, AND MAKES THE LARGE SHIFTS DUE TO MOORING MOTION CLEARLY VISIBLE.



If the arrival of one of the peaks in a pattern is subtracted from the others in the pattern for each day of the record, the resultant "ray differential" times show only the distortions of the arival pattern (Figure (8.4)). Ray differentials are thus immune to instrument clock shifts, which just displace the arrival pattern. Because much of the mooring displacement causes the arrival pattern to translate with minor distortions, the ray differentials also screen out much of the noise due to mooring motion. Although both the ocean and the movement of the mooring both translate and deform the pattern of ray arrivals for a given source-receiver pair, the modes of change can be at least partially distinguished, and this is the key factor in allowing useful inversions in the presence of large, uncorrected mooring motions.

Each tracked peak presumably corresponds to a distinct ray path through the ocean, and the next step in the data reduction is to determine the ray paths for the arrivals observed in the data. This procedure, called "ray identification", also depends on the pattern of the ray arrivals. Rays are traced numerically using a typical sound speed state for the area, range dependent or independent, and the pattern of numerical ray arrivals is compared with the tracked peak pattern on a given day or series of days (Figure (2.6)). The identification can be done manually or automatically, provided that the pattern contains enough information to make an unambiguous match.

FIGURE 8.4 A TIME SERIES OF RAY-DIFFERENTIAL TRAVEL TIMES FOR S2-R2. ON EACH DAY, THE TRAVEL TIME OF THE LATEST TRACKED PEAK HAS BEEN SUBTRACTED FROM ALL PEAKS IN THE PATTERN. THIS REDUCES THE EFFECT OF MOORING MOTION AND CLOCK DRIFT.



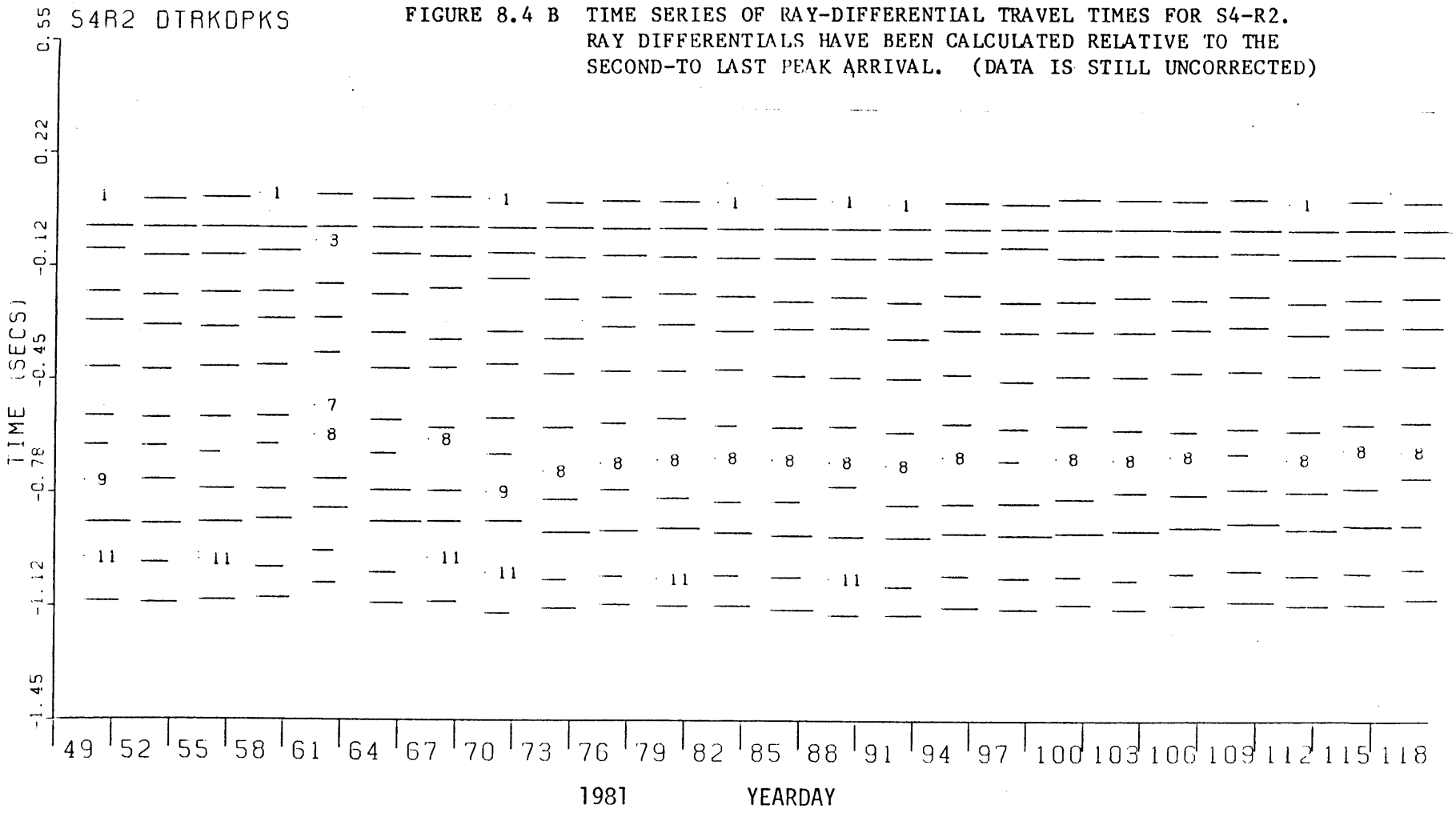


FIGURE 8.4 B TIME SERIES OF RAY-DIFFERENTIAL TRAVEL TIMES FOR S4-R2. RAY DIFFERENTIALS HAVE BEEN CALCULATED RELATIVE TO THE SECOND-TO LAST PEAK ARRIVAL. (DATA IS STILL UNCORRECTED)

If the pattern is not detailed enough to make ray identification certain, then several alternatives are available. The Scripps receivers used a vertical array of 4 hydrophones, allowing beam forming to estimate arrival angles of the rays corresponding to the peaks in the pattern. If this information is not available, an approximation to beam forming can still be done using the motion of the mooring. The travel time shifts for a given ray due to mooring motion are sensitive to the angle that the ray makes with the horizontal at the instrument which is moving. If the mooring moves on a short time scale, as a result of inertial waves or tides, for example, then the shifts of the tracked paths provide a consistency check on a tentative ray identification, provided mooring motion tracking is available. In the future, a generalized beam-forming routine could be used to resolve the angles in an optimal way, capitalizing on the motion of the mooring.

If mooring tracking is not available, then the inversion will provide the check on ray identification through an examination of residuals. Different modes of variation of the travel times in a pattern correspond to different physics, and the residual noise level can be robustly identified. Systematic errors above this level will show up in the "residuals" calculated by removing the effects of mooring motion and clock offsets from the travel

time data. If the rays have been incorrectly identified for a particular source-receiver pair, the residuals for that pair will reveal the mismatch. This technique was used to correct some of the preliminary identifications in the first stage of processing the tomography data.

8.3 PLANNED IMPROVEMENTS

Some of the techniques described above are by no means final, and will be improved for the "ultimate" inverse or for future experiments. The interpolation and peak finding steps could be replaced by a maximum entropy algorithm, treating the 254-point arrival pattern as a spectrum. Fourier transforming the pattern yields 254 "lagged covariances", which are then fed into a maximum entropy algorithm to produce the poles of the "spectrum", which correspond to the peaks of the arrival pattern, with resolution equivalent to an infinite number of interpolated points (J. Catipovic, personal communication, 1982).

At the same time, the simple averaging scheme employed in the first pass processing will be discontinued, so that peak finding is done for the hourly returns. This is necessary to allow the mooring motion beamforming mentioned above, and avoids possible problems with a rapidly shifting peak, which may appear as two peaks if the simple summation is used. An hourly time series of peaks could be tracked in the same way that the daily peaks were, and then the averaging to remove tides and internal waves would take place path by path, weighted by the uncertainty of each peak. The un-averaged time series would be useful if the inversion was to be extended to the shorter time scales.

CHAPTER 9

ESTIMATORS USED FOR THE 1981 TOMOGRAPHY EXPERIMENT

9.1 THE MODEL

Most of the discussion of inverse methods presented so far has been general, in an attempt to show the interconnections and justifications of methods which often seem quite distinct. I will now discuss in detail the inversion techniques used with the data from the 1981 tomography experiment, after data processing as described in Chapter 8. The formalism of the stochastic inverse will be used throughout the following since it allows considerable flexibility, including a continuous representation of the unknown field. In any case, it was shown (in Chapter 5) that the stochastic inverse is equivalent to several other forms of linear least-squares inversion, so there is no reason to use a different form.

At this stage, only travel time data have been used in the inverse, to allow independent comparison with the conventional measurements taken during the experiment, but any and all of the other data types can be included, and will be used in the future. The transmissions in the 1981 experiment were one way only, so that the travel time changes due to ocean currents were not specially resolved,

and have so far been neglected in comparison with the travel times due to sound speed changes. The travel time errors incurred by this assumption should be order 2 msec, comparable to the other error sources. As the processing of the data improves, currents will be incorporated as part of the inverse, although the resolution will not be great.

In order to use the stochastic formalism, it is necessary to define a mean state for the sound speed and the expected covariance around this basic state. Because we are interested in deriving reliable snapshots of the evolution of the sound speed anomalies due to mesoscale dynamics, we are more interested in the minimum variance properties of the estimator than in its possible bias. For this reason, the basic state need only be specified near enough to the true state to avoid problems with linearization. This means that most any archived estimate of the local mean sound speed is adequate for use as a mean state, although the closer the assumed mean state is to the true mean the smaller the variance around the mean will be, increasing the effectiveness of the estimator.

For the initial estimates from the 1981 experiment, a simple average of the CTD casts during the first NOAA survey of the area was chosen to be the basic state,

(Figure 1.1), more for convenience in coordinating between institutions than any other reason. The basic state was taken to be stationary and horizontally homogeneous, $C_o(\underline{x},t) = C_o(z)$, both for simplicity and because the data available to date are inadequate to support any assumptions to the contrary.

The estimate of covariance for the sound speed anomaly is also derived from archived data, and is then used with the forward problem to calculate the expected data-data covariance matrix. The decomposition into vertical modes with horizontally varying amplitudes has been discussed above, and this model will be used throughout the inversions to follow:

$$C'(\underline{x},t) = C(\underline{x},t) - C_o(z) = \sum_{i=1}^{M} F^c{}_i(z) \cdot \eta_i(x,y,t) \qquad (1)$$

The modes chosen as a basis are the empirical orthogonal functions of sound speed variation for the MODE experiment (Figure 9.1). This basis was chosen before the data from the experiment were available, so that the model for vertical structure would be independent of the traditional measurements made during the experiment. Because the MODE EOFs were calculated relative to the average sound speed profile from the MODE experiment, it would have been more logical to use the MODE averaged sound speed profile as a reference, rather than the average of

FIGURE 9.1 A,B,C,D: FIRST 4 EMPIRICAL ORTHOGONAL FUNCTIONS FROM THE MODE CTD SURVEY. CALCULATED FOR SOUND SPEED PERTURBATIONS RELATIVE TO THE AVERAGE C(Z) FOR MODE.

FIGURE 9.1 A FIRST SOUND SPEED EMPIRICAL ORTHOGONAL FUNCTION FROM MODE

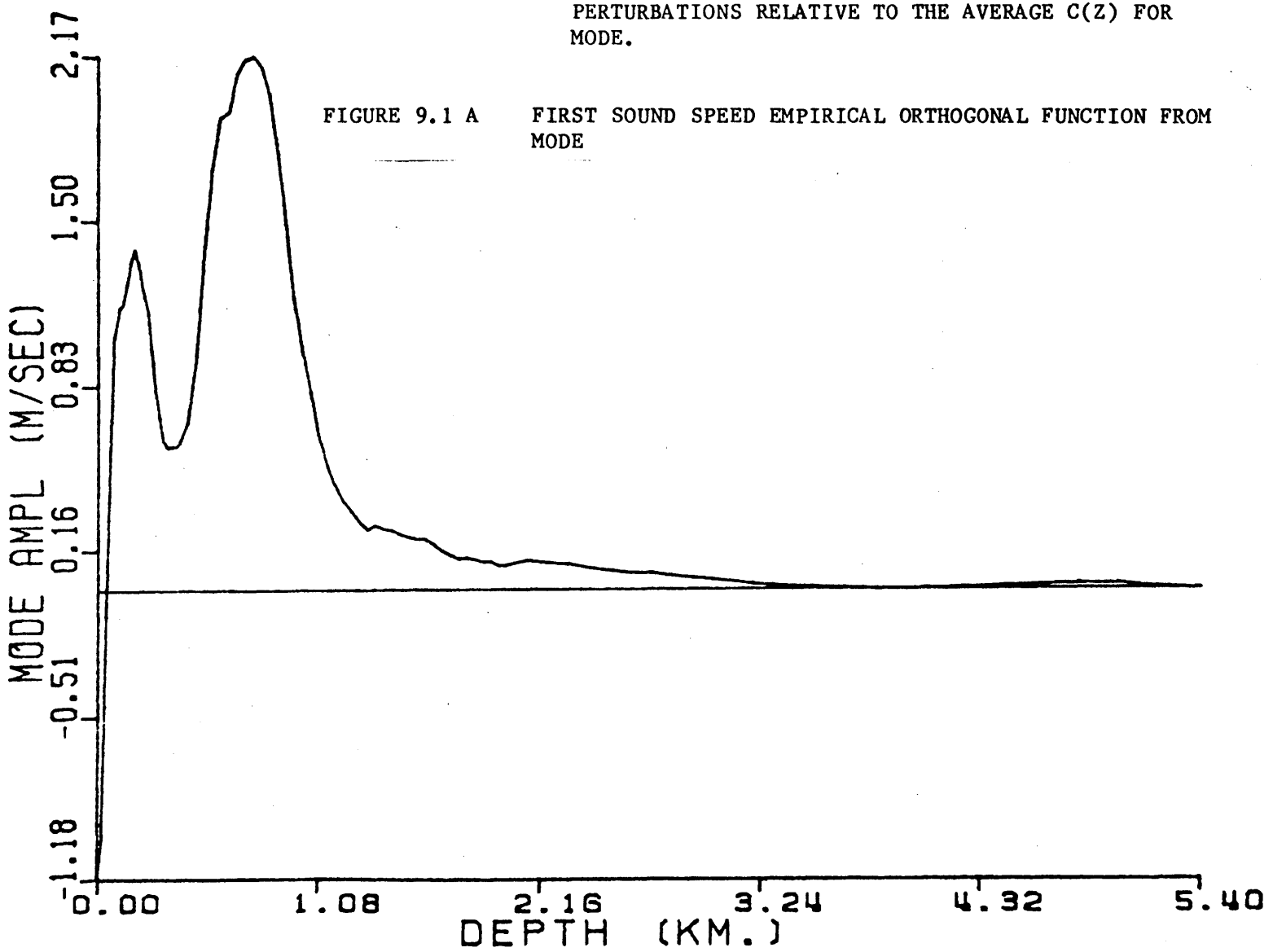


FIGURE 9.1 B SECOND SOUND SPEED EMPIRICAL ORTHOGONAL FUNCTION FROM MODE

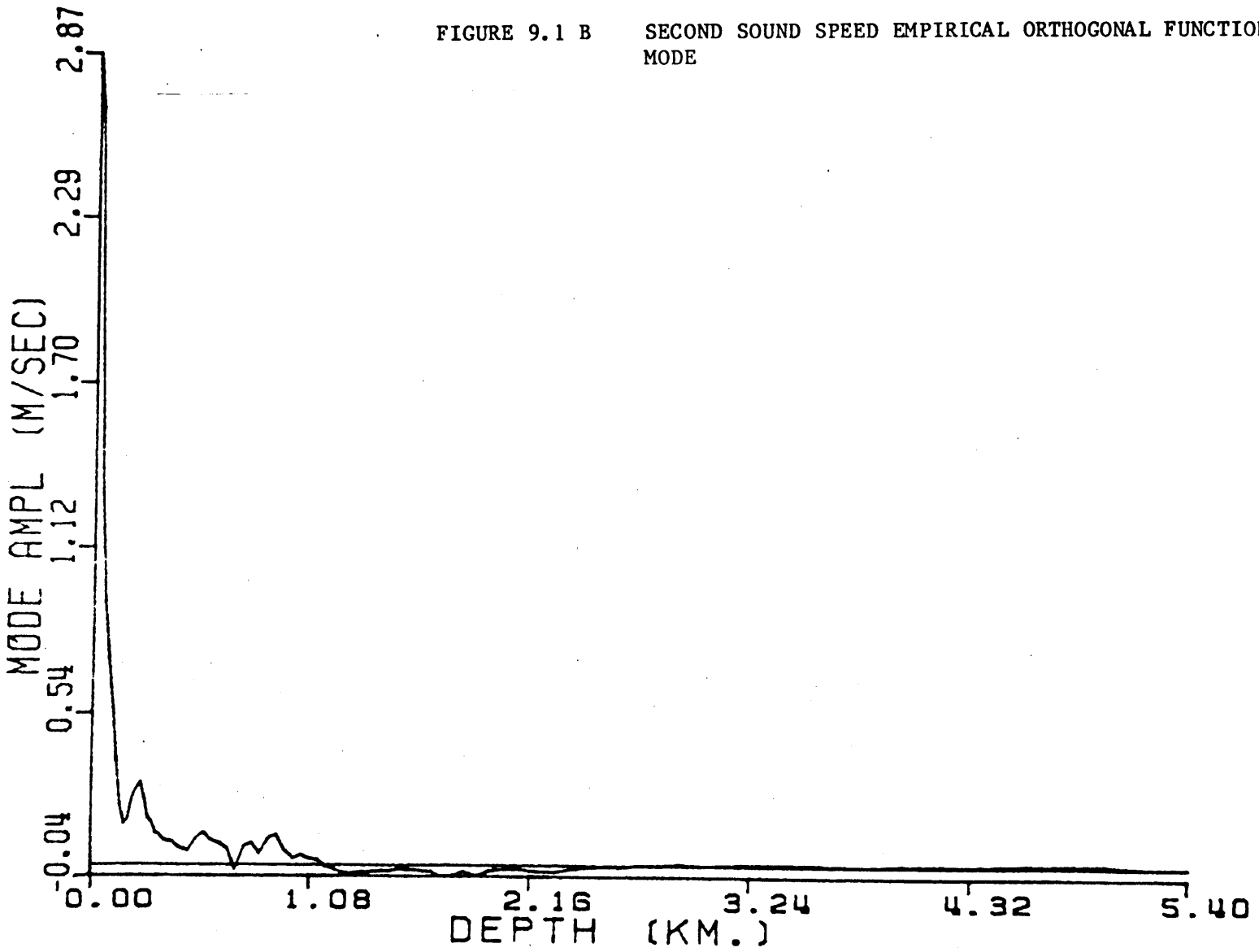


FIGURE 9.1 C THIRD SOUND SPEED EMPIRICAL ORTHOGONAL FUNCTION FROM MODE

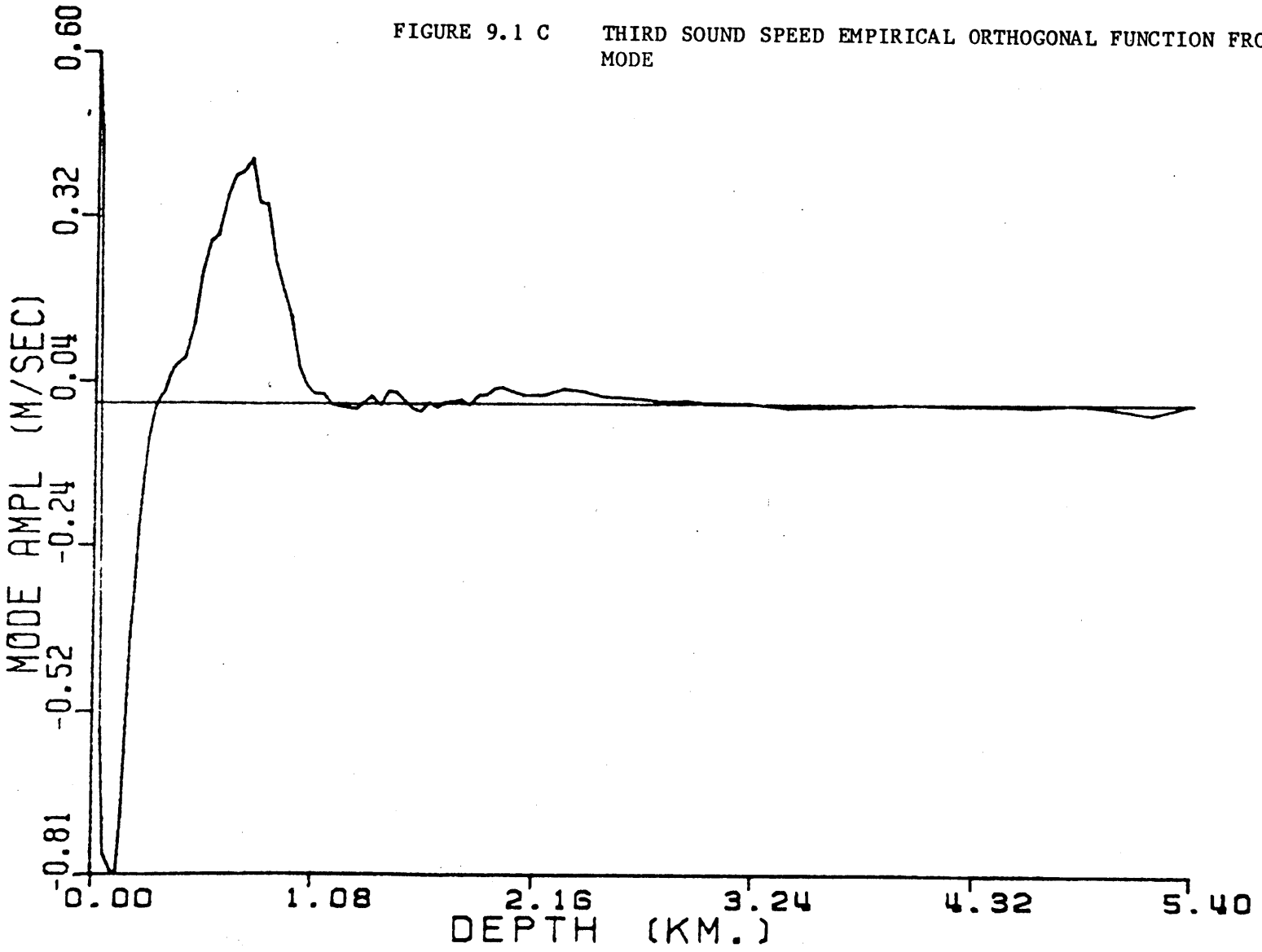
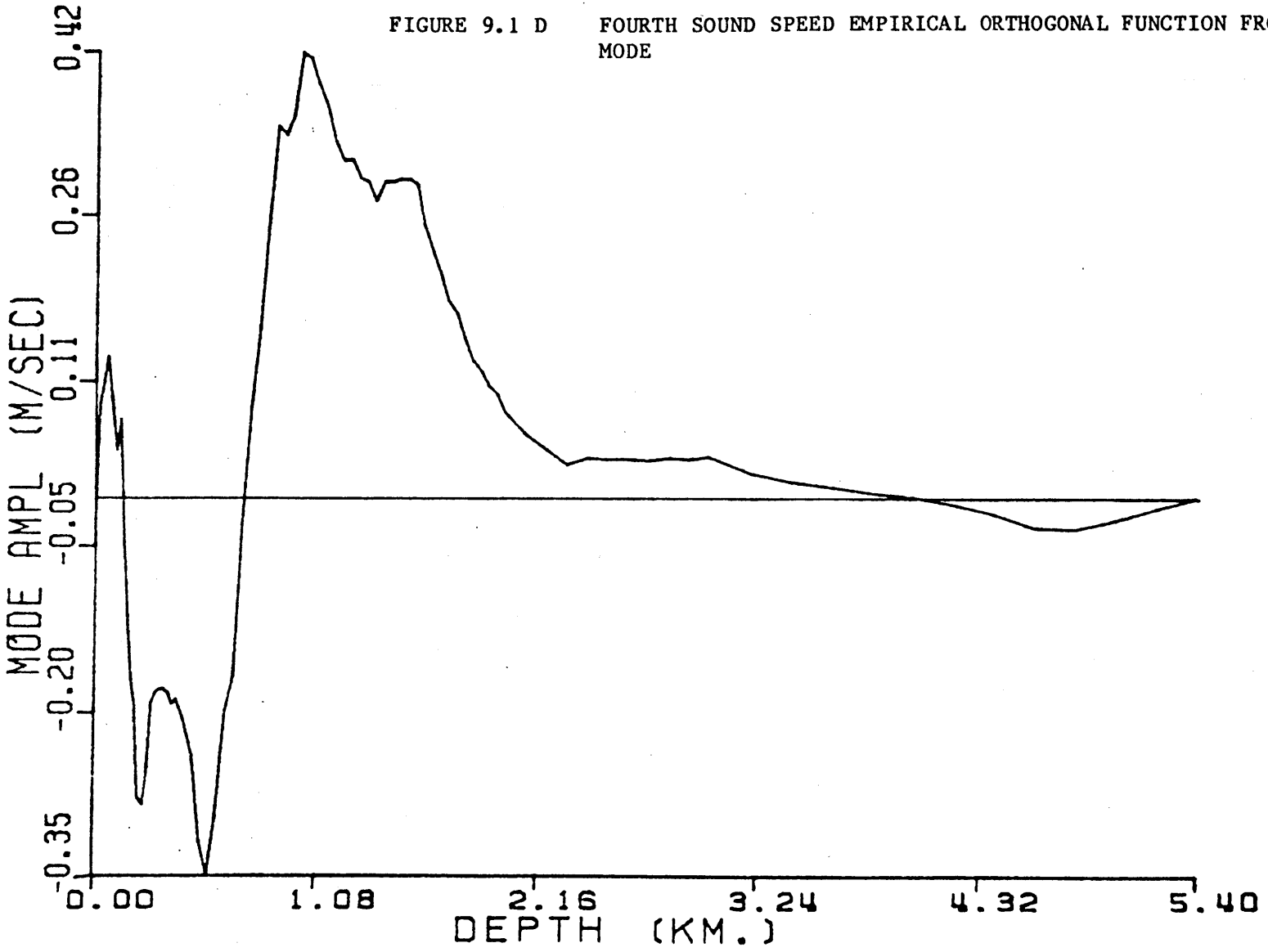


FIGURE 9.1 D FOURTH SOUND SPEED EMPIRICAL ORTHOGONAL FUNCTION FROM MODE



the first NOAA CTD survey. In future inverse calculations, the MODE $C_o(z)$ profile will be used with the MODE EOFs, or else analytical modes will be used, relative to an appropriate basic state.

The "analytical" modes (solutions of the vertical structure equation discussed in Chapter 3) should be calculated using an estimated climatological mean buoyancy frequency profile. Given a basis set of displacement modes, conversion to density modes or sound speed modes is possible, given mean temperature and salinity profiles (Chapter 6 above). The EOFs allow variance in the upper layers of the ocean, presumably due to seasonal effects, (see Figure 9.1), while the analytic modes have nodes at the surface by construction (Figures 9.2 A-D). If an analytical mode basis is used, then surface-intensified modes must be added to those calculated using quasi-geostrophy. These may either be specified in some ad hoc way, such as layers, or modes derived from mixed layer or climate models might be incorporated.

The expected variances of the modes as derived from MODE CTD data are listed in Table (9.1), and were used to construct the total data-data covariance matrix. The overall energy level is arbitrary, so the weighting by expected variances need only yield a correct signal to

FIGURE 9.2 A FIRST BAROCLINIC MODE (IN TERMS OF DENSITY VARIATIONS) CALCULATED BASED ON THE AVERAGED BUOYANCY FREQUENCY PROFILE FROM MODE.

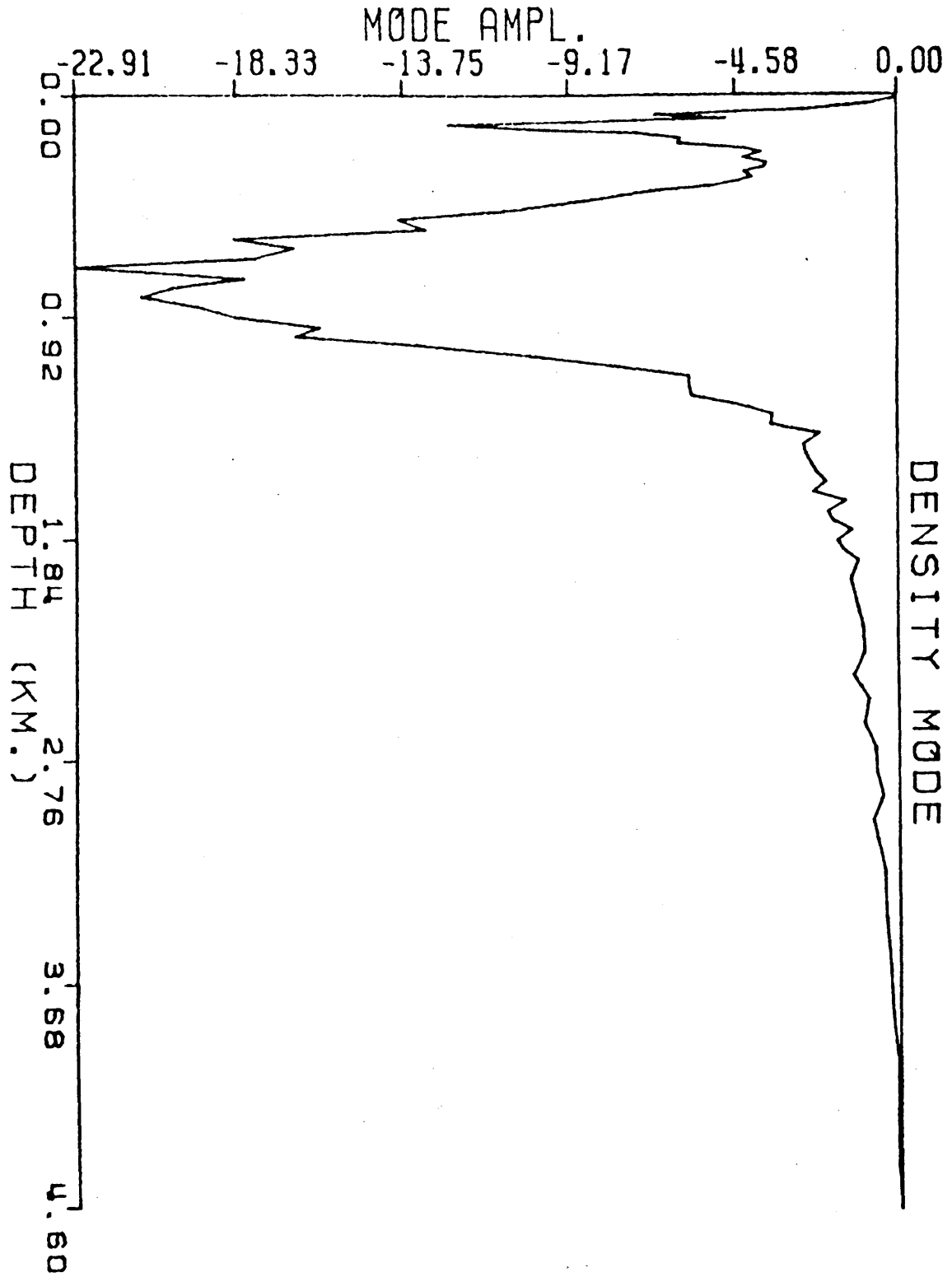


FIGURE 9.2 B 2ND BAROCLINIC MODE (IN TERMS OF DENSITY VARIATIONS) CALCULATED BASED ON THE AVERAGED BUOYANCY FREQUENCY PROFILE FROM MODE.

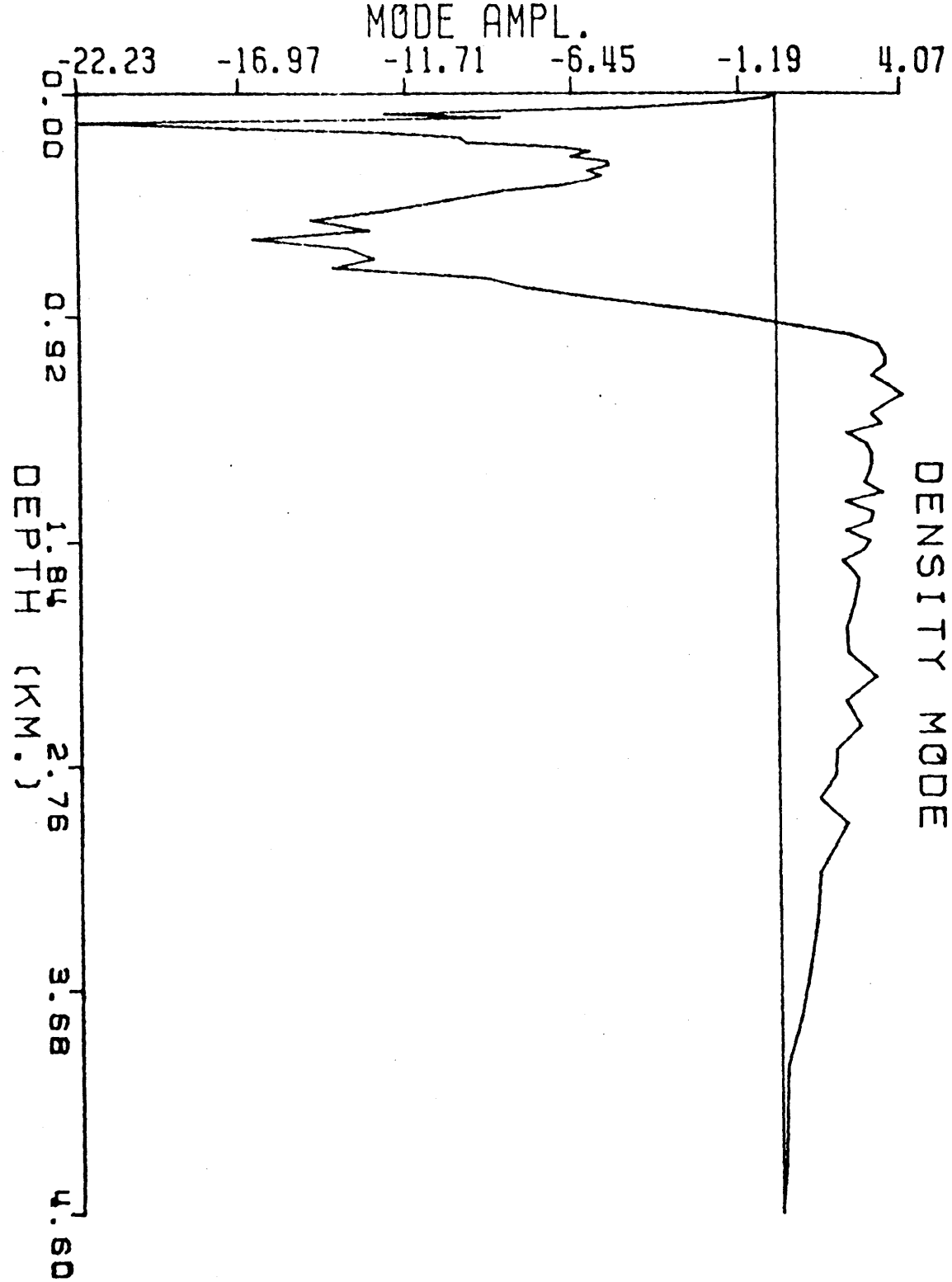


FIGURE 9.2 C FIRST BAROCLINIC MODE (IN TERMS OF SOUND SPEED VARIATIONS) CALCULATED BASED ON THE AVERAGED BUOYANCY FREQUENCY PROFILE FROM MODE.

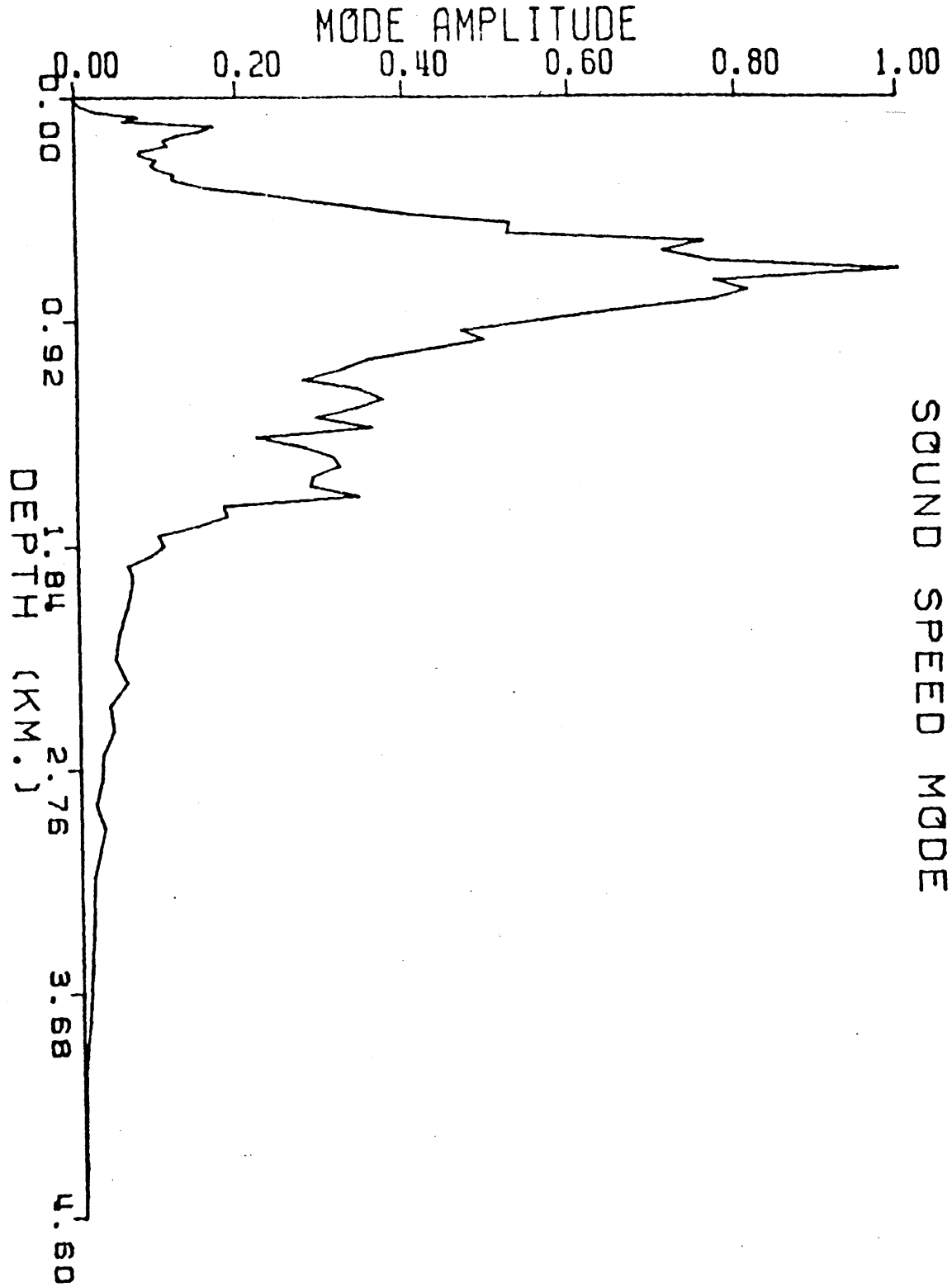


FIGURE 9.2 D 2ND BAROCLINIC MODE (IN TERMS OF SOUND SPEED VARIATIONS) CALCULATED BASED ON THE AVERAGED BUOYANCY FREQUENCY PROFILE FROM MODE.

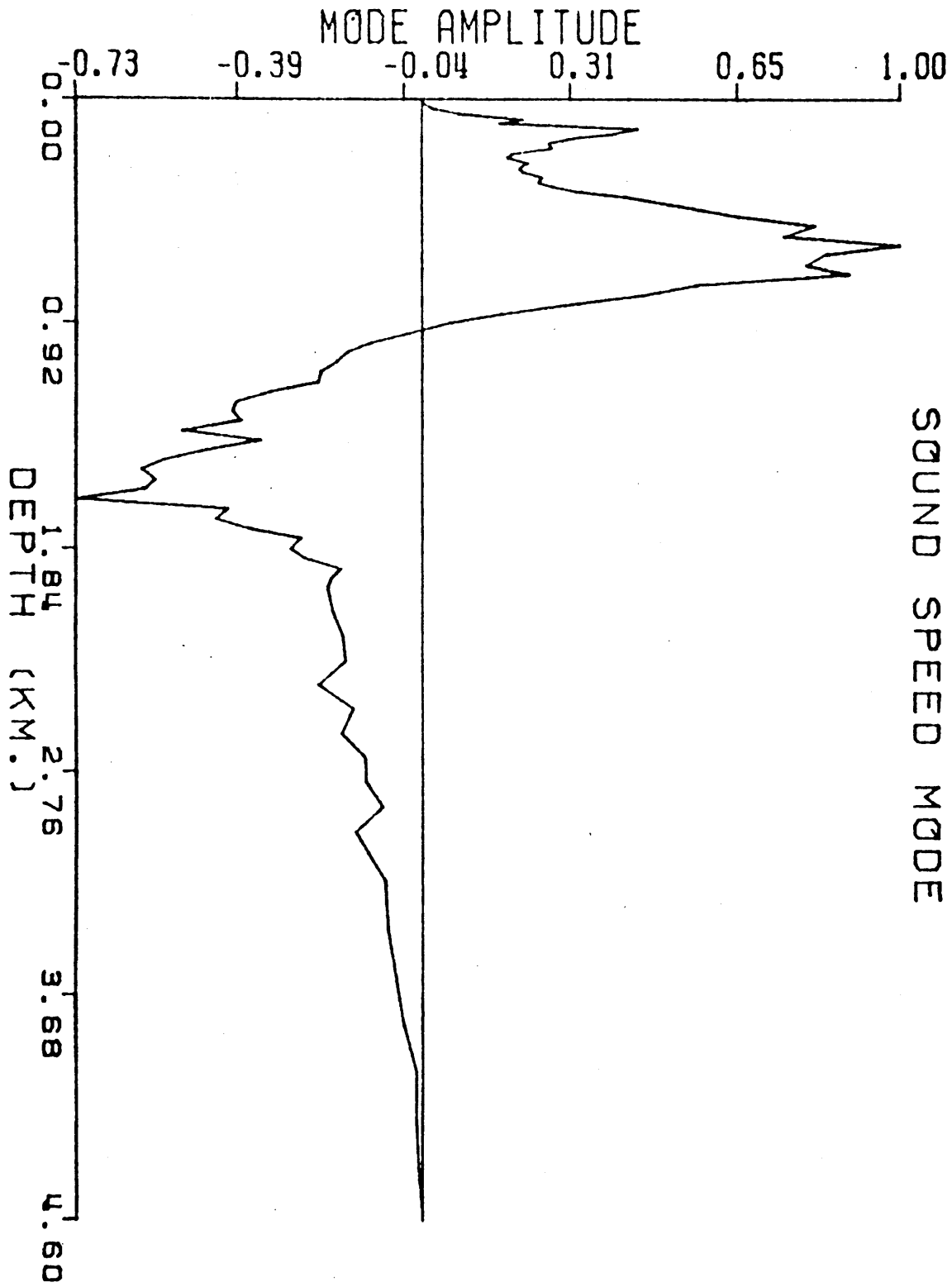


TABLE 9.1: EOF VARIANCES

Mode	1	2	3	Total
Variance of Particular mode $(m/sec)^2$				
MODE CTD data	.421	.057	.025	
Inverse	.4	.1	.1	

TRAVEL TIME

The numbers in the table are the expected standard deviations of the travel time anomalies (in msec) given for 5 different rays and divided into individual mode contributions.

	Mode 1	2	3	Total
ray (arb. index) 1	32.	1.3	3.9	32.4
21	40.1	2.1	12.2	42.3
41	46.5	3.1	16.7	49.5
51	25.0	1.4	8.2	26.4
55	17.0	3.4	4.1	18.

noise ratios. The inverse operators (estimators) derived in the course of playing with the data were not sensitive to these weightings, but order of magnitude increases in the estimated error variances can significantly decrease the resolution of the corresponding estimator.

Although the vertical structure has been parameterized by a finite number of modes, the horizontal structure has been left continuous, so that only the horizontal covariance function for the amplitude of each mode has been specified in advance (Figure (9.3 A)). The covariance was specified analytically, as a time-independent gaussian with an e-folding range of 100 km., and is homogeneous and isotropic, so the covariance between two points depends only on the magnitude of their horizontal separation.

$$\langle C'(\underline{x}_1,t_1)C'(\underline{x}_2,t_2)\rangle =$$

$$\left\langle \sum_{i=1}^{M} F^{c}_{i}(z_1)\cdot\gamma_i\cdot\eta_i(x_1,y_1,t_1) \cdot \sum_{j=1}^{M} F^{c}_{j}(z_2)\cdot\gamma_j\cdot\eta_j(x_2,y_2,t_2) \right\rangle \tag{2}$$

$$= \sum_{i=1}^{M} \gamma_i^{\,2}\cdot\langle\eta_i(x_1,y_1,t_1)\eta_i(x_2,y_2,t_2)\rangle\cdot F^{c}_{i}(z_1)\cdot F^{c}_{i}(z_2) \tag{3}$$

$$= \sum_{i=1}^{M} \gamma_i^{\,2}\cdot H_i(x_1,y_1,t_1,x_2,y_2,t_2)\cdot F^{c}_{i}(z_1)\cdot F^{c}_{i}(z_2) \tag{4}$$

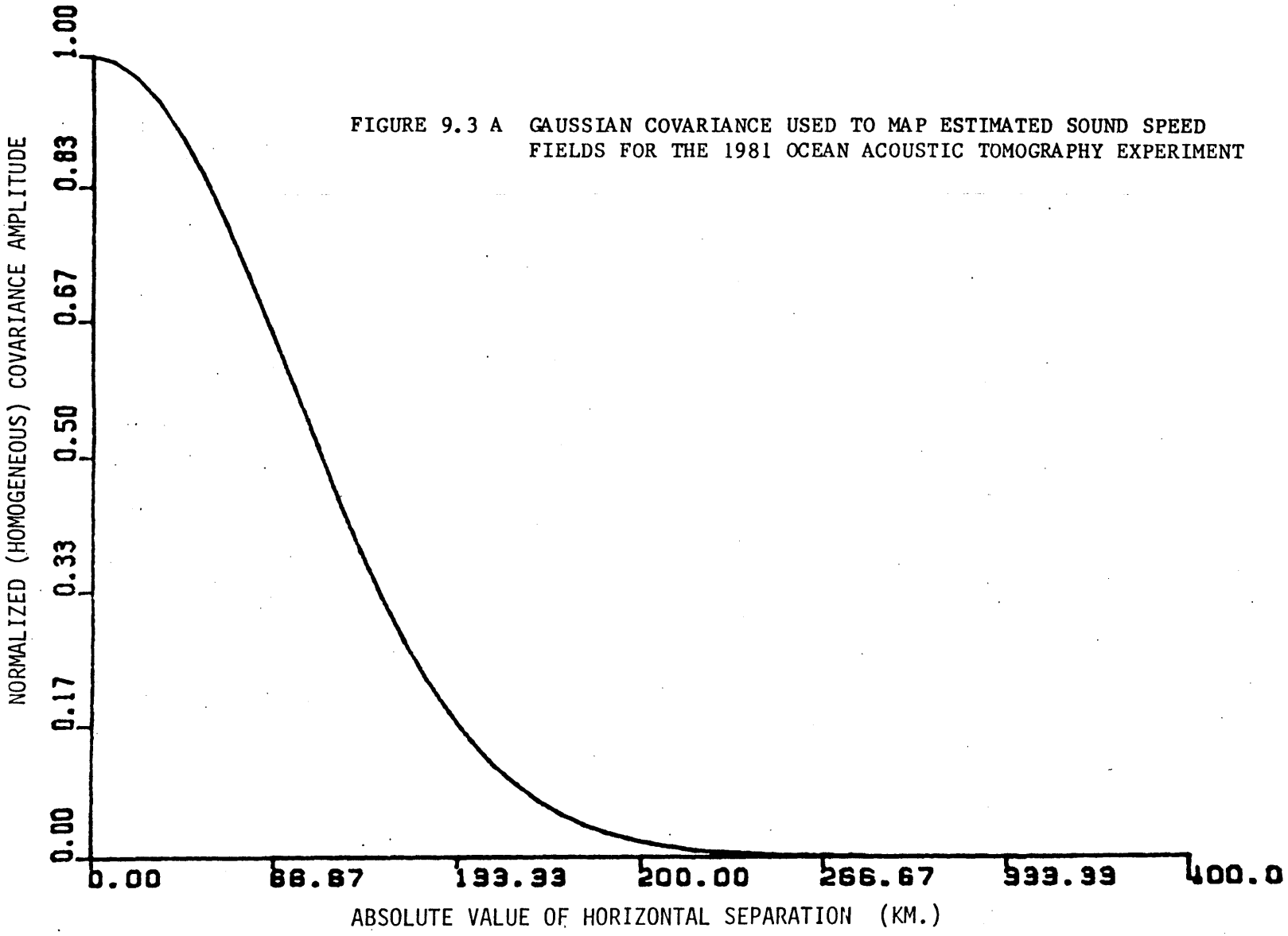


FIGURE 9.3 A GAUSSIAN COVARIANCE USED TO MAP ESTIMATED SOUND SPEED FIELDS FOR THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT

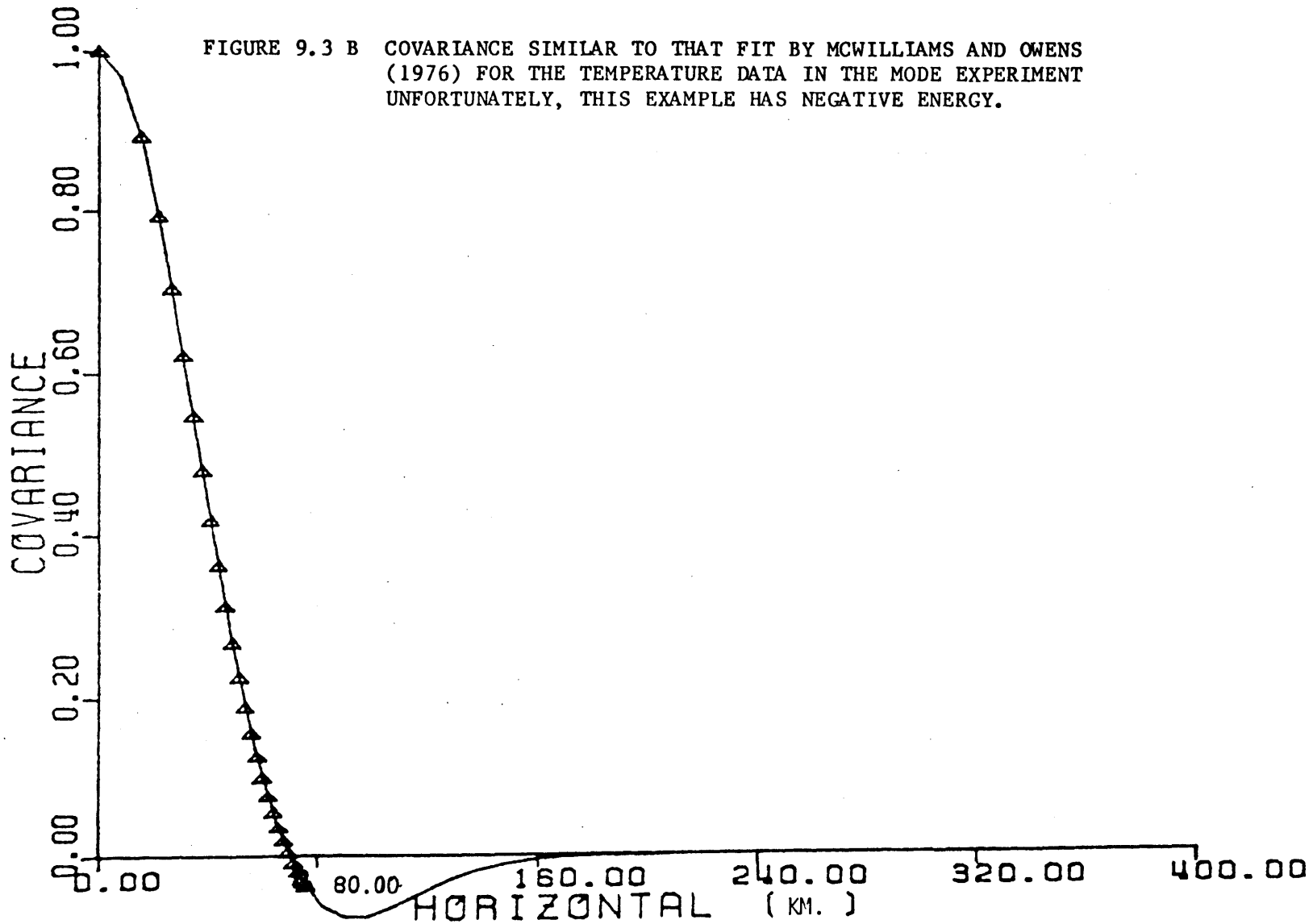


FIGURE 9.3 B COVARIANCE SIMILAR TO THAT FIT BY MCWILLIAMS AND OWENS (1976) FOR THE TEMPERATURE DATA IN THE MODE EXPERIMENT UNFORTUNATELY, THIS EXAMPLE HAS NEGATIVE ENERGY.

$$= \sum_{i=1}^{M} \gamma_i^2 \cdot H_i(R_{12}) \cdot F_i^c(z_1) \cdot F_i^c(z_2) \cdot \delta(t_1 - t_2) \qquad (5)$$

$$\text{where } R_{12} = [\ (x_1 - x_2)^2 + (y_1 - y_2)^2\ ]^{1/2} \qquad (6)$$

$$\text{and } H_i(R_{12}) = H(R_{12}) = \exp[-R_{12}^2/(100 \text{ km.})^2] \qquad (7)$$

In this case, the same covariance function was used for all of the vertical modes, although the inverse framework allows independent functions for each mode. At present the sound speed structure is the desired output of the estimator, so the "barotropic" mode, which does not displace the isopycnals, and thus cannot produce sound speed changes, has been removed from the inverse. If current meter data were used, then it would be necessary to include the barotropic mode in the model, and the covariance function for the horizontal structure of this mode would be significantly different from that used for the baroclinic modes, due to the much larger radius of deformation for the lowest mode (Hua and Owens, 1982).

Covariance shape becomes most important when estimating quantities like velocity or vorticity, which require differentiation of the fields (and, therefore, the covariance function). It is perhaps easier to understand this by considering spectral space--looking at the transform of the covariance. Taking the derivative

multiplies the energy in each wavenumber by the wavenumber itself, amplifying the energy at the small scales. Two covariances which look roughly similar may have differing amounts of small-scale energy, and each differentiation will enhance the difference. The most obvious effect of this "cascade" is in the error estimator returned by the estimation procedure.

The acoustic observations are averages, so that the data-inverse system tends to lack resolution at small scales. Thus, if two covariances have the same total energy but one has half its energy in scales too small to resolve, then at best that estimator will resolve 1/2 the expected energy as defined by the covariance function. When comparing inverse methods, the mutability of the error maps must be considered, since the sizes of the calculated error bars depends directly on the models used and the expected noise power. The error bars calculated using only data error are not as sensitive to the covariance shape, but do of course depend on the assumed error levels.

The covariance function does not need to be analytic, isotropic or homogeneous, but there is no reason to add complications not required by the archived data, in this case the MODE experiment. The energy field is certainly non-homogeneous, but it was modelled as uniform, again because of the lack of a reliable alternative model. The

temperature covariance derived from the data from MODE shows a zero crossing, indicating a wavelike character (McWilliams and Owens, 1976), Figure 9.3 B, but it is not clear that this is a robust feature. Care must be taken to choose a covariance function which corresponds to a real spectrum with positive energy, because the matrix algebra requires the covariance matrices to be positive definite. The gaussian corresponds to a gaussian spectrum, and is clearly positive definite besides being satisfyingly red.

9.2 BUILDING THE ESTIMATORS

Once the model covariance (desribing the unknown field, in this case the sound speed anomaly) has been obtained, the model-data covariance and data-data covariance matrices can be constructed as described in chapter 6. The model-data covariances were constructed for mapping to 65 points in the horizontal, at the station locations of the 65 casts in the first CTD survey. This was done to ease comparisons between the estimates of the sound speed from acoustic data and those calculated from the CTD stations. The travel times used in the inverse are selected from the set of all resolved, identified rays which are available on the day for which the inverse is to be calculated. The inverse is at present time-independent, so that the maps are assumed to have no coherence between them, and each uses only data on a single day. The number of rays available changes day by day, so each map is made from a different set of rays, weighted using the error estimates for that day.

The model-data covariance matrix is calculated for all data and then saved, so that columns are selected to match the data available on any given day. In the same way, data-data covariance matrices for each of the vertical modes and the mooring motion are saved, and a properly

weighted combination is constructed for each day to match the expected noise, mooring motion variances, and to conform to the available data. The inverse operator is thus specific to a single day, even though the basic covariances were specified without time dependence. Time dependent covariance functions were not used in the demonstration inverses on the 1981 data because they require assumptions which can be controversial, and might render the resulting maps suspect, in spite of (or because of) the increased resolution and data error immunity that such assumptions foster. The assumption that sucessive maps are independent snapshots is certainly robust, but it is clear that the mesoscale ocean changes little on that time scale, and future work will explore the use of time-dependent covariances for improving the inversions.

The travel-time data has so far been used in two forms; as differences between "corrected" travel times observed for the same path on different days (called "day differentials"), (corrected for all available recorded mooring motion and clock drift), and as uncorrected (for mooring motion) travel times referenced to numerically calculated travel times for the basic state. The day differentials were used in the initial inversions presented by the Ocean Tomography Group (1982) since they are simple and robust, and could be quickly fed to inverse operators calculated before the 1981 moorings were recovered.

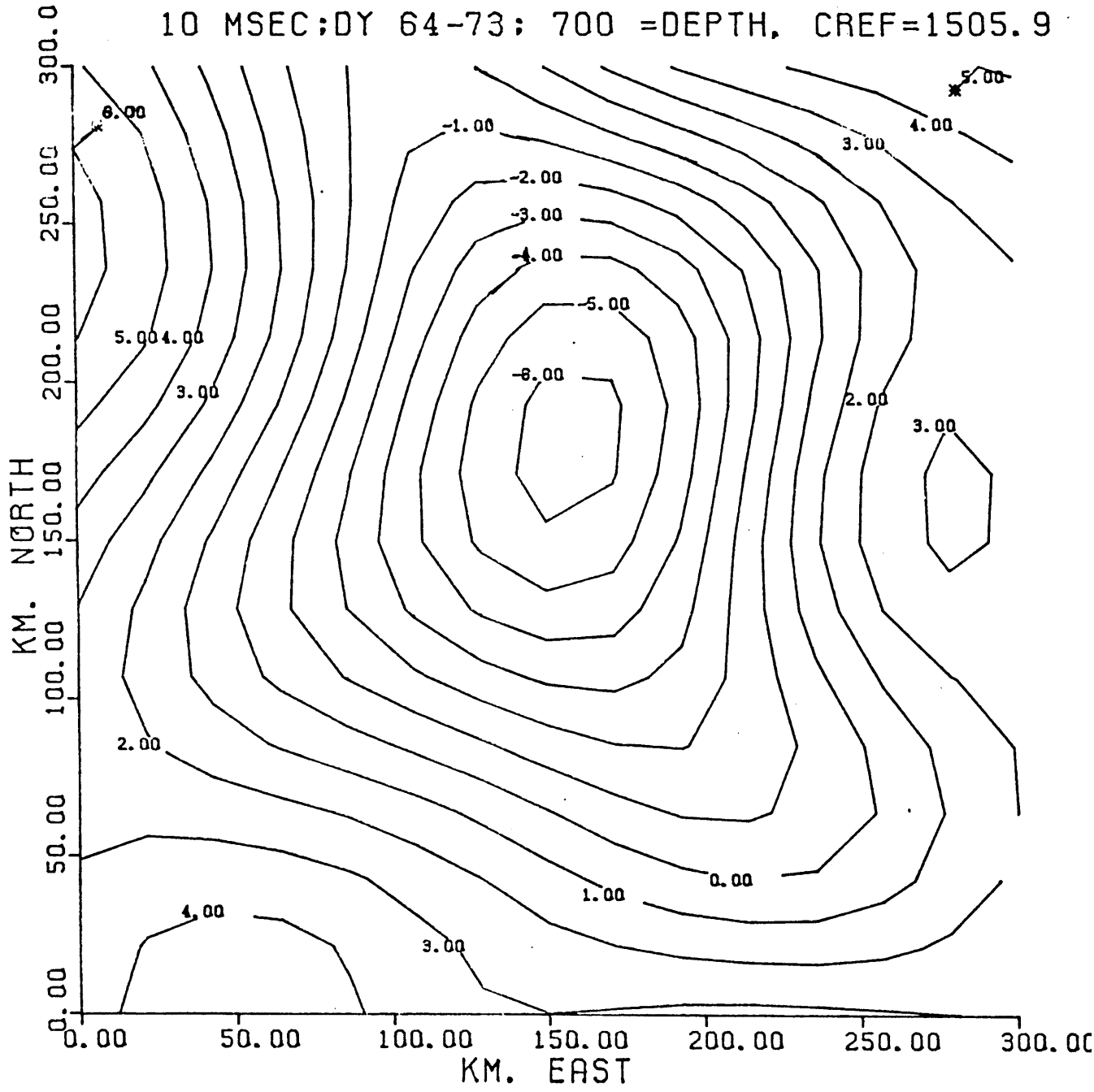
Because day differential travel times are referenced to the observed travel times on a given day of the experiment, they are not affected by the uncertainties of the mooring anchor positions. In fact, the true positions of the moorings do not need to be known, provided that the relative motions have been tracked and removed from the travel time data set. The model used to produce the expected data-data covariances for day differential travel times can thus be made very simple since the times depend only on mesoscale sound speed changes plus measurement errors. The original plan for the tomographic inversions was to use only these data, counting on the availability of mooring motion data to correct the travel times before invoking the inverse operator.

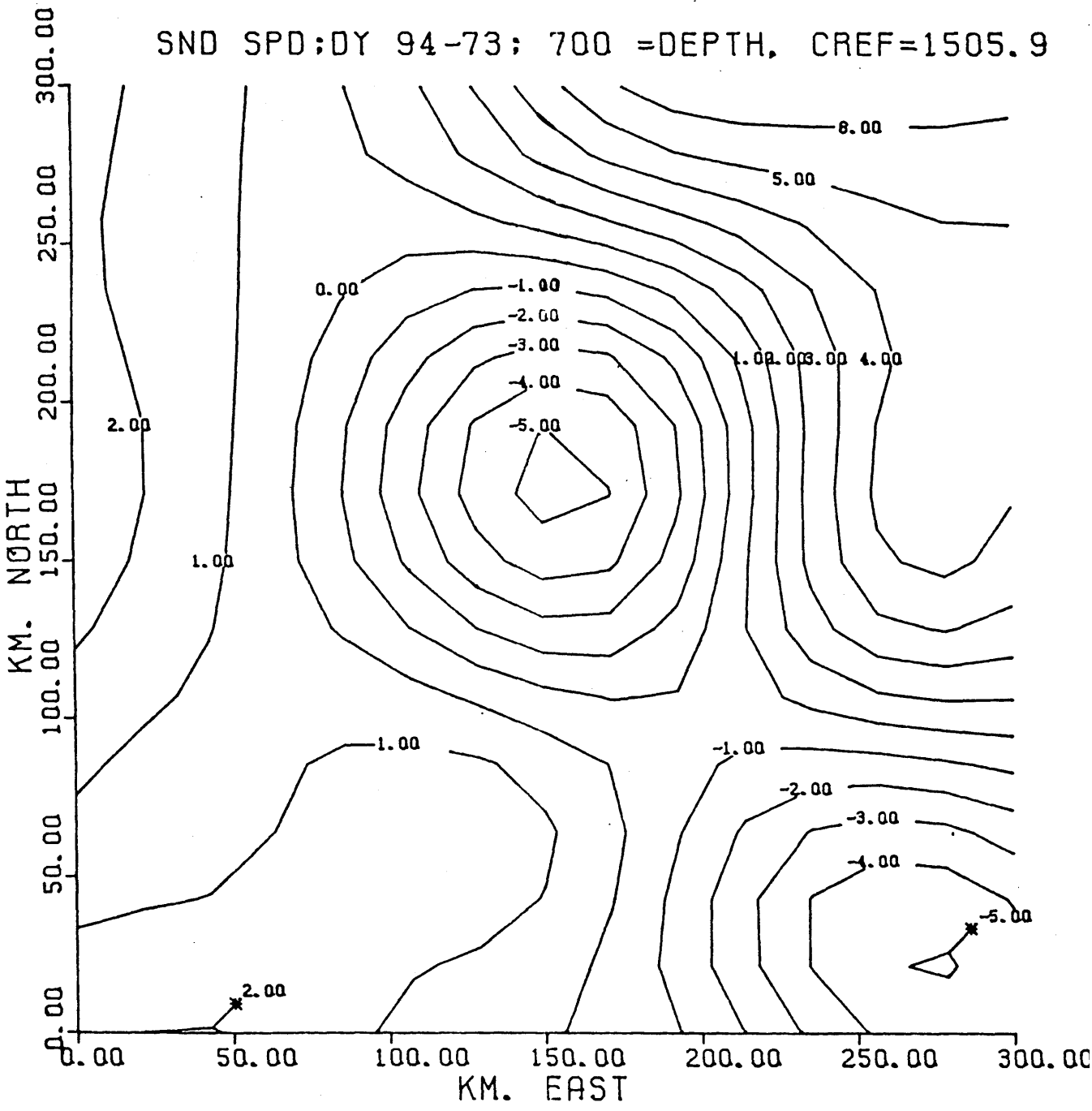
9.3 THE DAY DIFFERENTIAL ESTIMATOR

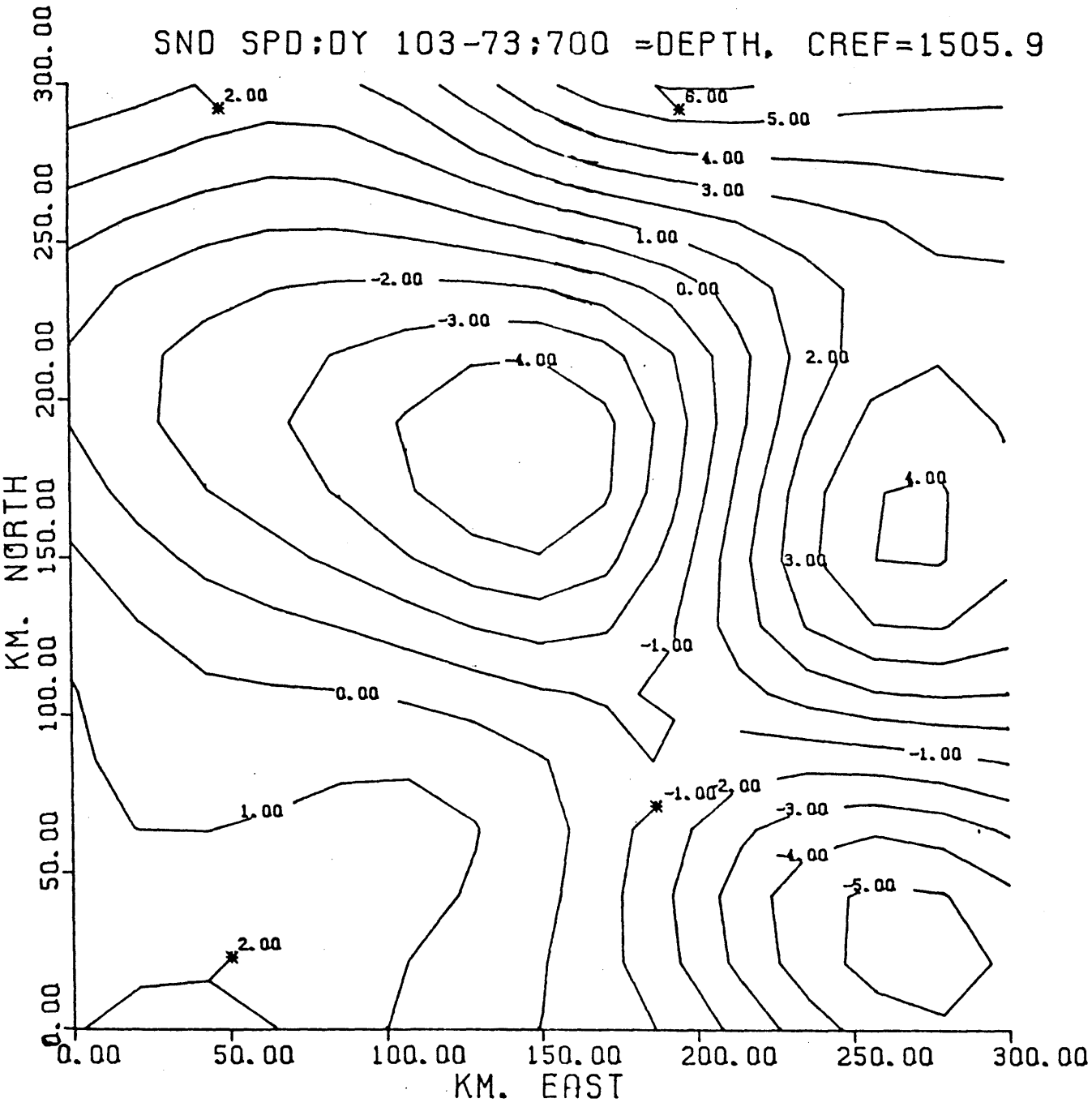
Day differentials are insensitive to errors in the ray identification, and to uniform clock offsets or other systematic errors in the data, so there is less worry in using a preliminary data set. On the minus side, because day differentials require mooring motion data, maps can only be made for the days when enough of the transponders were in operation to give a reliable set of corrections. There are random errors present on all days, but day differentials have twice the expected error variance of the original times. The day differentials produce maps of the sound speed anomalies relative to the reference day of the travel time differences. In the OTG paper, this was overcome by picking a reference day during the first NOAA CTD survey, so that the computed sound speed anomalies were added to the field calculated from the CTD survey to produce total maps. (Figure 9.4). The day differential travel times were used to calculate estimated sound speed mode amplitudes at the 65 CTD station locations. The mode amplitudes were used to linearly combine the vertical modes to produce an updated survey, which could be objectively mapped for plotting in the same way that the original stations had been.

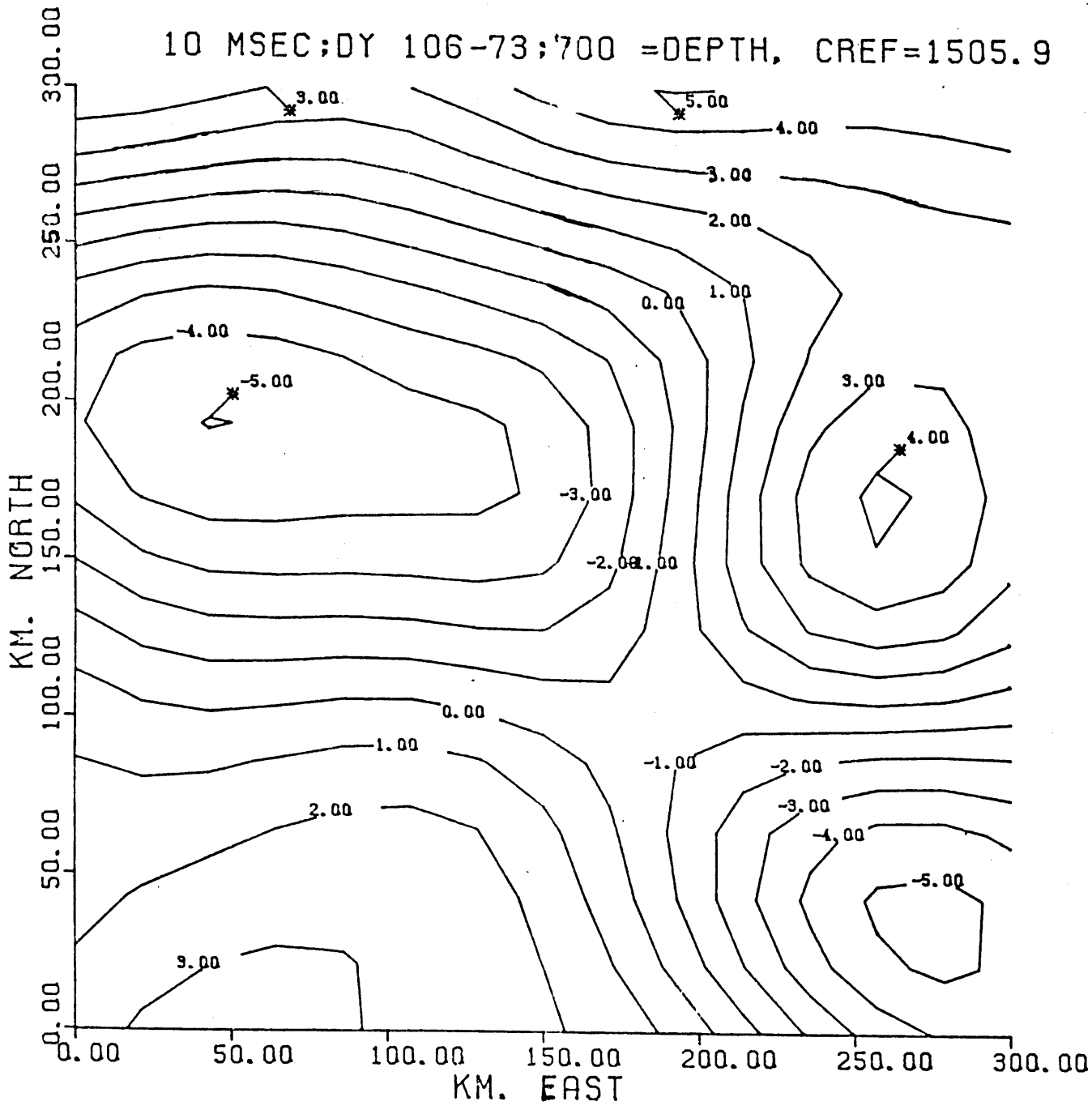
FIGURE 9.4 A,B,C,D: MAPS OF SOUND SPEED ANOMALY GENERATED USING DAY-DIFFERENTIAL TRAVEL TIMES REFERENCED TO DAY 73, DURING THE FIRST NOAA CTD SURVEY. CONTOURS ARE OF SOUND SPEED ANOMALY RELATIVE TO THE REFERENCE C(Z). CONTOUR INTERVAL IS 1 M/SEC.

FIGURE 9.4 A MAP FOR DAY 64




SND SPD;DY 94-73; 700 =DEPTH, CREF=1505.9
8.00
5.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
1.00
2.00
3.00
4.00
2.00
1.00
1.00
-1.00
-2.00
-3.00
-4.00
-5.00
2.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00


SND SPD;DY 103-73;700 =DEPTH, CREF=1505.9
2.00
6.00
5.00
4.00
3.00
1.00
0.00
-2.00
-3.00
-4.00
2.00
4.00
3.00
-1.00
0.00
-1.00
1.00
-1.00
2.00
-3.00
-4.00
-5.00
2.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


10 MSEC;DY 106-73;700 =DEPTH, CREF=1505.9
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00
3.00
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00

The same techniques can be used with any inhomogeneous basic state, and so iteration is simple. The initial estimate of the true sound speed field is mapped to produce a "continuous" ocean in which numerical rays are traced. The inverse is re-computed following the scheme in Chapter 6 and the data are adjusted to conform to the iteration scheme outlined there. Each inverse result is mapped to update the previous ocean estimate, so the cycle can be repeated endlessly, if desired. During the 1981 experiment the ocean perturbations were far too weak to deform the paths enough to require iteration (See Figure 2.5).

This was fortunate, because while the iterative procedure is simple, calculation of the travel time data covariance matrices can require significant computer time, since the double integration over two ray paths can require the computation of the covariance estimate upwards of 10^4 times per matrix element. This is not a problem on a large computer, but for a megameter array, with 500 to 1000 computed points per ray, 10^6 covariance computations per matrix element may raise issues of computational efficiency, forcing compromises in the generality of the inverse form.

9.4 DATA ERROR AND INFORMATION

The maps shown are made for those days on which enough corrected data were available to give adequate resolution. If too few rays are used, then the inverse maps do not have much detail. On the other hand, adding rays to the inverse beyond a certain point will not greatly increase the resolving power of the estimator, because no additional independent information is being added. This break-even point is dependent on the amount of random error in the measured travel times. If the random error is large, then similar rays may be indistinguishable within the limits imposed by the error, so that a supplemental ray is less "valuable" than if the error level was smaller (Figure (9.5)).

Figure 9.5 A is a plot of information content vs. the number of rays used in the inverse. The slope of each curve is the marginal gain in information per additional ray datum, given a particular level of random error and no expected mooring offsets. The dotted curve represents an ideal case where there are absolutely no errors in the data, so that each additional ray datum adds independent information. In a real case, with finite errors, the curves deviate from this ideal line when the newest ray added to the inverse samples the ocean very much like some

FIGURE 9.5 A INFORMATION CONTENT OF DATA SET AS A FUNCTION OF THE RANDOM ERROR LEVEL AND THE NUMBER OF RAYS USED. NOTE THAT ADDING RAYS TO THE INVERSION HAS LITTLE EFFECT IF ERRORS ARE STRONG.

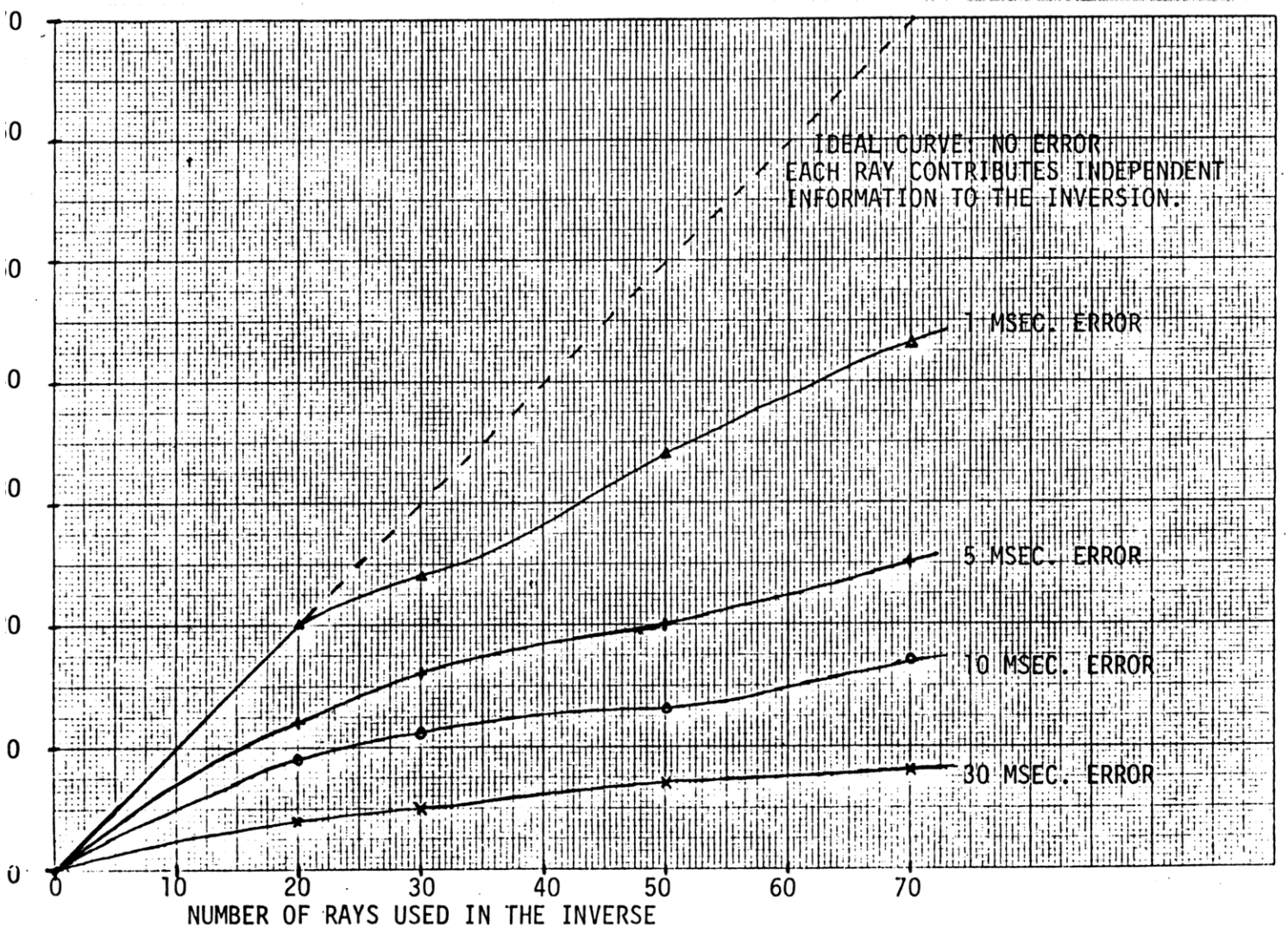
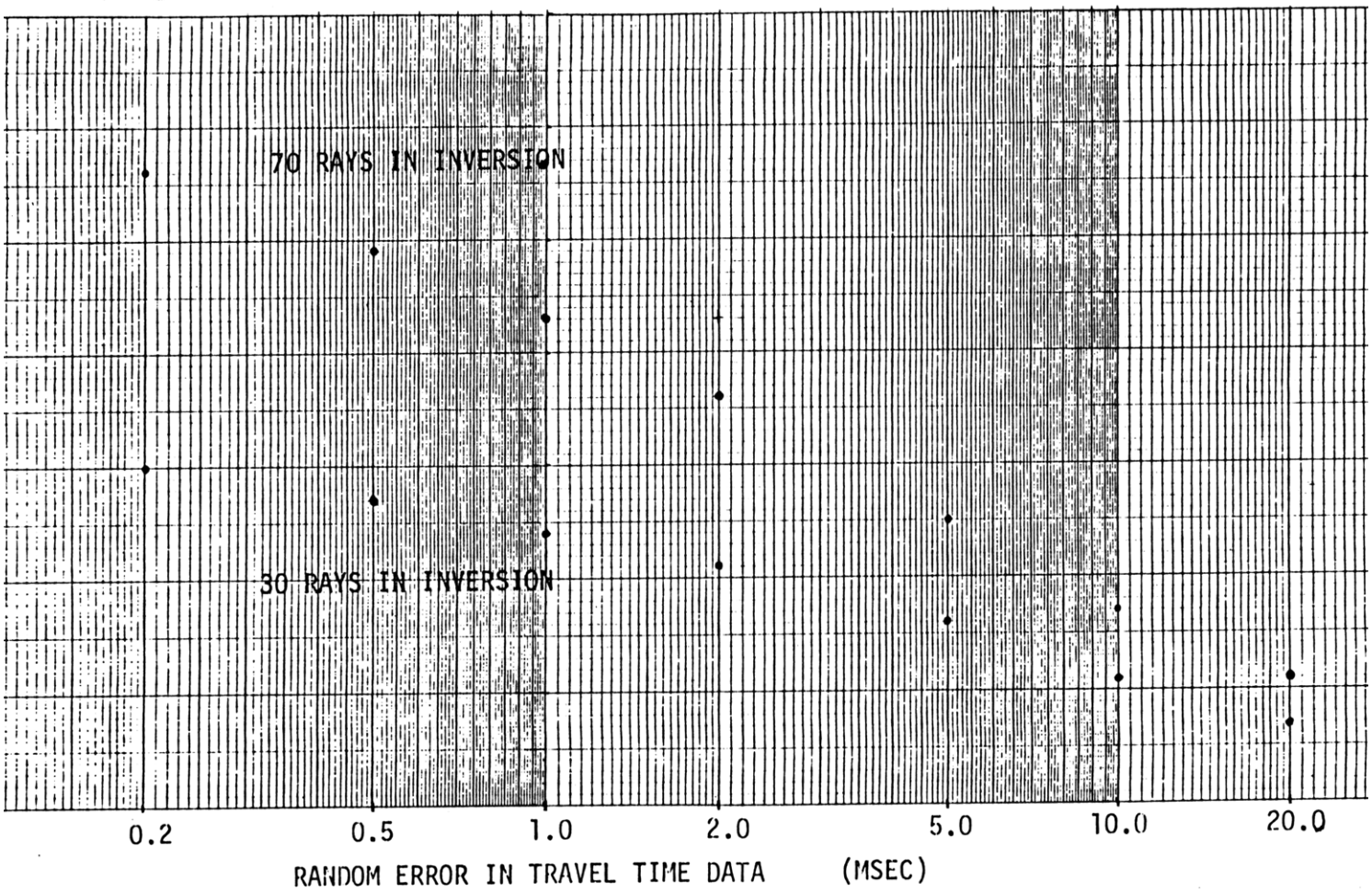


FIGURE 9.5 B INFORMATION CONTENT VS ERROR LEVEL FOR INVERSIONS WITH 30 AND 70 RAYS IN THE DATA SET. SHOWS HOW LOGARITHMIC DECREASES IN ERROR ARE REQUIRED TO GIVE CONSTANT INCREASES IN INFORMATION.



combination of rays already included, with the differences swamped by the random errors. The curves in Figure 9.5 A are not smooth because the rays are added haphazardly, so that several rays from a given source-receiver pair may be added at once. In all cases, the slopes decrease for large numbers of rays, showing the lessening benefit from added data at a constant error level. This type of curve can be used to analyse the amount of range information available in the rays of a single source-receiver pair.

Figure 9.5 B shows the decrease in independent information available to the estimator as the random error in the data is increased. At the low error extremes, the curves end at the number of rays used, while for large errors they tend toward zero. Figures 9.5 A and B can be used to bound the performance of the inverse as the number of rays used is increased beyond the 73 used for the maps in this thesis. If the random errors in the data cannot be reduced below 5 msec, no dramatic improvements in the results can be expected, while if an error level of 1 msec can be attained, the maps shown herein should improve significantly.

The use of figures of this type during array design simplifies the tasks of choosing engineering parameters and estimating the eventual performance bounds on the system.

Note that logarithmic decreases in the error level are required to maintain constant increases in the amount of independent information. Adding dependent rays increases the error immunity of the inverses somewhat, but does not produce the same improvements in resolution that independent rays yield. For the preliminary maps, about 73 rays were used, less than half of the number seen as stable arrivals at the receivers.

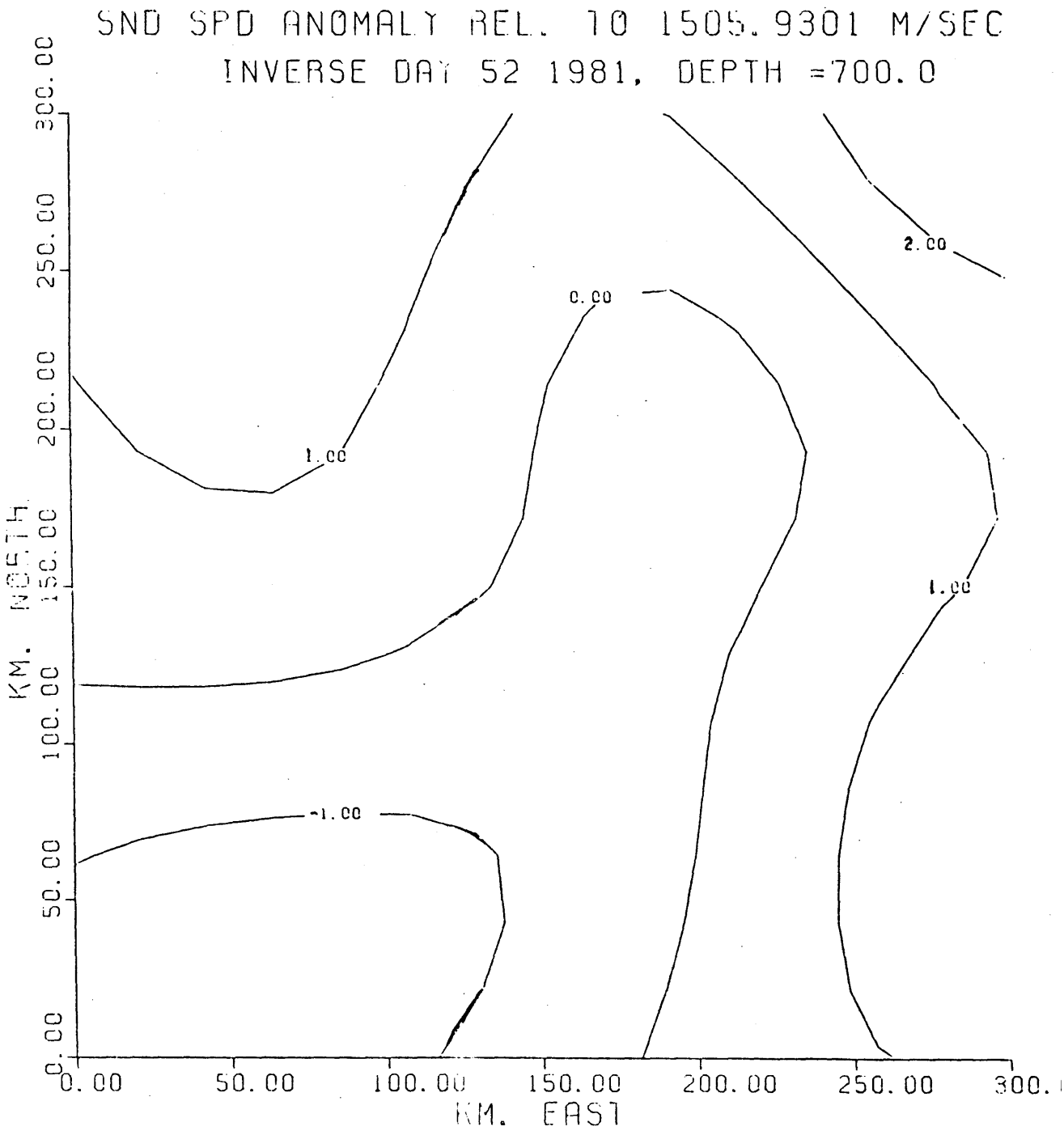
For the OTG paper, some uncorrected data were included as ray differentials (see chapter 7), referencing all the rays in the arrival pattern for a source-receiver pair to one of the rays in the pattern. The subtraction doubles the noise variance, so a travel time constructed as both day and ray differential has about 4 times the expected error variance as a single travel time. The process of forming ray differentials reduces the expected level of mesoscale-induced travel time changes, from order 40 msec to order 5 msec, so that the signal to noise ratio for ray differentials is less favorable. About 30% of the data used in the OTG maps were these "day, ray differentials", and these had very little effect on the maps.

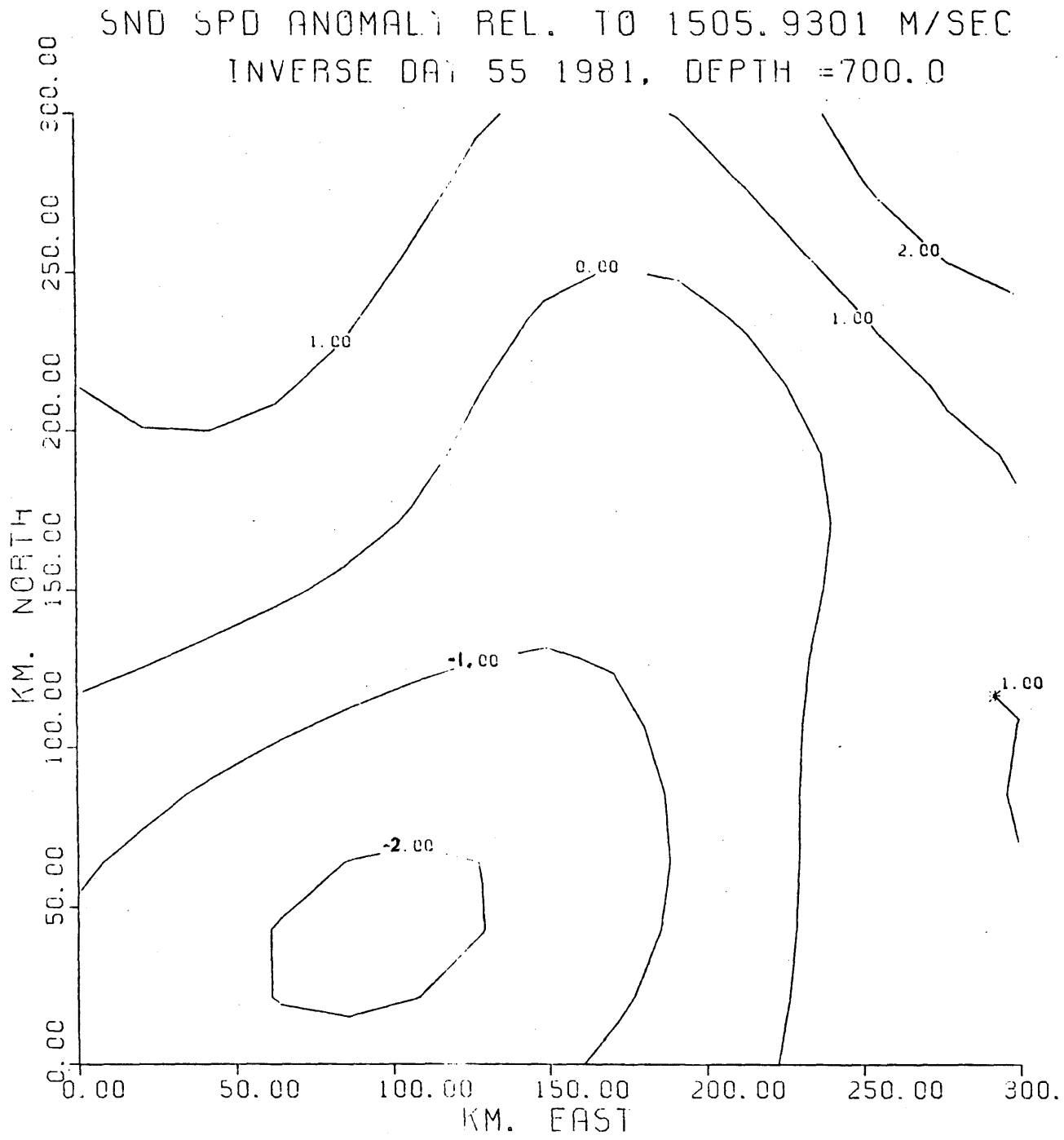
9.5 THE ESTIMATOR FOR UNCORRECTED DATA

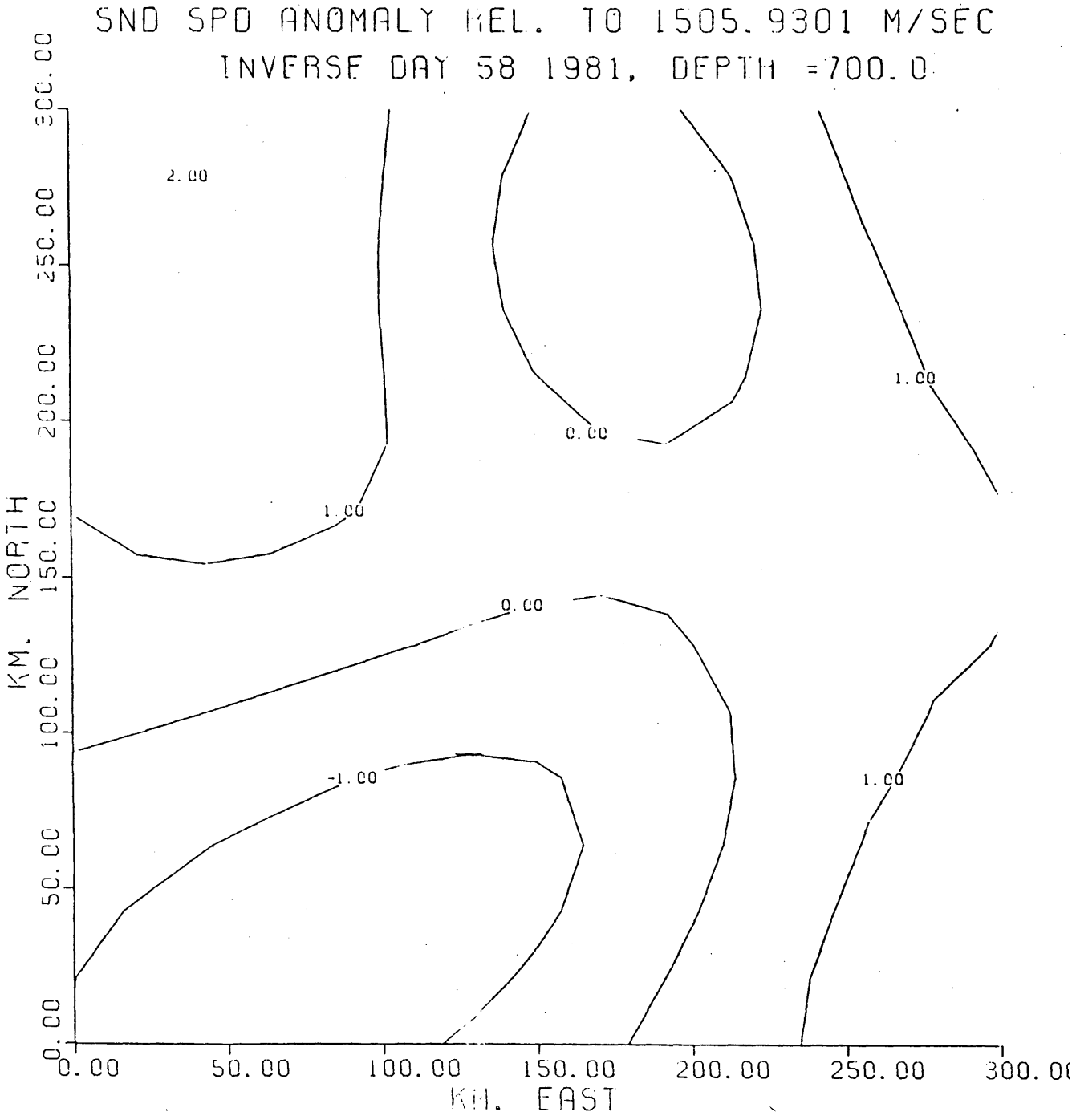
In order to more fully use the data set, it was necessary to abandon the simple day differential framework, and deal with the uncertainties in anchor position and mooring motion directly, as described in chapter 7. Parameterization of the mooring offsets is useful even if full mooring positions are available. The initial data corrections were done before the ray pattern was separated into arrivals and identified, so the entire pattern was shifted uniformly. The horizontal mooring motions were converted to line-of-sight range changes and divided by an estimated local sound speed to obtain an approximate travel time, which was then used to shift the time base of the arrivals. The true travel time effects of mooring position change depend on ray angle and, more critically, on depth changes, so that quasi-random errors are generated in this correction process. The errors introduced in this way can easily be order 5 msecs. The initial corrections must therefore be removed once ray geometry is known.

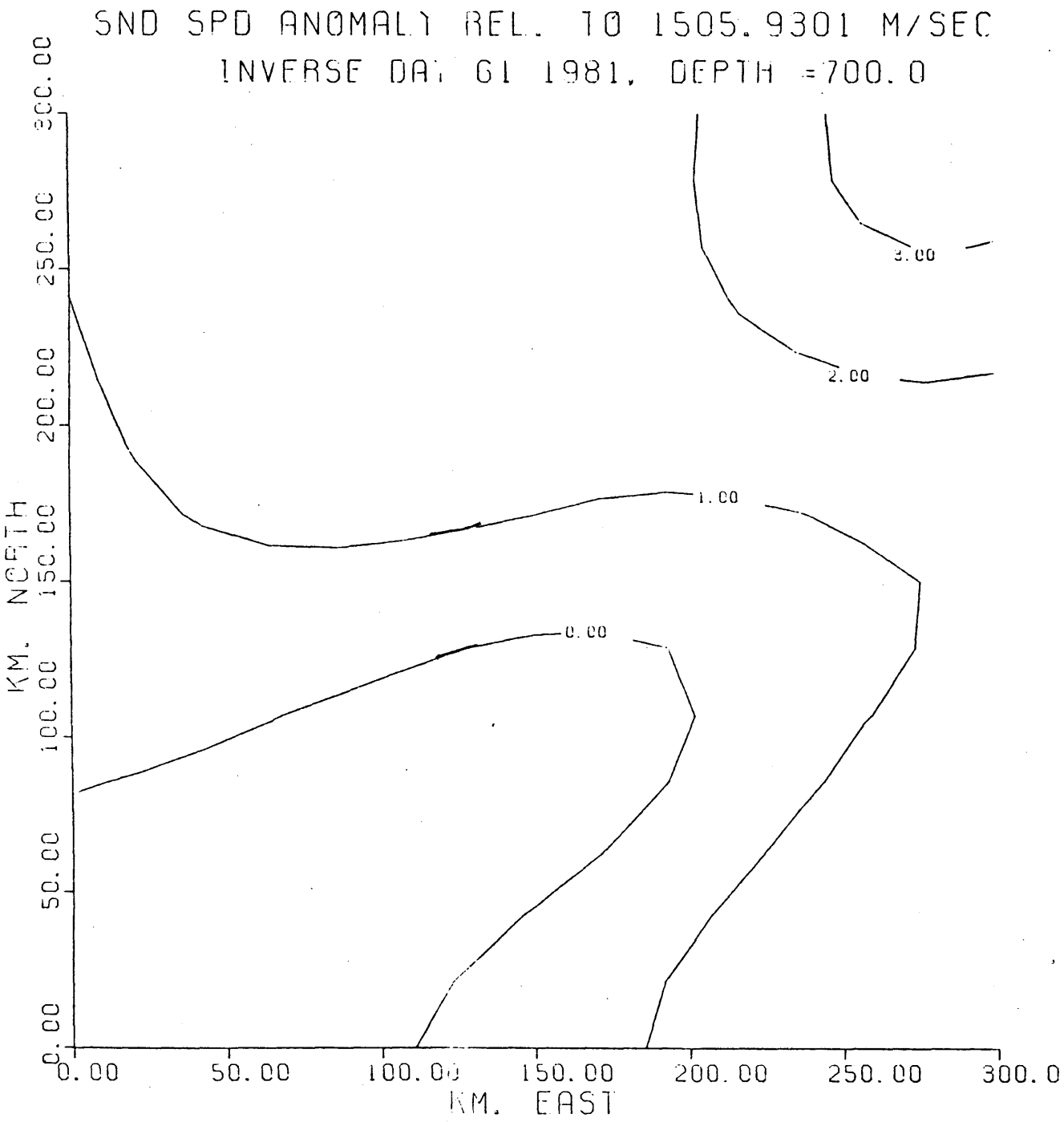
The maps shown as Figures 9.6 (A-DD) were made using data with the initial mooring motion corrections removed, and the inverse estimated mooring position in addition to constructing sound speed maps. Clock errors were also

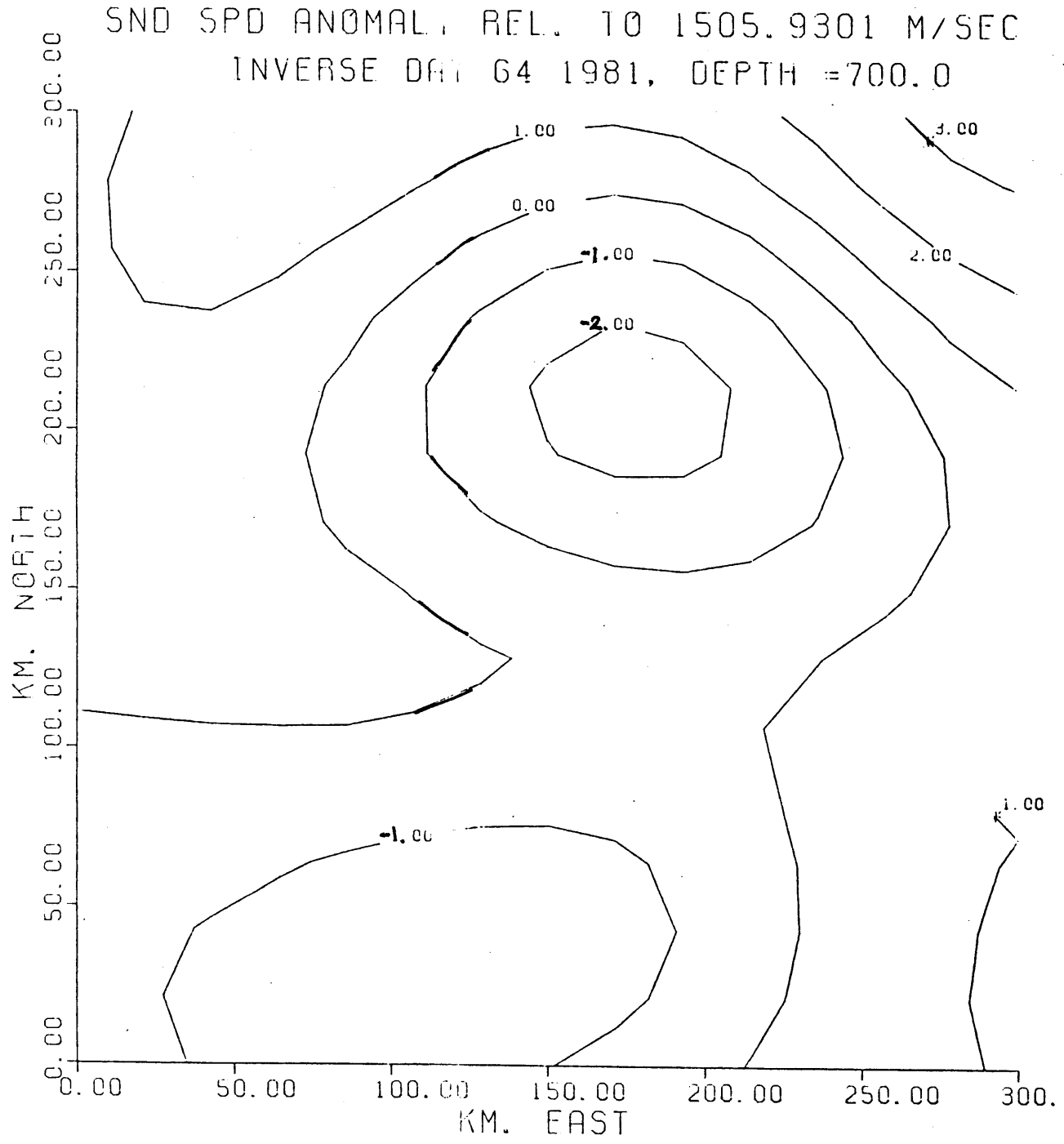
FIGURE 9.6 A-Z,AA-DD: MAPS OF SOUND SPEED ANOMALY AT 700 METERS DEPTH REFERENCED TO THE AVERAGE C(Z) PROFILE. CALCULATED FROM UNCORRECTED DATA, WITHOUT USE OF THE NOAA CTD SURVEYS. MAPS ARE PLOTTED FOR EVERY THIRD DAY. C.I. = 1 M/SEC.

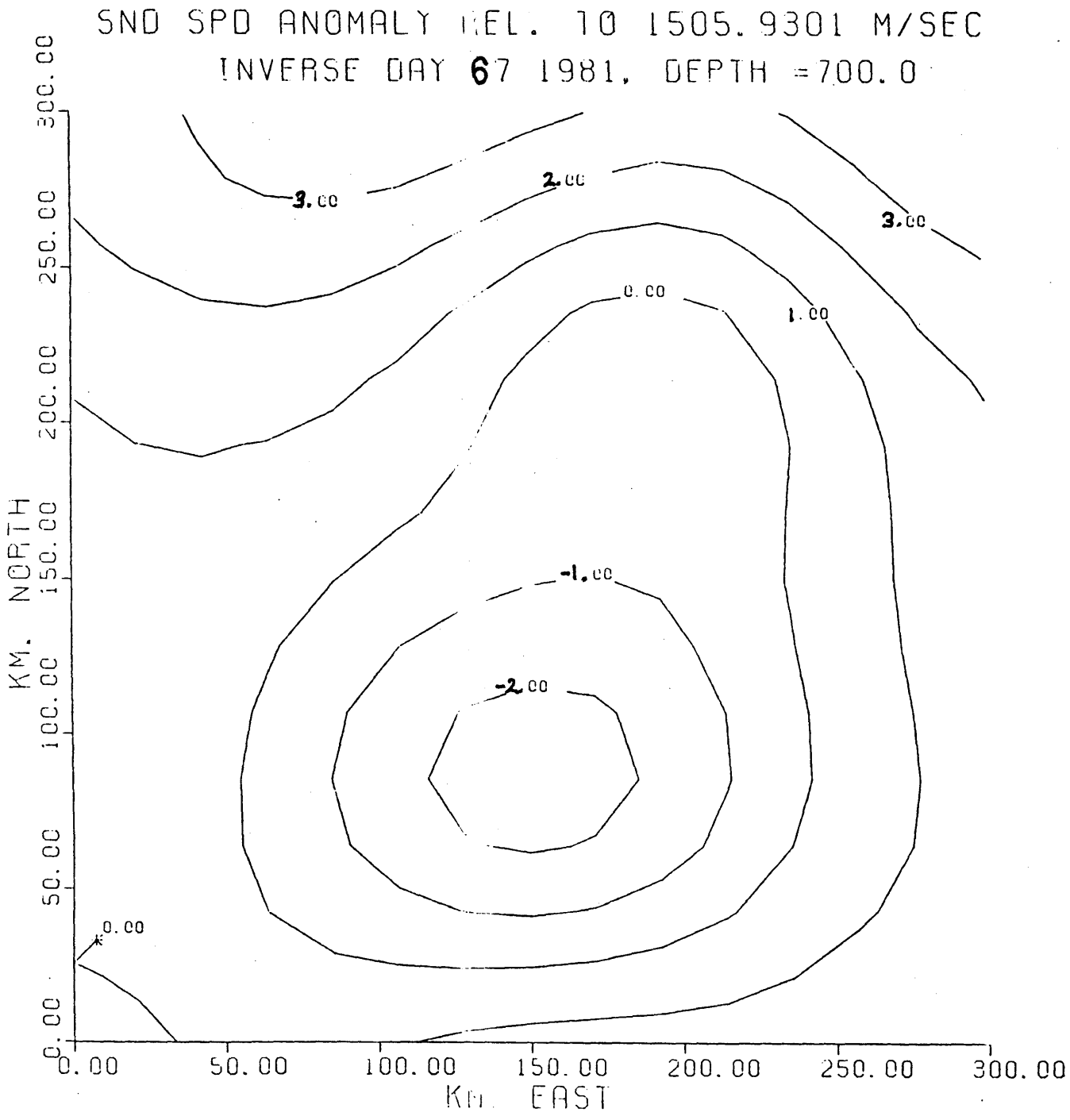


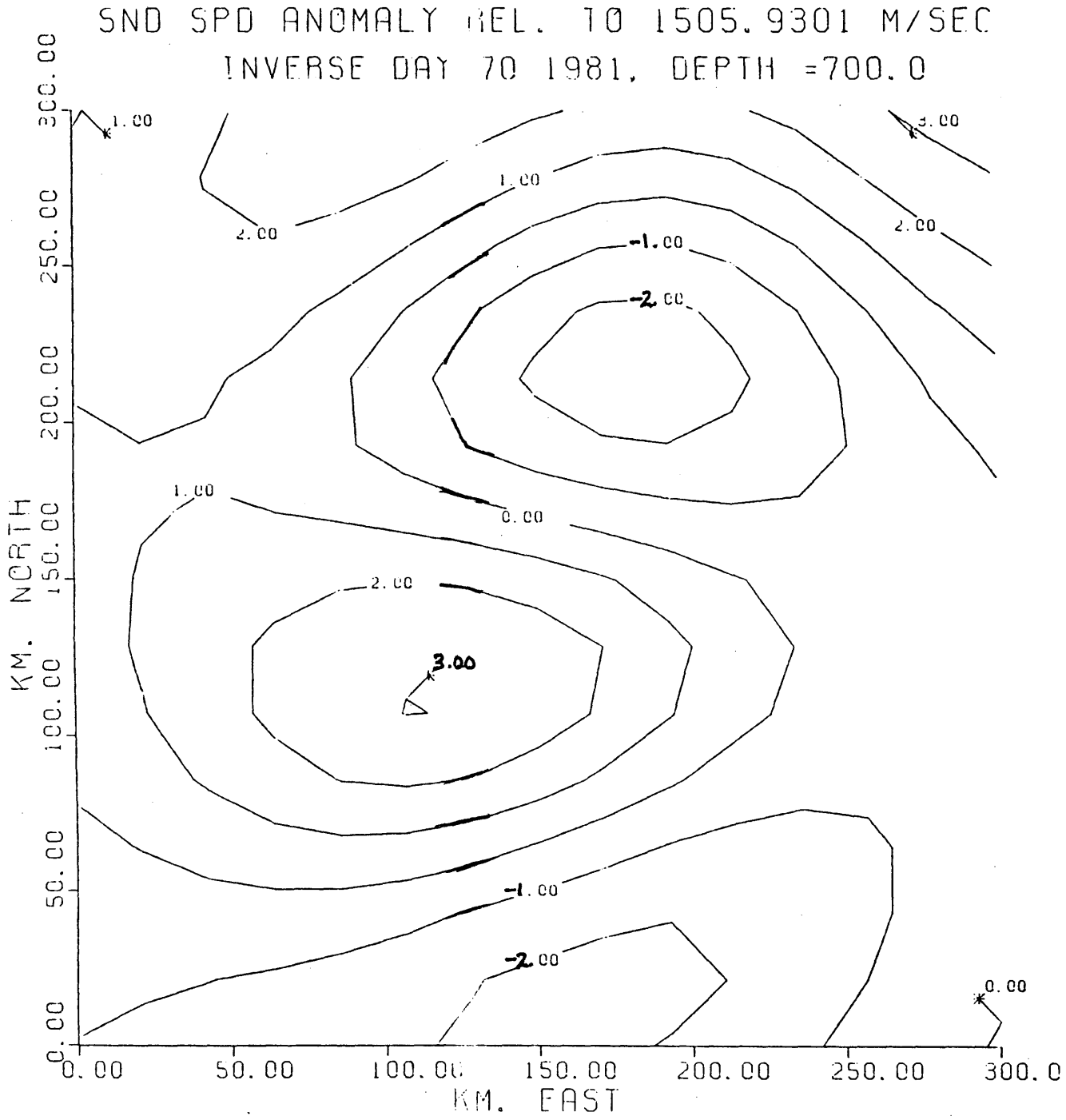

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 55 1981, DEPTH =700.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.
2.00
1.00
0.00
-1.00
-2.00

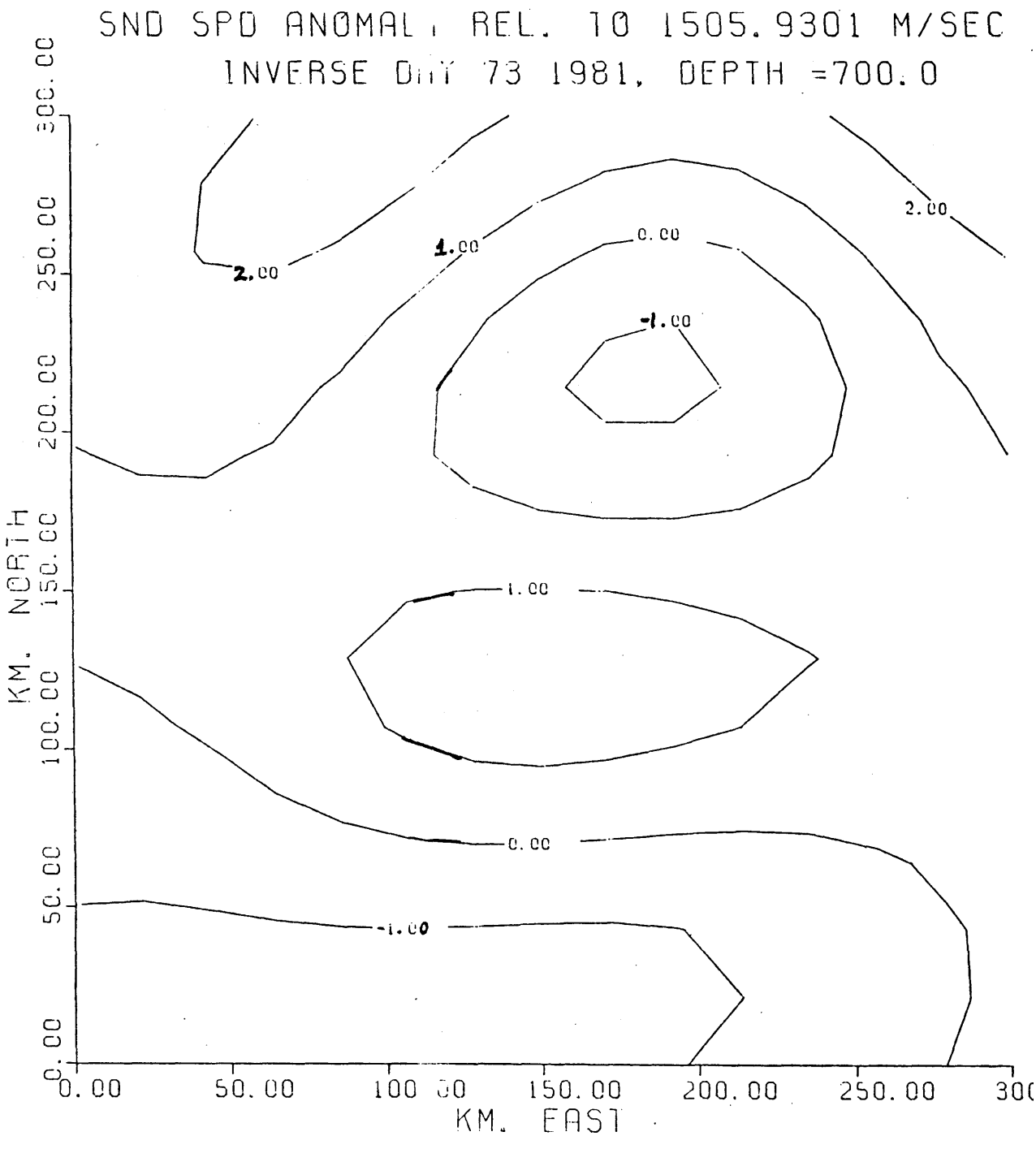

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 58 1981, DEPTH =700.0
2.00
1.00
0.00
1.00
0.00
-1.00
1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

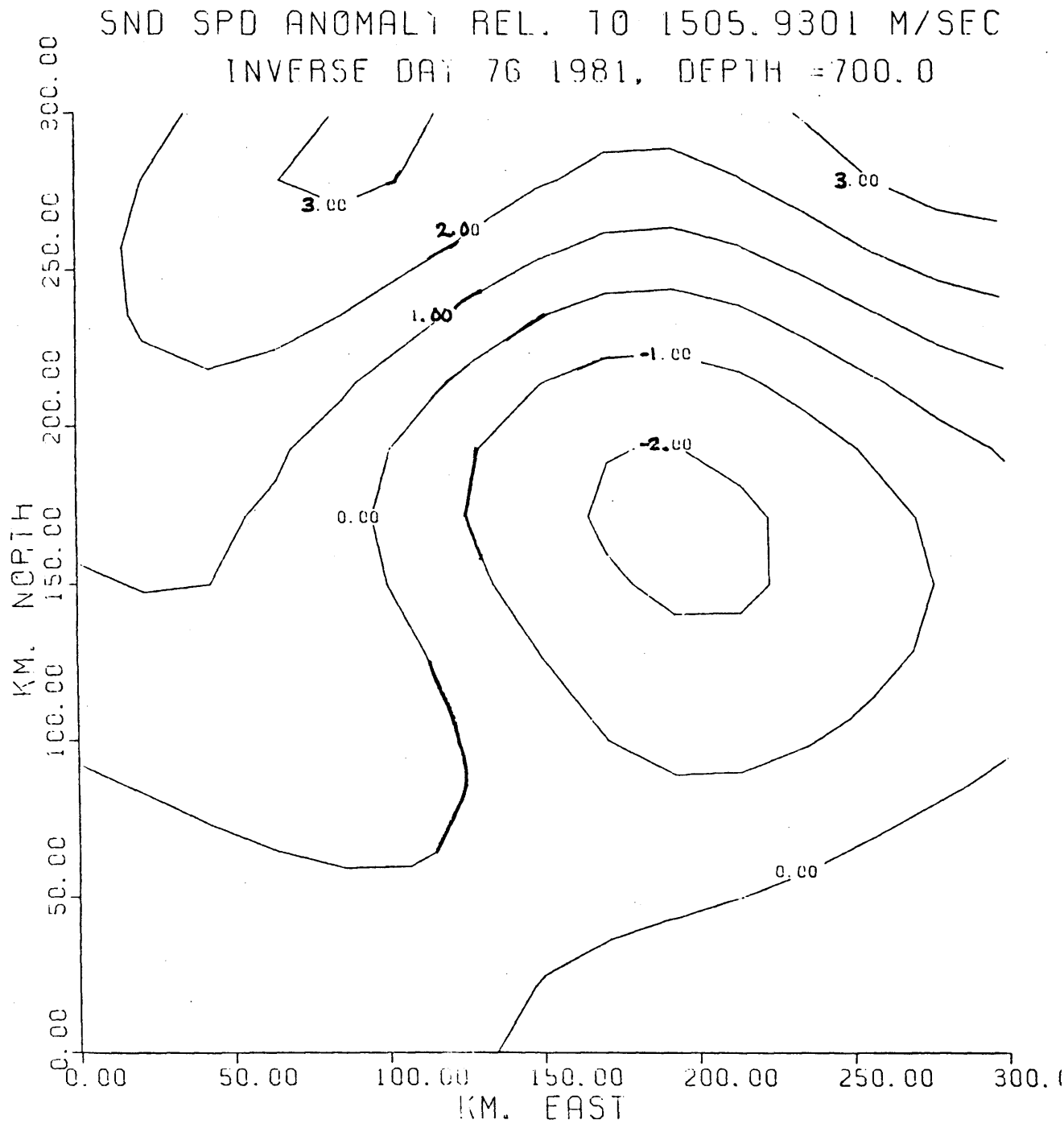

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 61 1981, DEPTH =700.0
3.00
2.00
1.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST

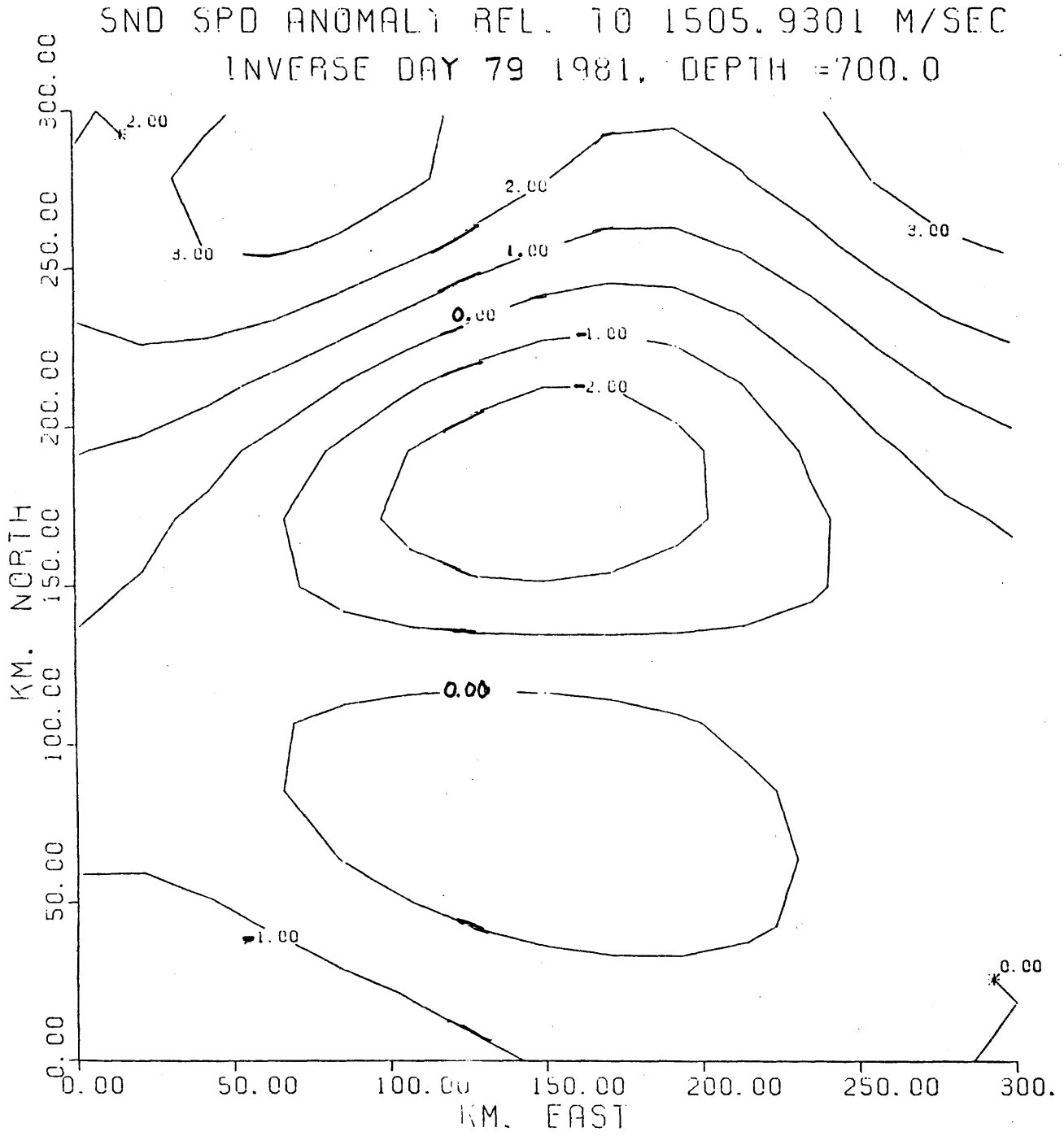

SND SPD ANOMAL, REL. TO 1505.9301 M/SEC
INVERSE DAY 64 1981, DEPTH =700.0
1.00
0.00
-1.00
-2.00
3.00
2.00
1.00
-1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.
KM. EAST

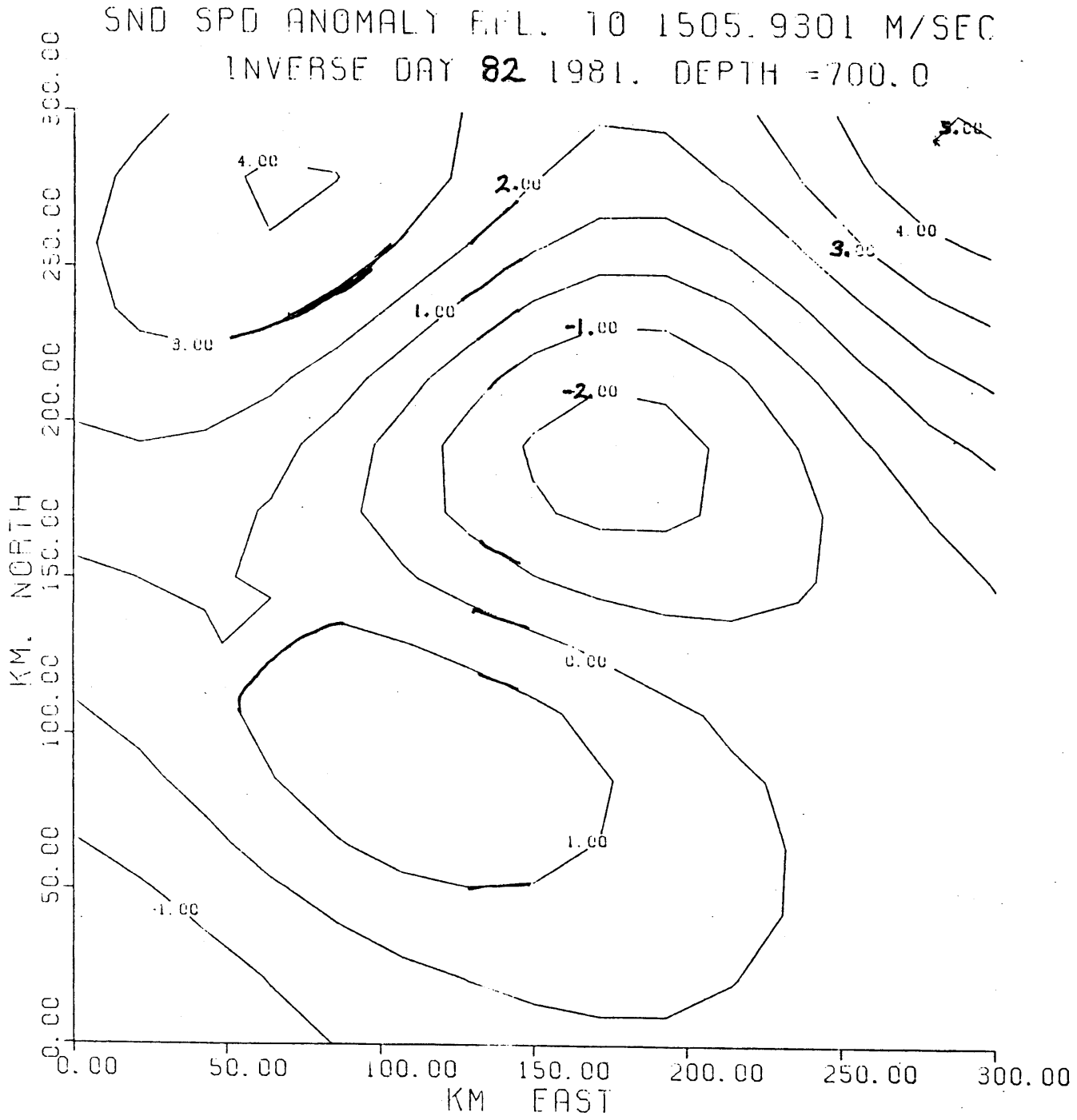

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 67 1981. DEPTH =700.0
3.00
2.00
3.00
0.00
1.00
-1.00
-2.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
Km. EAST

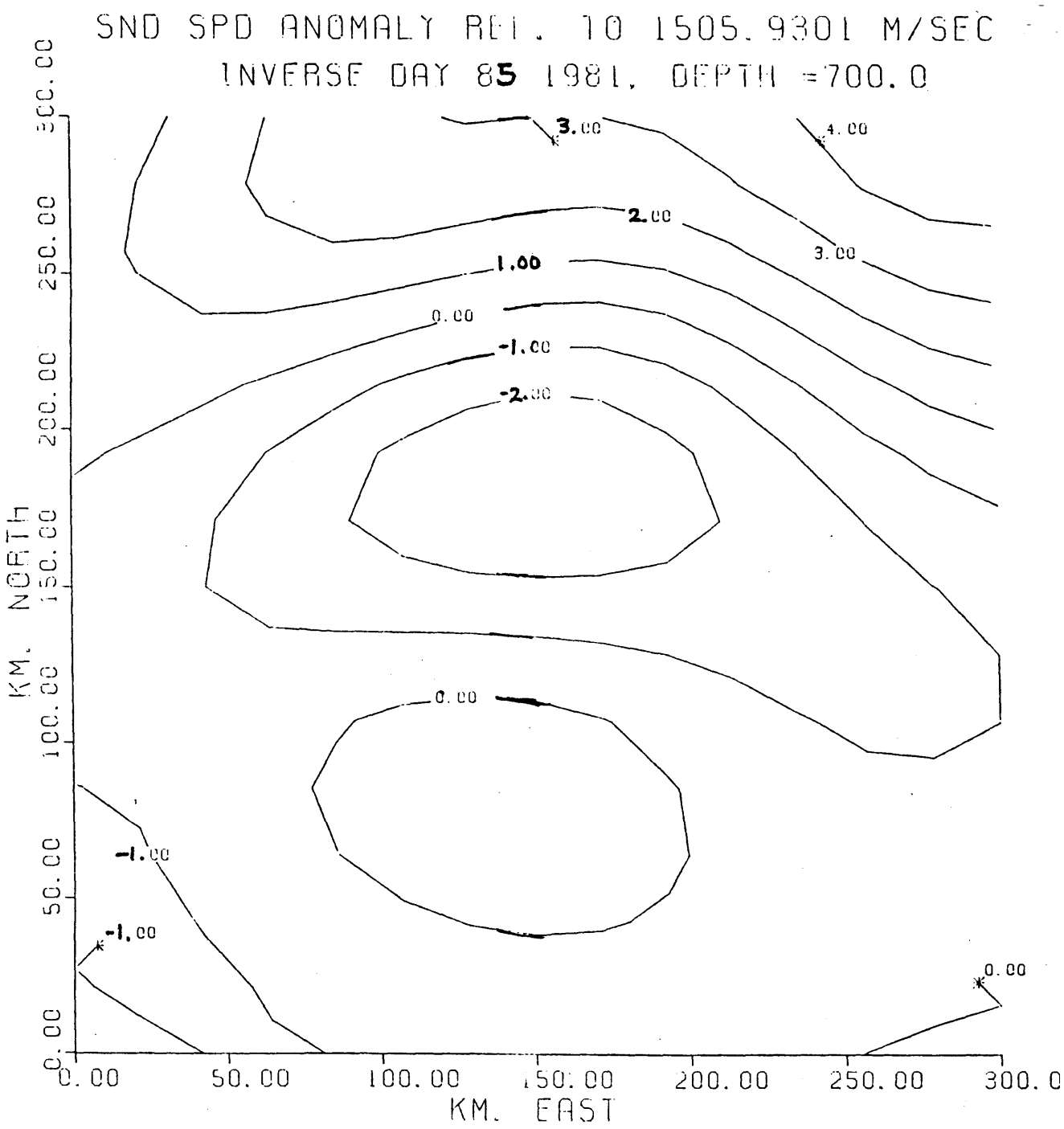

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 70 1981, DEPTH =700.0
1.00
2.00
1.00
-1.00
-2.00
2.00
1.00
0.00
2.00
3.00
-1.00
-2.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST

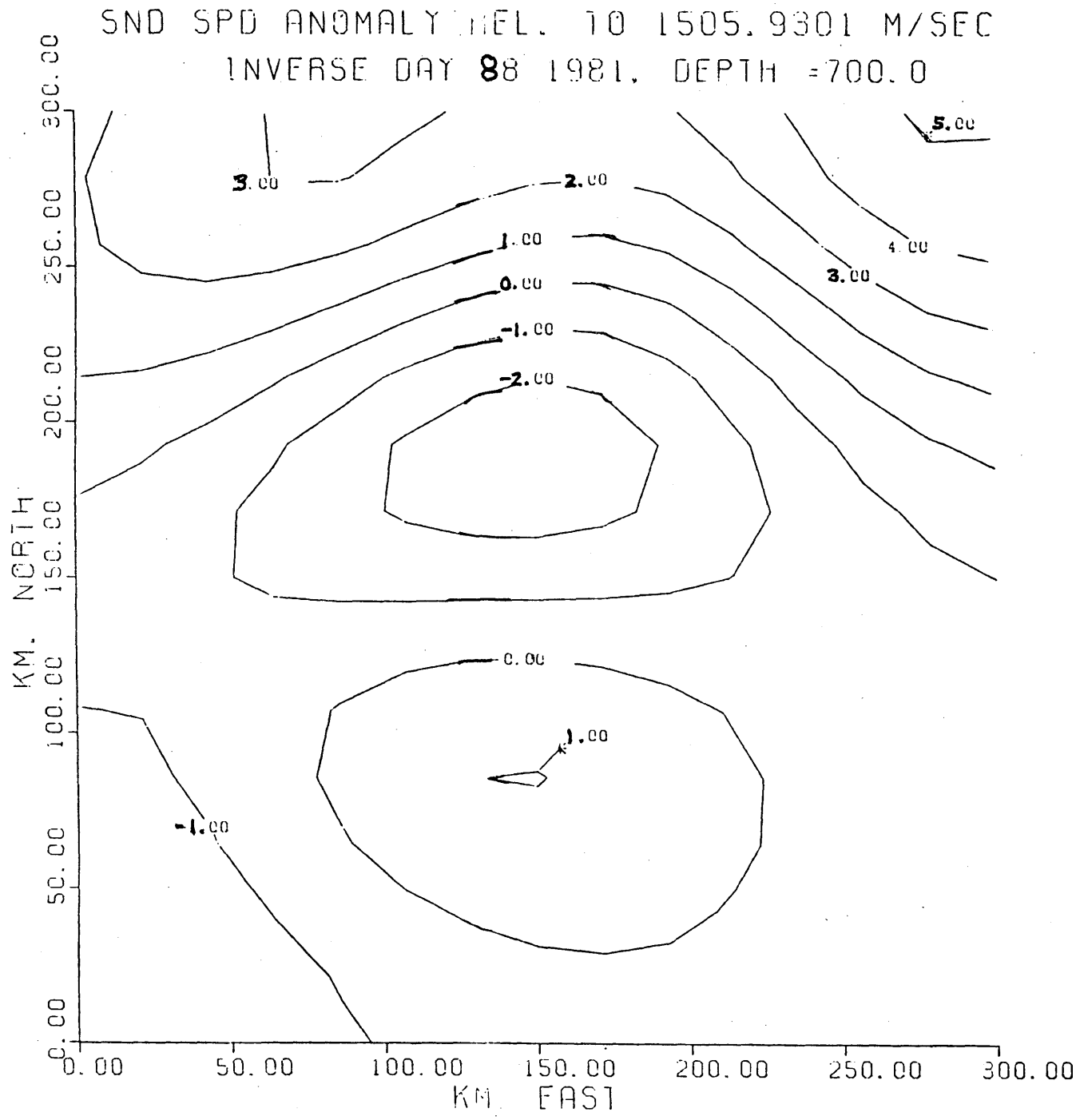

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 73 1981, DEPTH =700.0
2.00
1.00
0.00
-1.00
1.00
0.00
-1.00
2.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

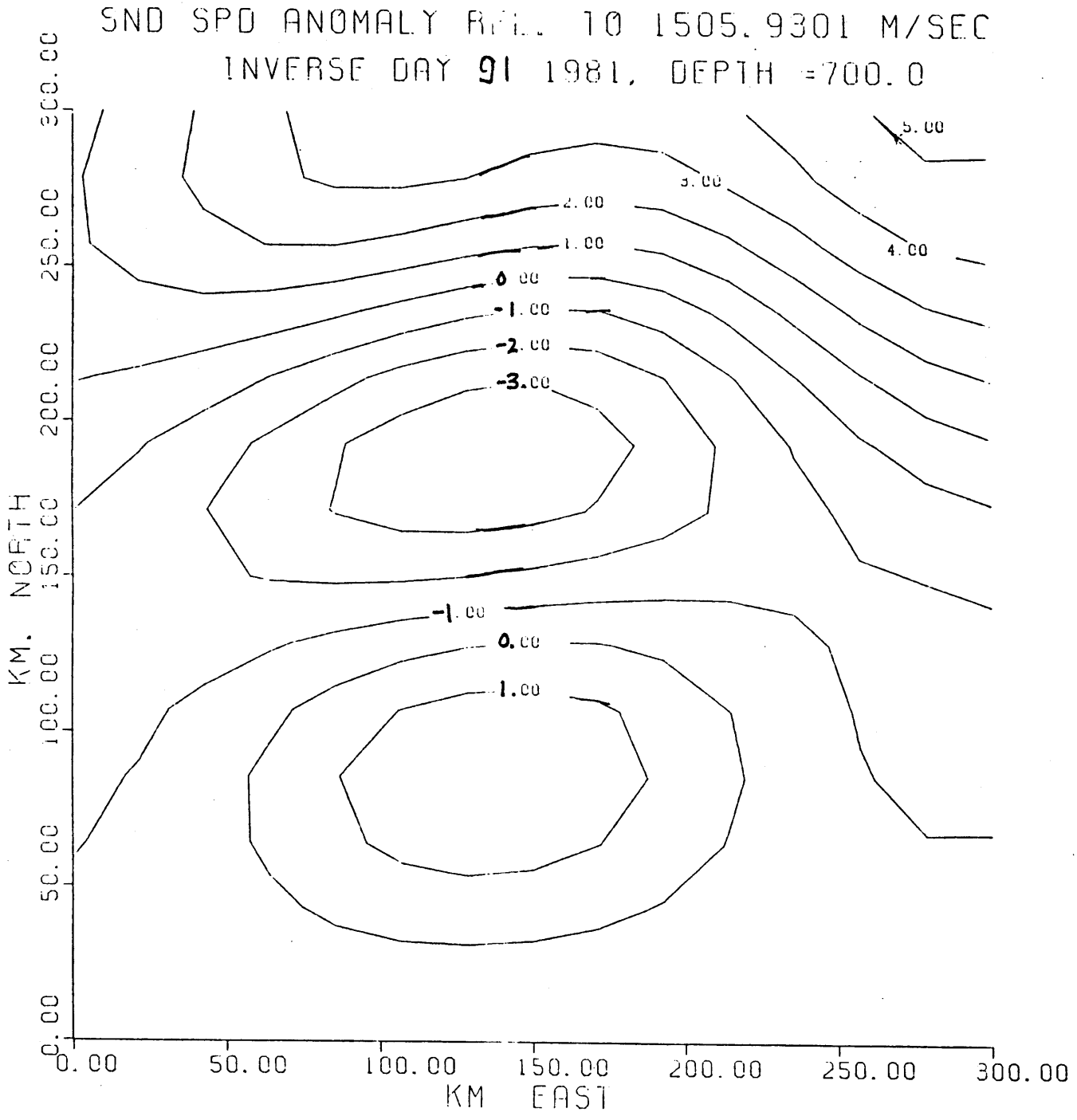

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 76 1981, DEPTH =700.0
3.00
3.00
2.00
1.00
-1.00
-2.00
0.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST

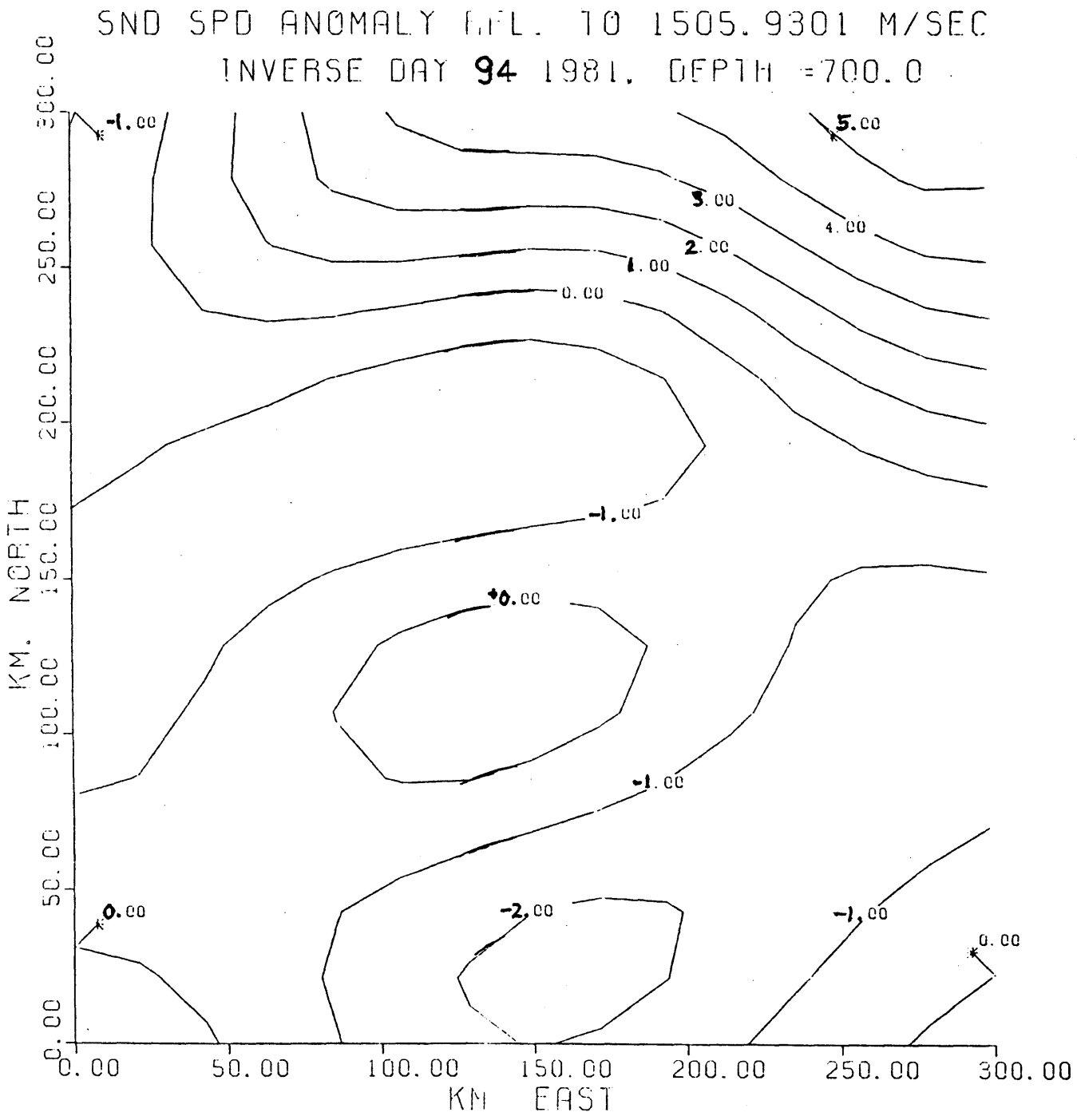

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 79 1981. DEPTH =700.0
2.00
3.00
2.00
1.00
3.00
0.00
-1.00
-2.00
0.00
-1.00
0.00
KM. NORTH
300.00
250.00
200.00
150.00
100.00
50.00
0.00
0.00
50.00
100.00
150.00
200.00
250.00
300.
KM. EAST

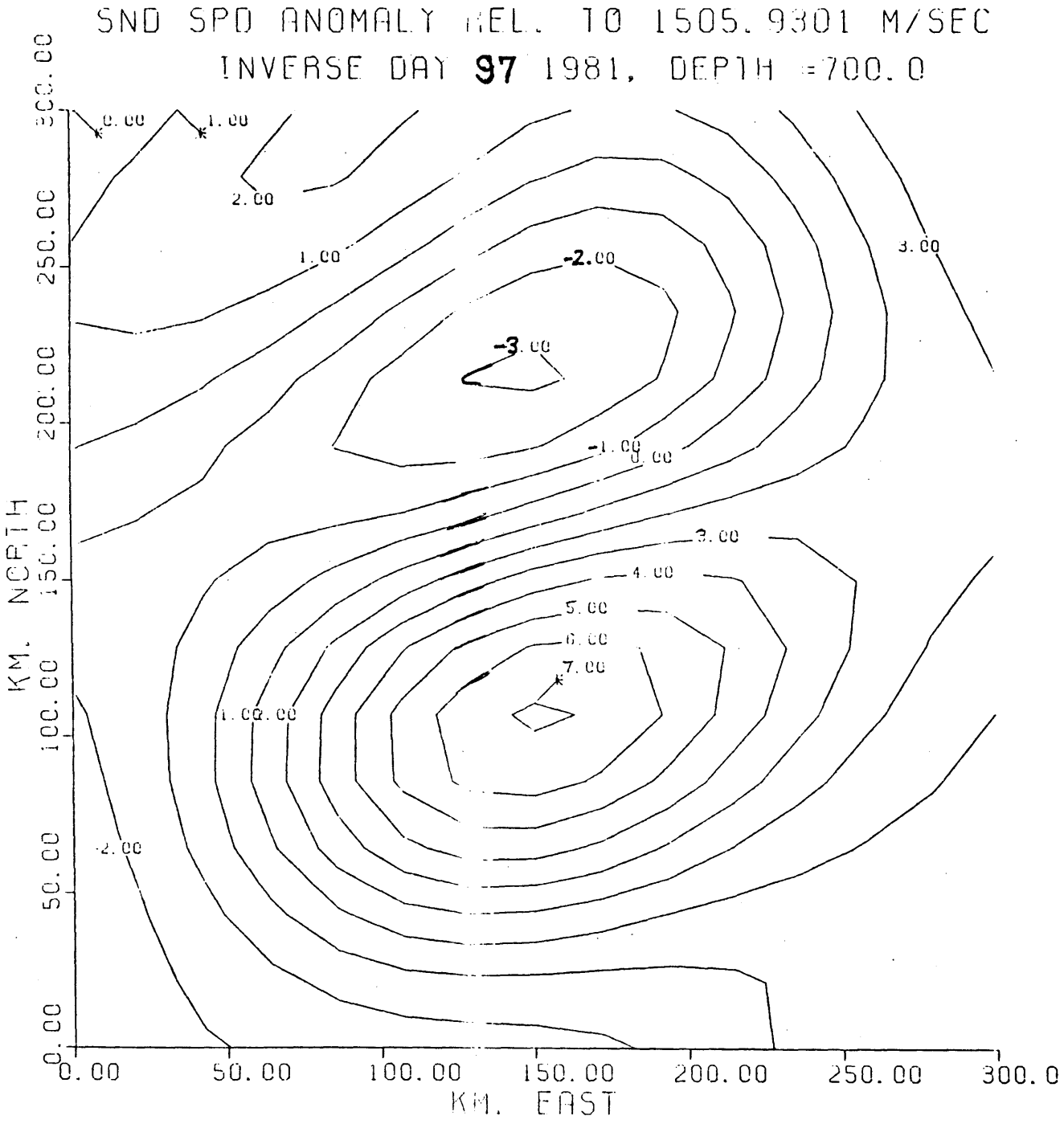

SND SPD ANOMALY R.P.L. TO 1505.9301 M/SEC
INVERSE DAY 82 1981. DEPTH =700.0
5.00
4.00
2.00
3.00
4.00
3.00
1.00
-1.00
-2.00
0.00
1.00
-1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM EAST

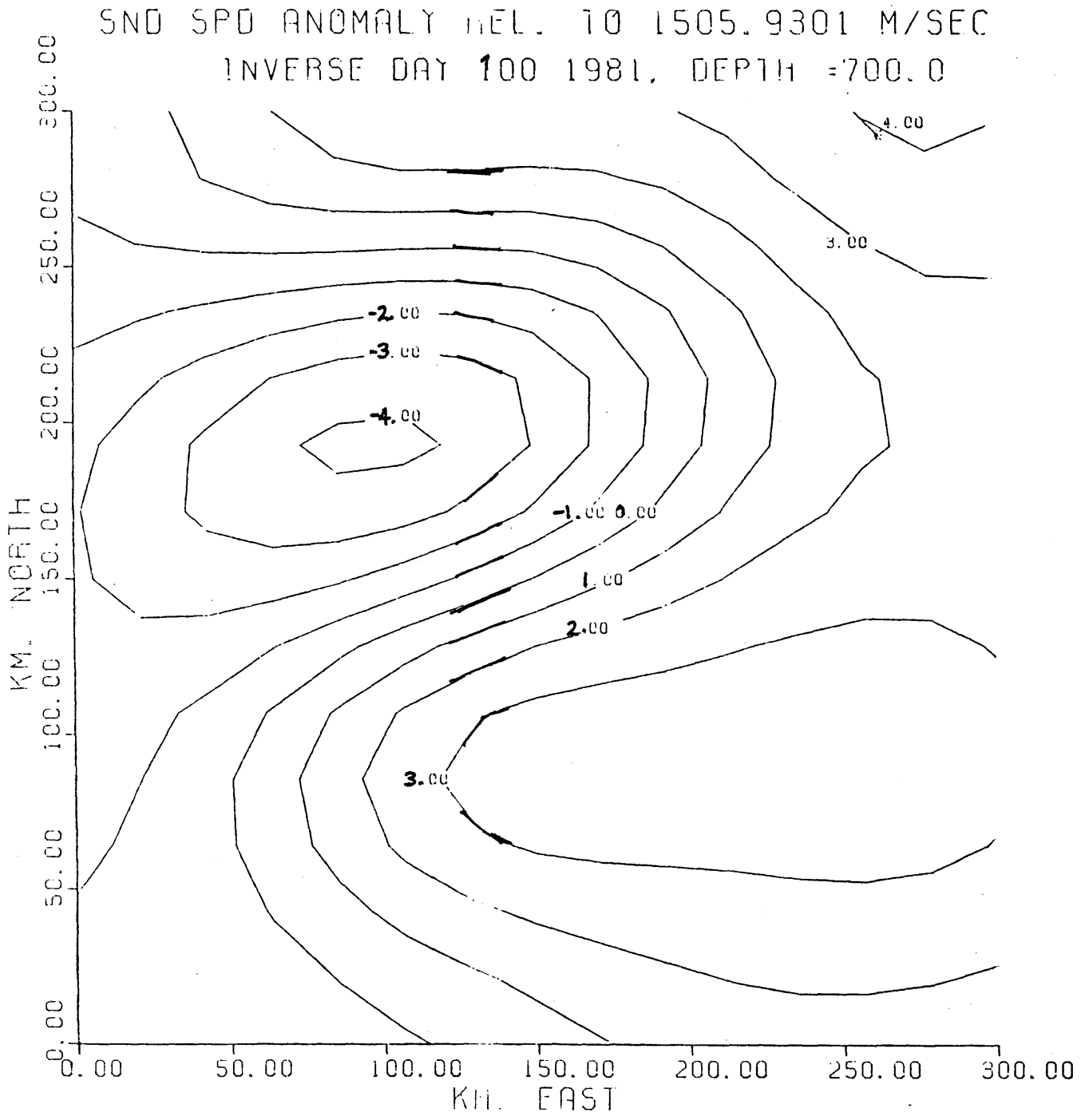

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 85 1981. DEPTH =700.0
3.00
4.00
2.00
3.00
1.00
0.00
-1.00
-2.00
0.00
-1.00
-1.00
0.00
300.00
250.00
200.00
150.00
100.00
50.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST

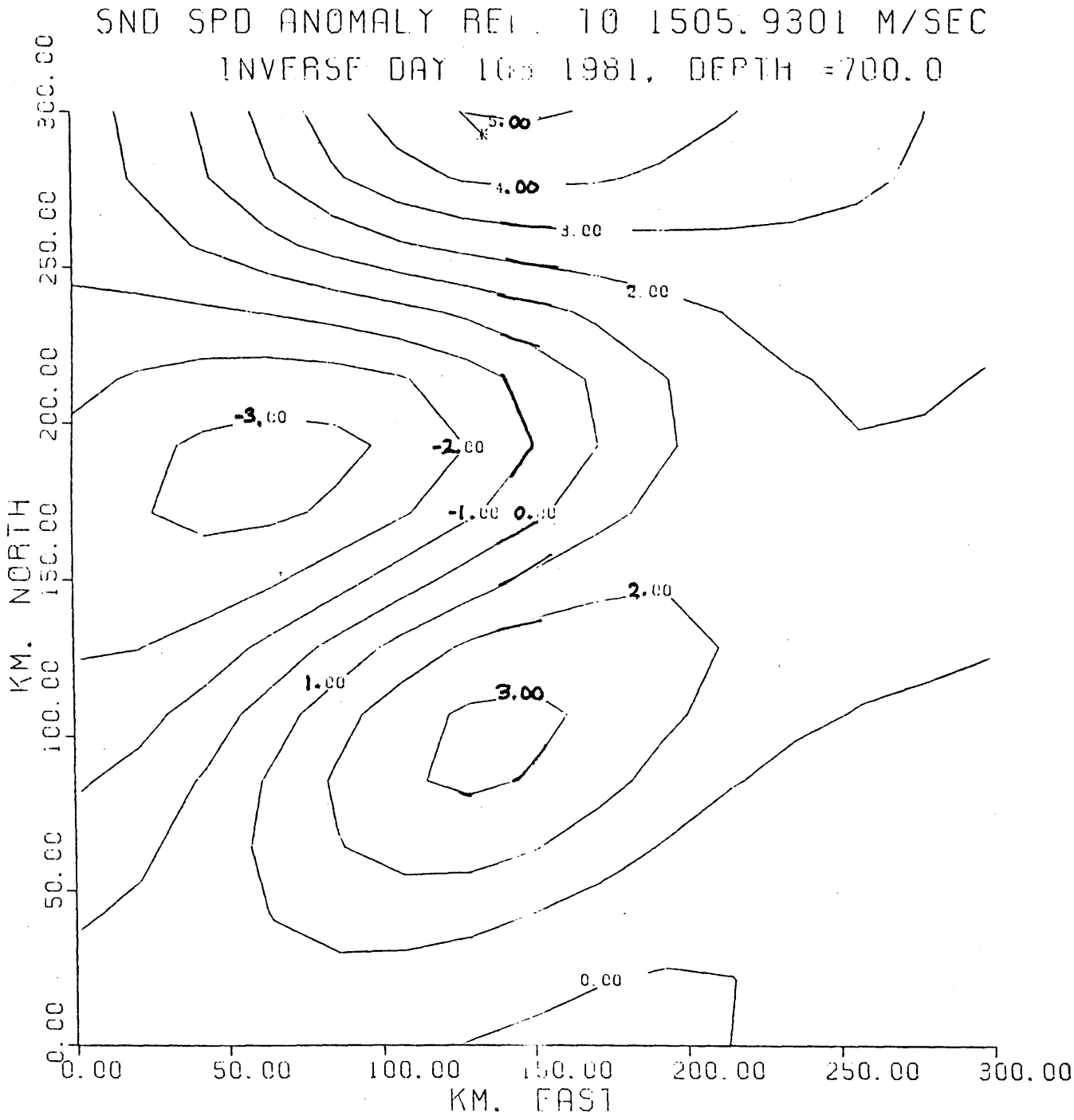

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 88 1981. DEPTH =700.0
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
KM. NORTH
300.00
250.00
200.00
150.00
100.00
50.00
0.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

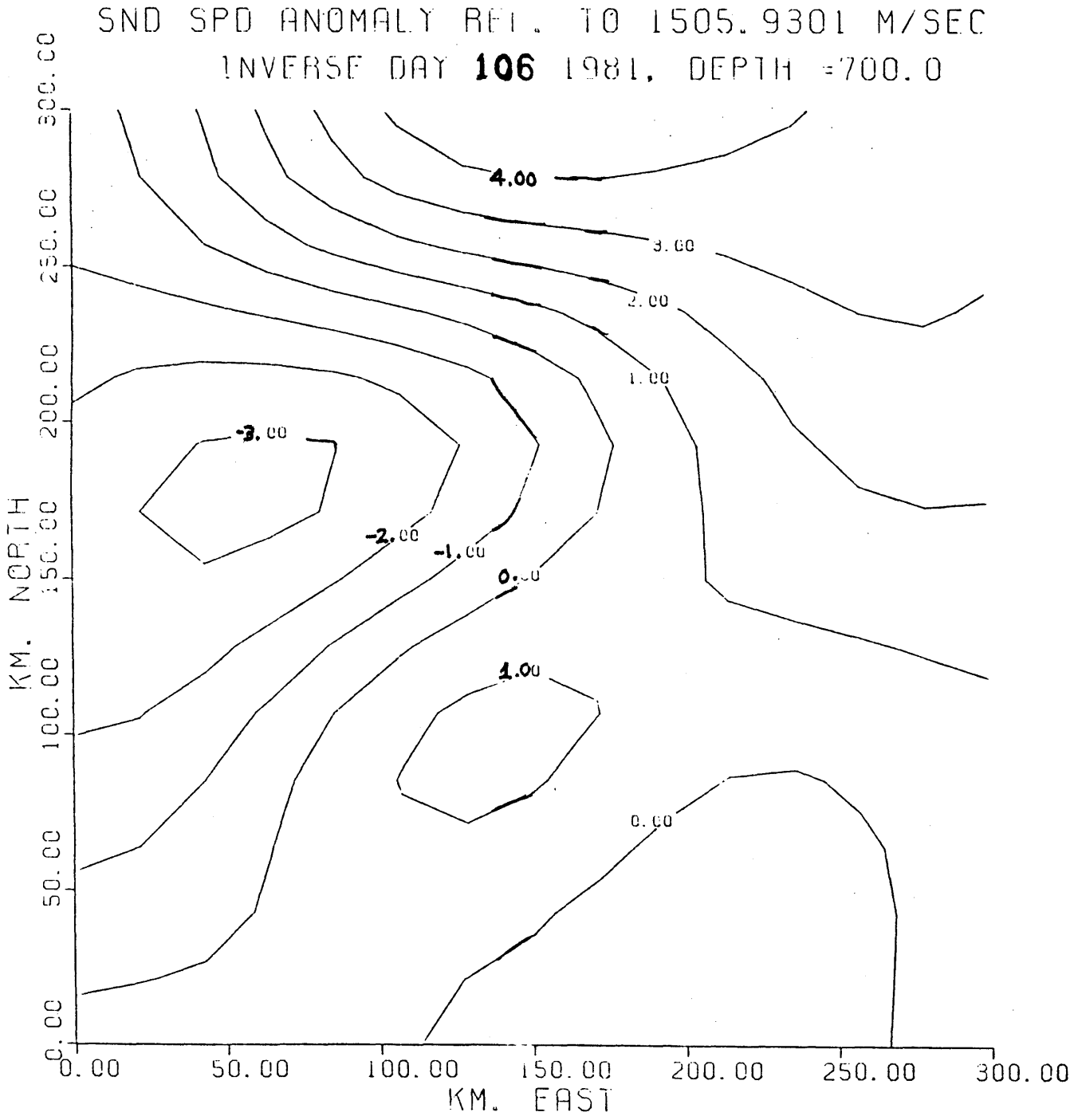

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 91 1981, DEPTH =700.0
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-1.00
0.00
1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM EAST

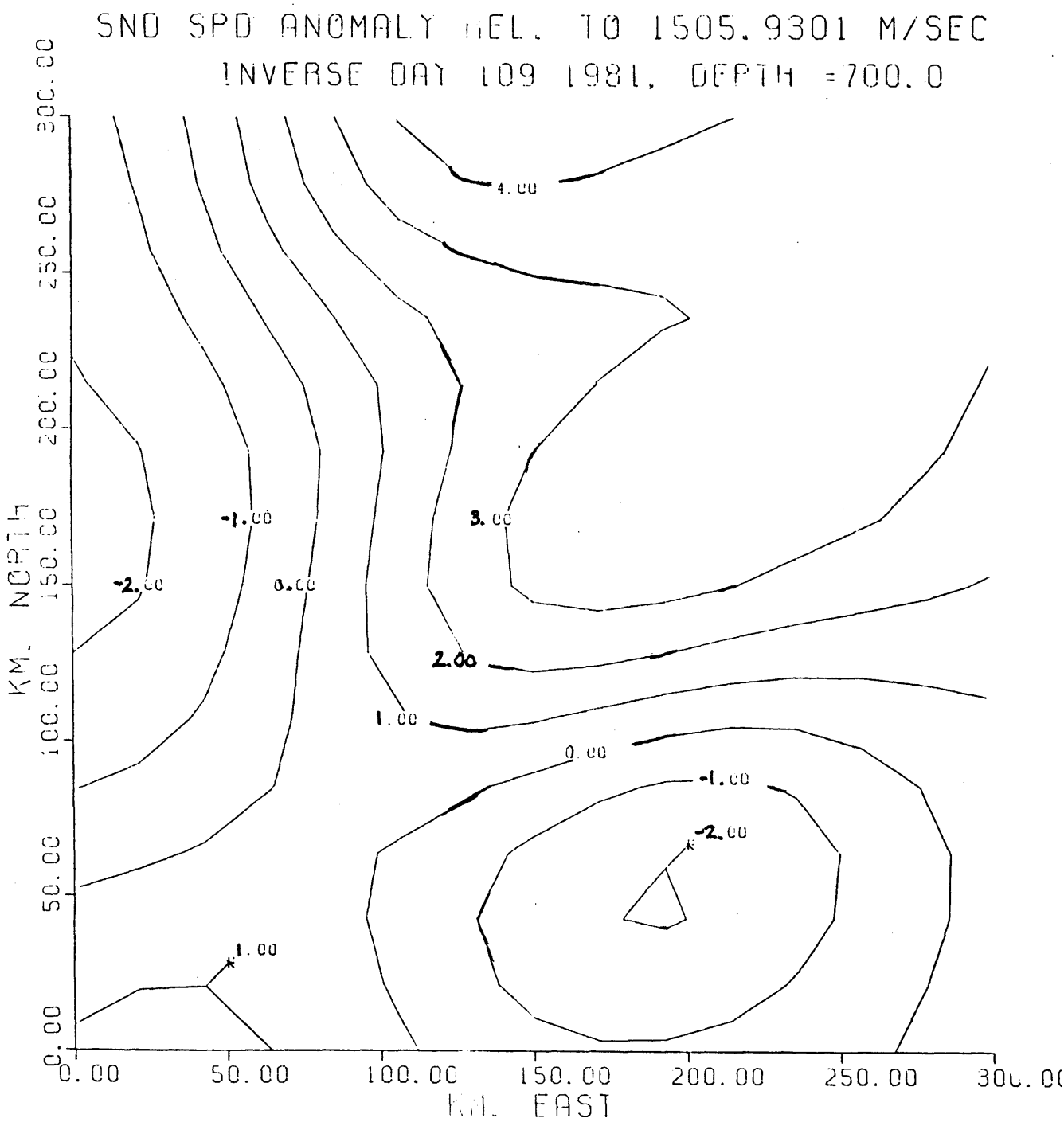

SND SPD ANOMALY FIL. TO 1505.9301 M/SEC
INVERSE DAY 94 1981. DEPTH =700.0
-1.00
5.00
3.00
4.00
2.00
1.00
0.00
-1.00
0.00
-1.00
0.00
-2.00
-1.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM EAST

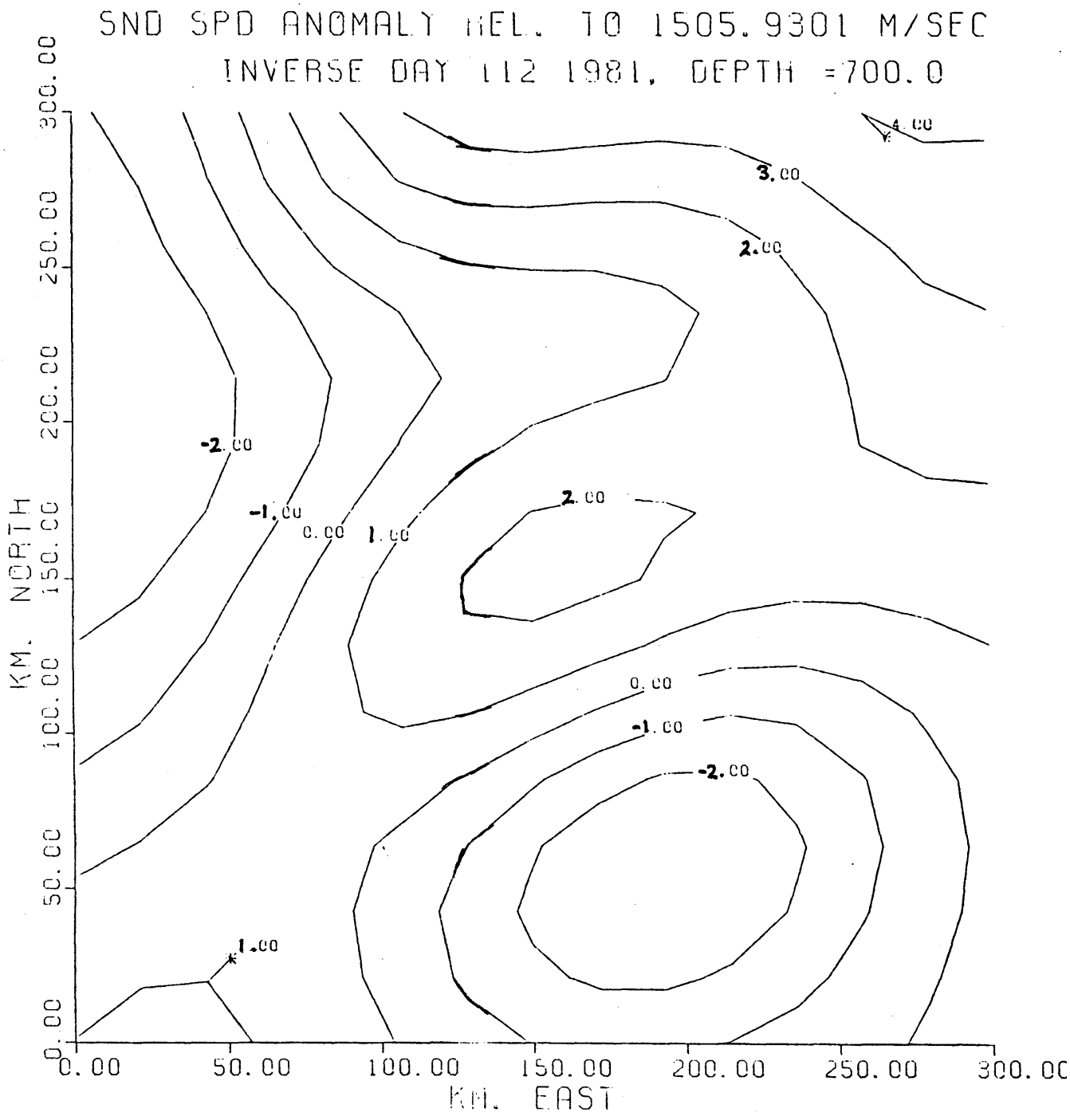

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 97 1981, DEPTH =700.0
0.00
1.00
2.00
1.00
-2.00
3.00
-3.00
-1.00
0.00
3.00
4.00
5.00
6.00
7.00
100.00
-2.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST

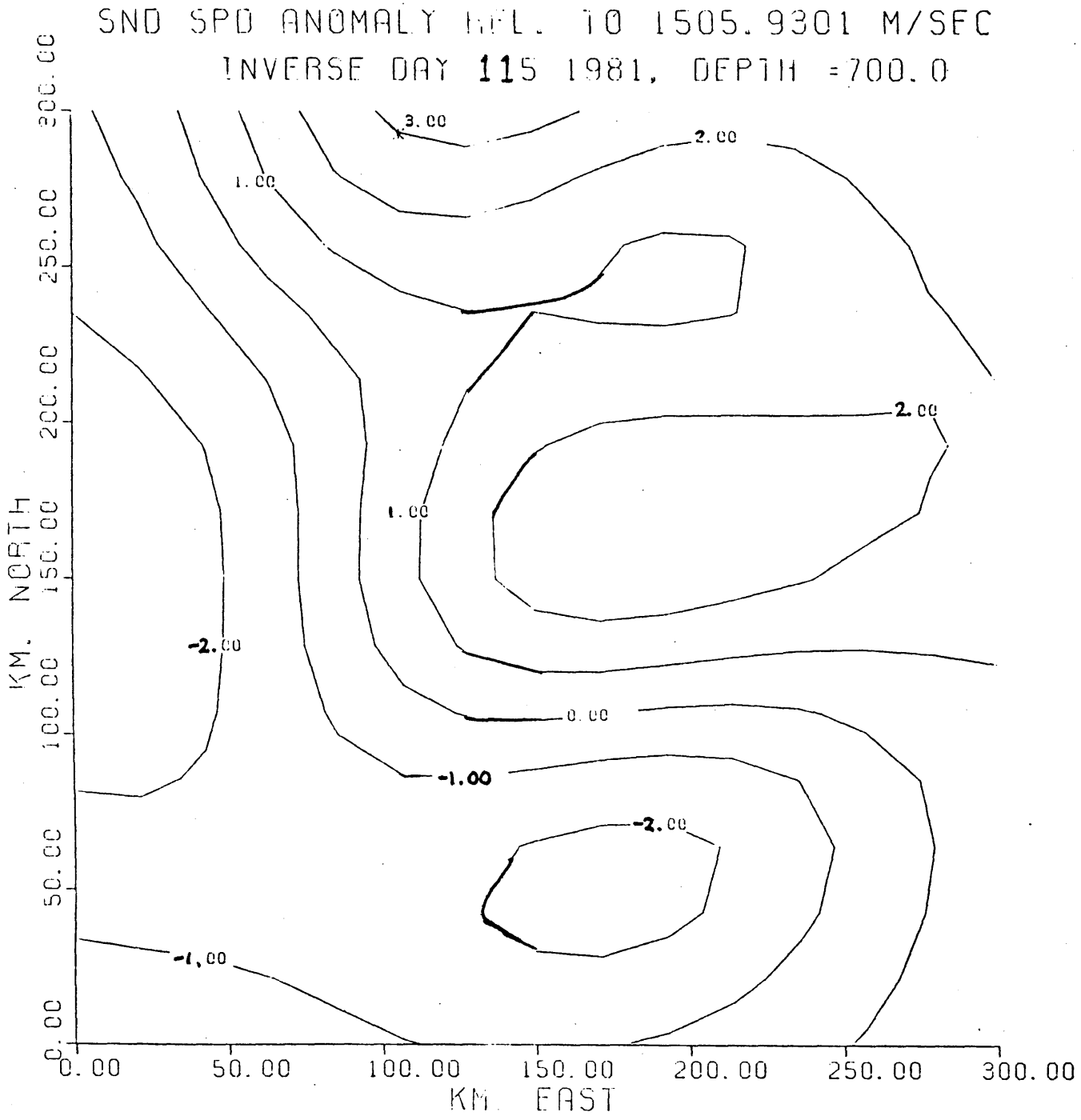

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 100 1981, DEPTH =700.0
4.00
3.00
-2.00
-3.00
-4.00
-1.00
0.00
1.00
2.00
3.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

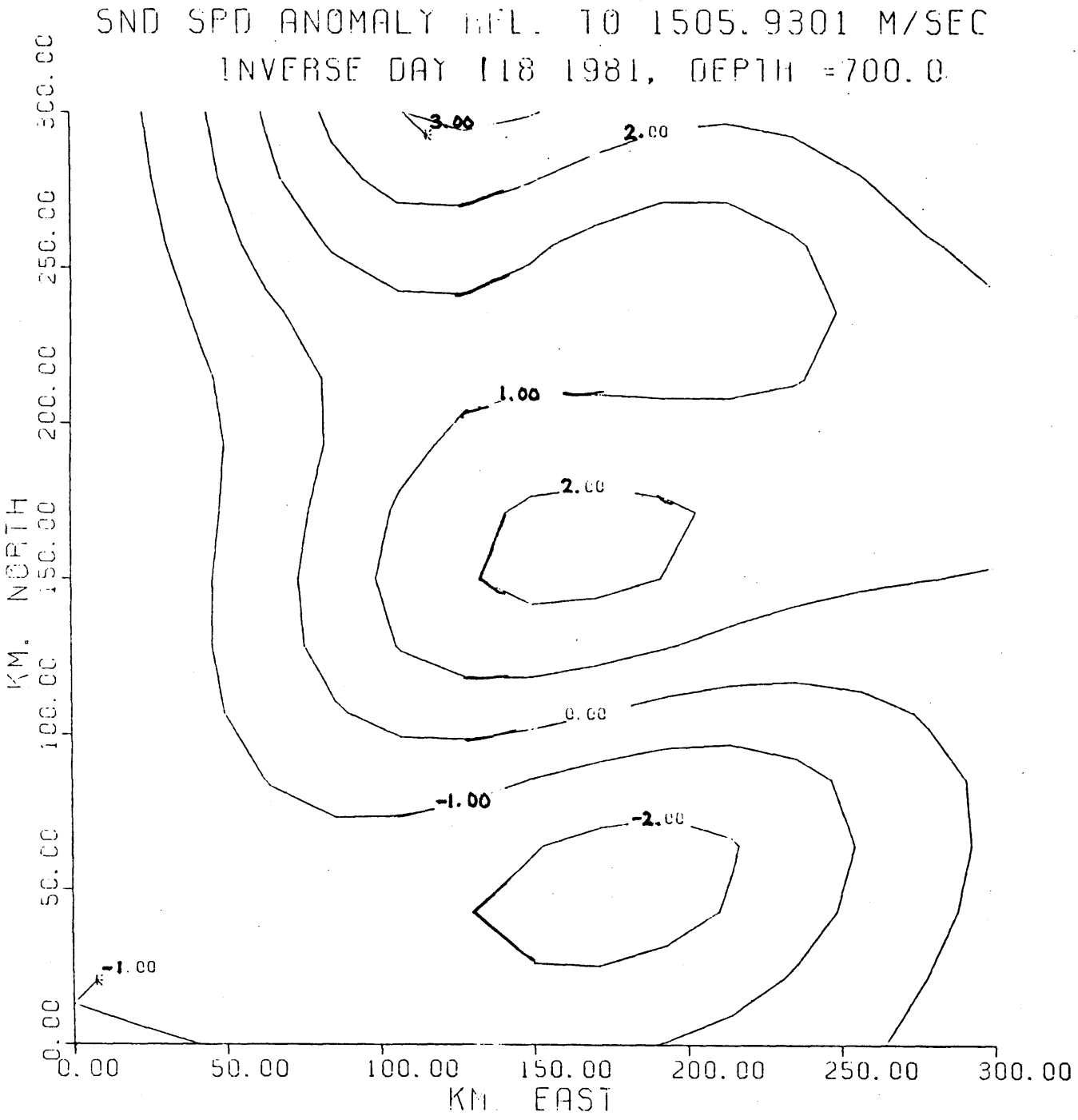

SND SPD ANOMALY REF. TO 1505.9301 M/SEC
INVERSE DAY 100 1981, DEPTH =700.0
5.00
4.00
3.00
2.00
-3.00
-2.00
-1.00
0.00
2.00
1.00
3.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

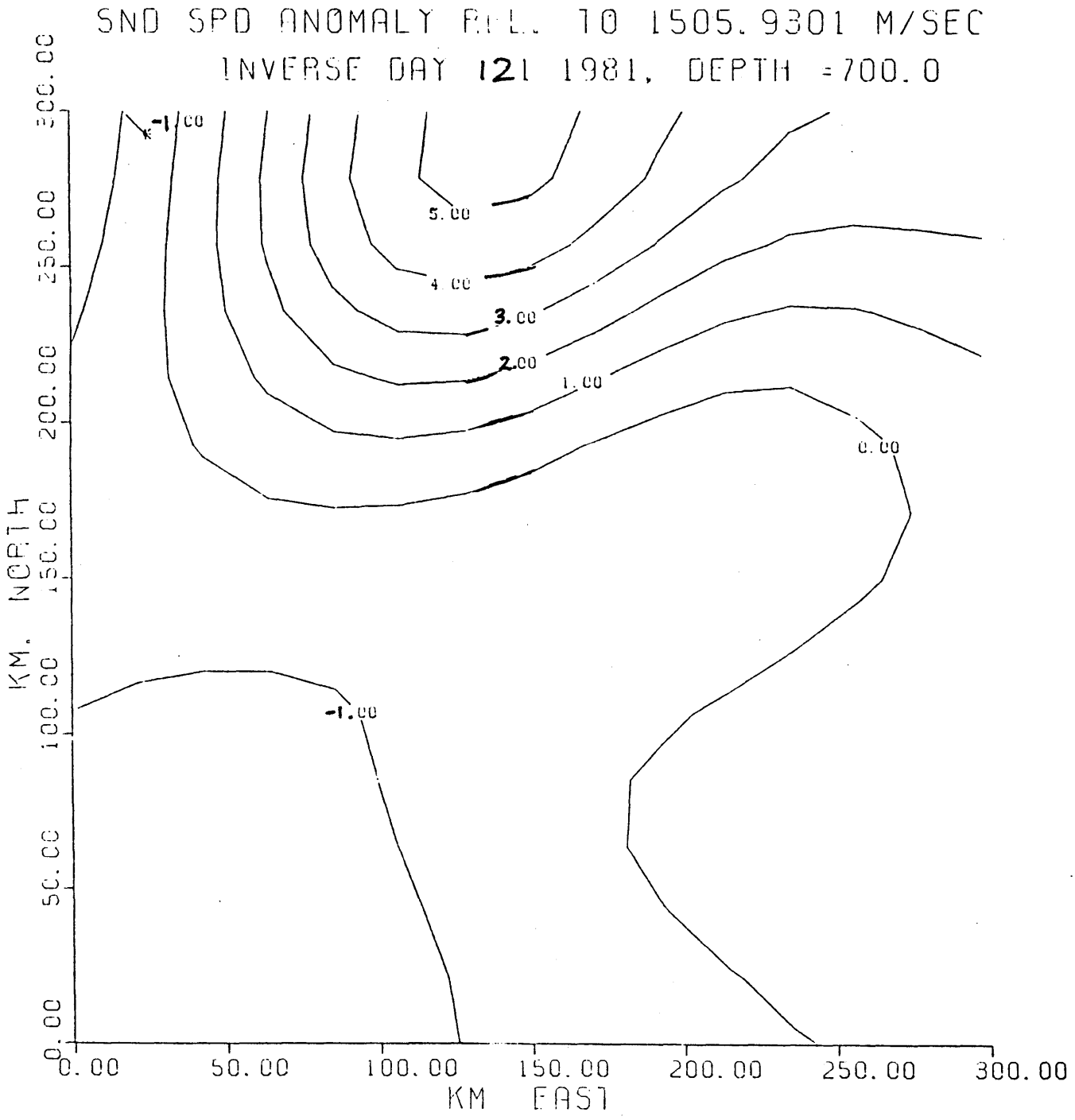

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 106 1981, DEPTH =700.0
4.00
3.00
2.00
1.00
-3.00
-2.00
-1.00
0.00
1.00
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

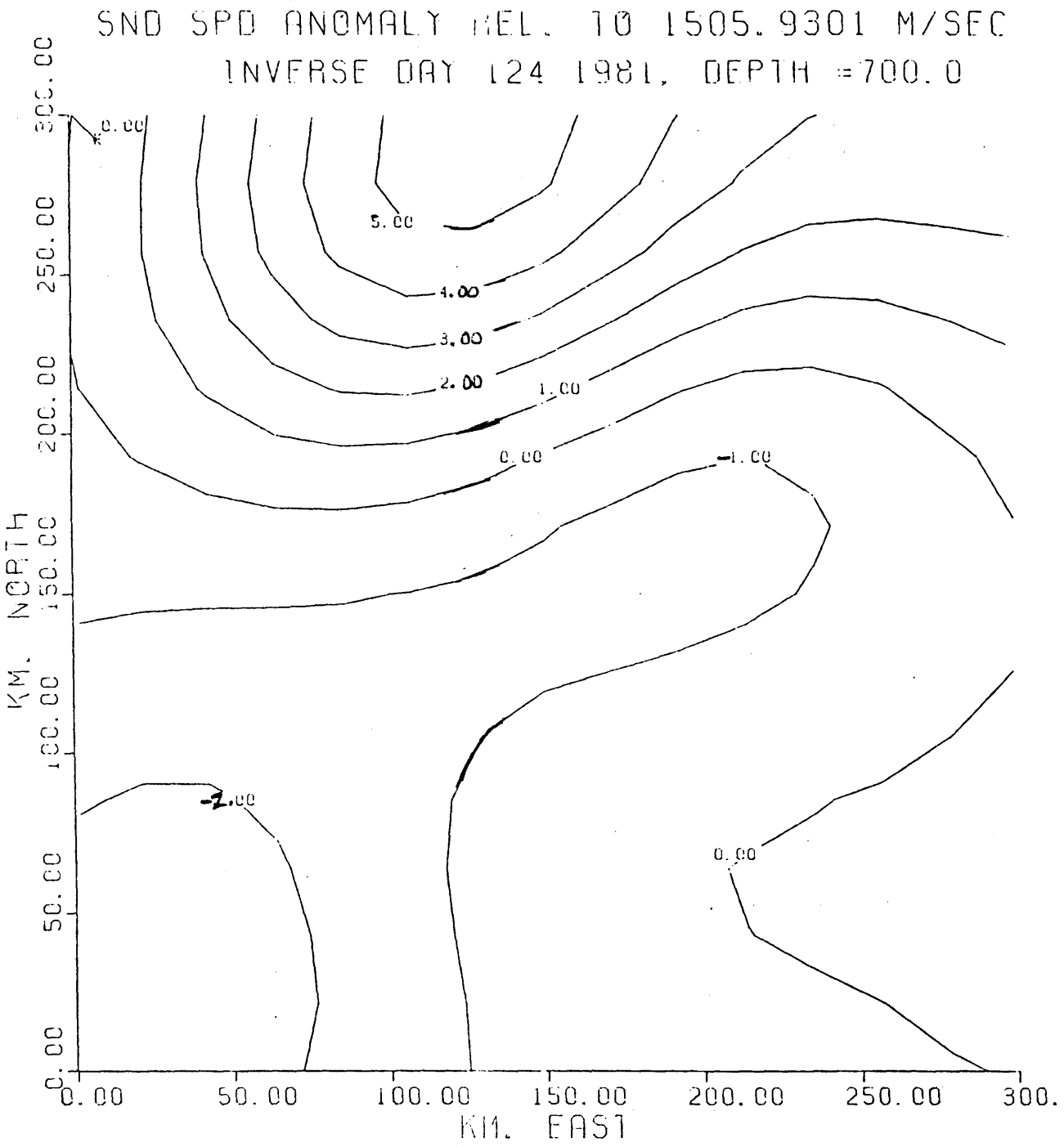

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 109 1981, DEPTH =700.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00

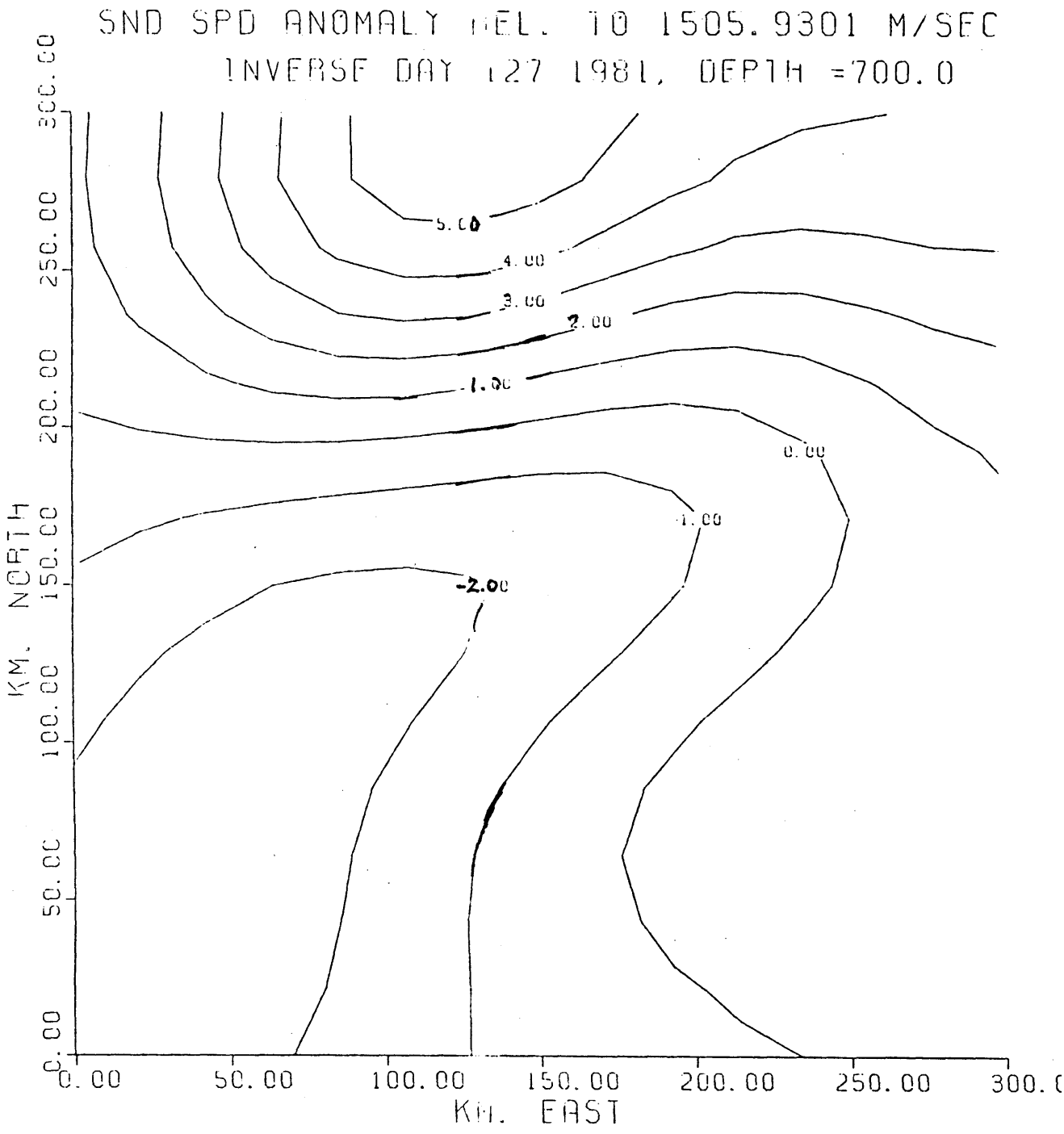

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 112 1981, DEPTH =700.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00
4.00
3.00
2.00
-2.00
-1.00
0.00
1.00
2.00
0.00
-1.00
-2.00
1.00

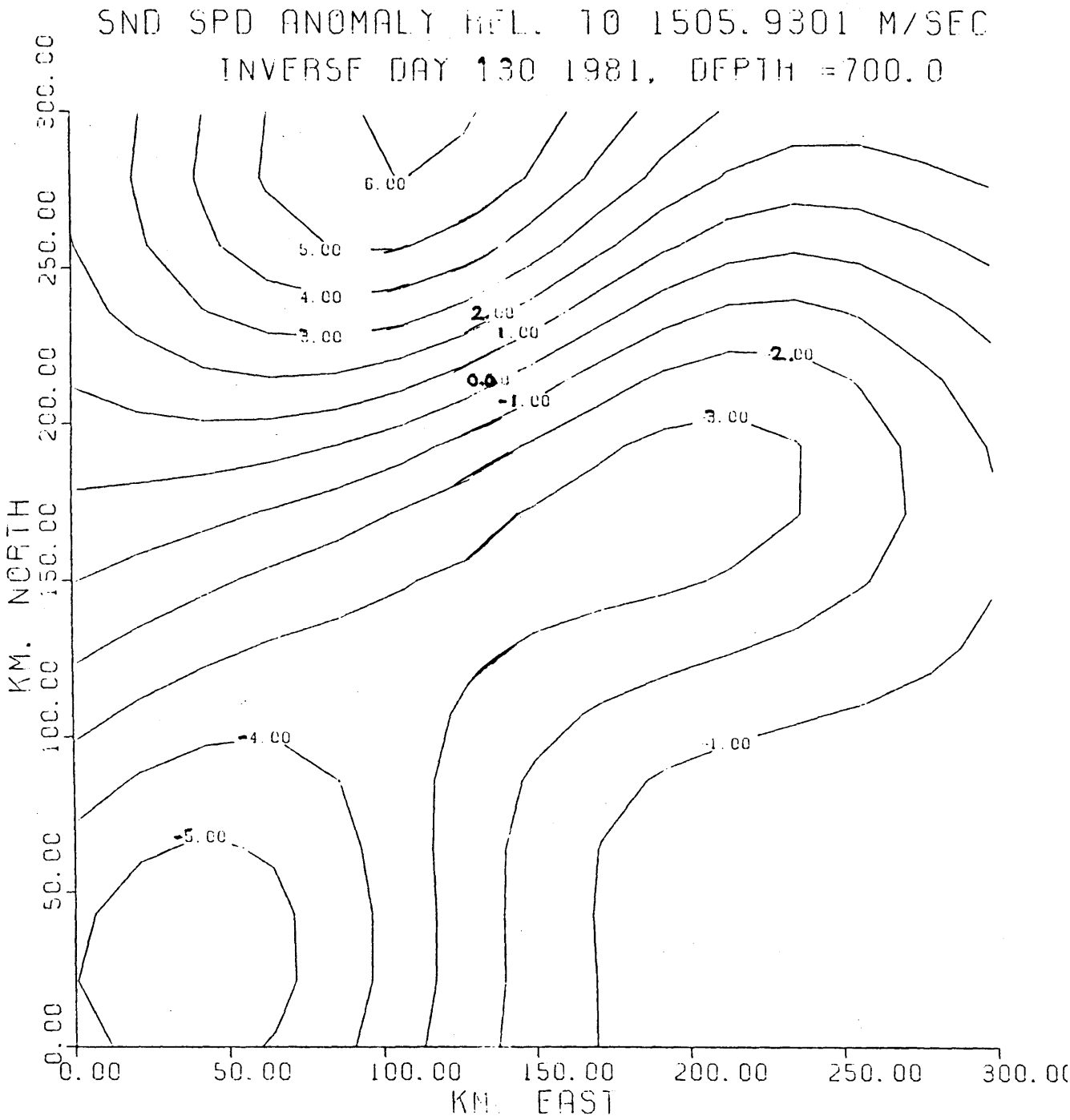

SND SPD ANOMALY INFL. TO 1505.9301 M/SEC
INVERSE DAY 115 1981, DEPTH =700.0
3.00
2.00
1.00
2.00
1.00
-2.00
0.00
-1.00
-2.00
-1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

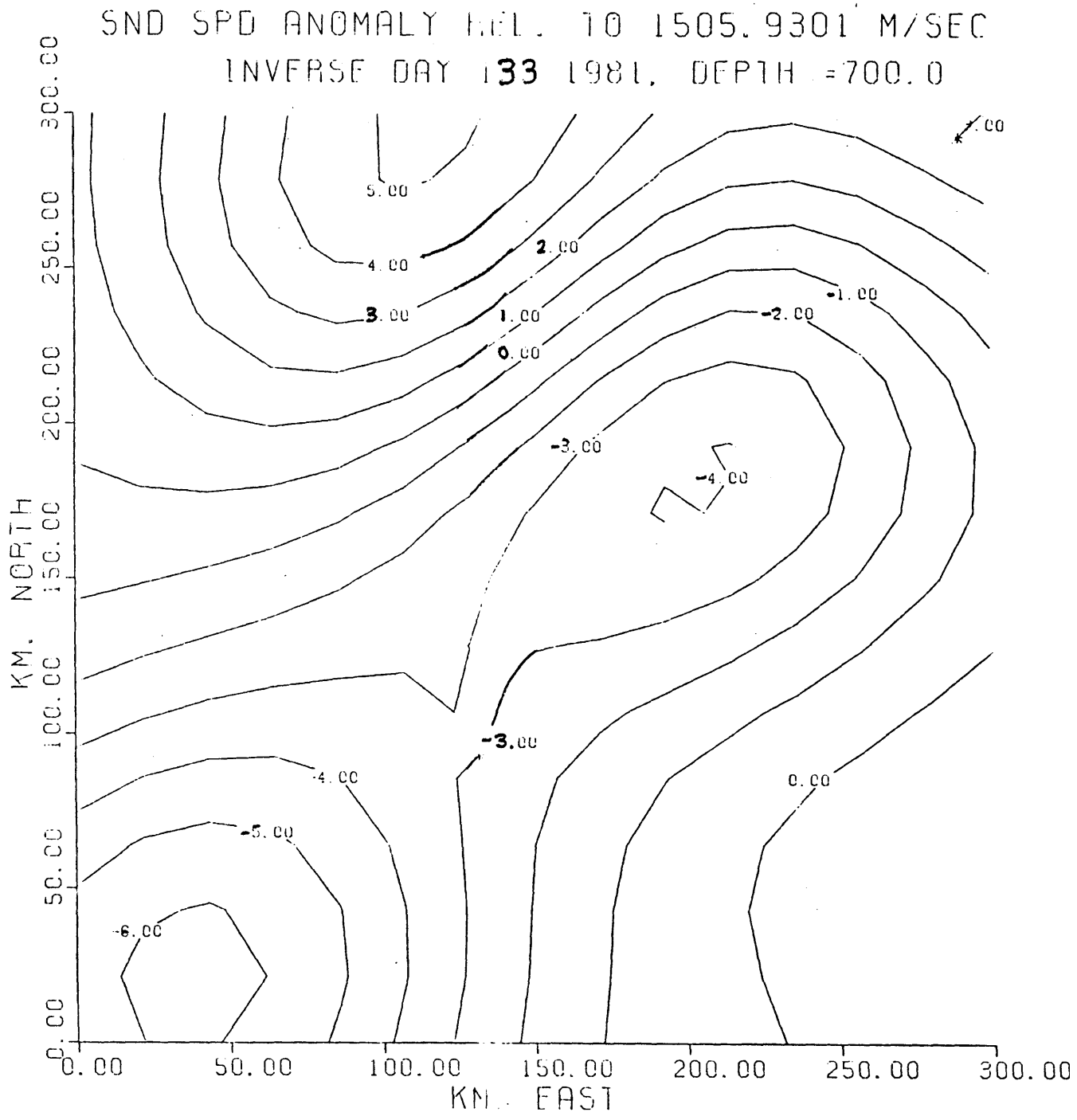

SND SPD ANOMALY INTL. TO 1505.9301 M/SEC
INVERSE DAY 118 1981, DEPTH =700.0
3.00
2.00
1.00
2.00
0.00
-1.00
-2.00
-1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

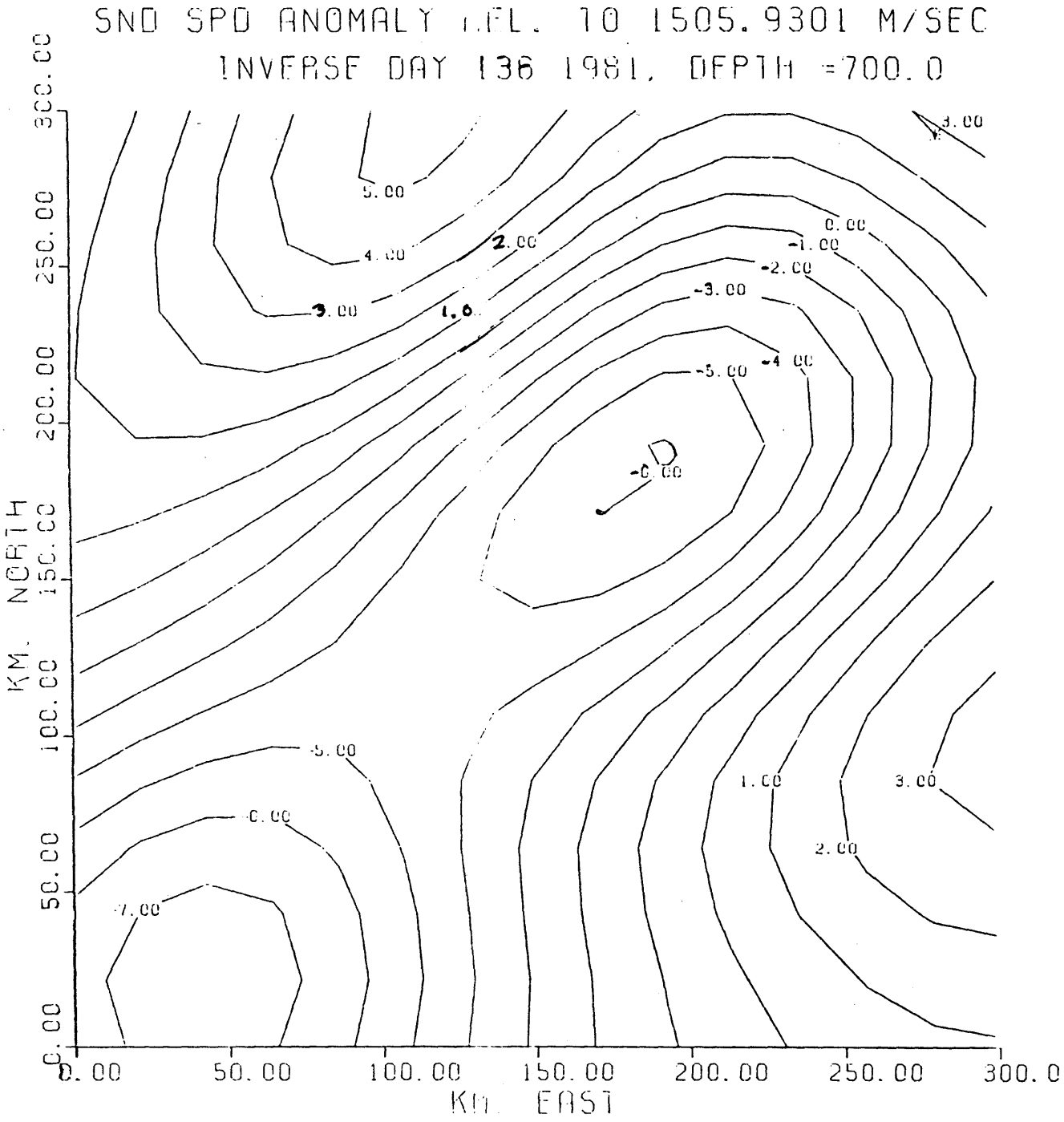

SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 121 1981, DEPTH =700.0
-1.00
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM EAST

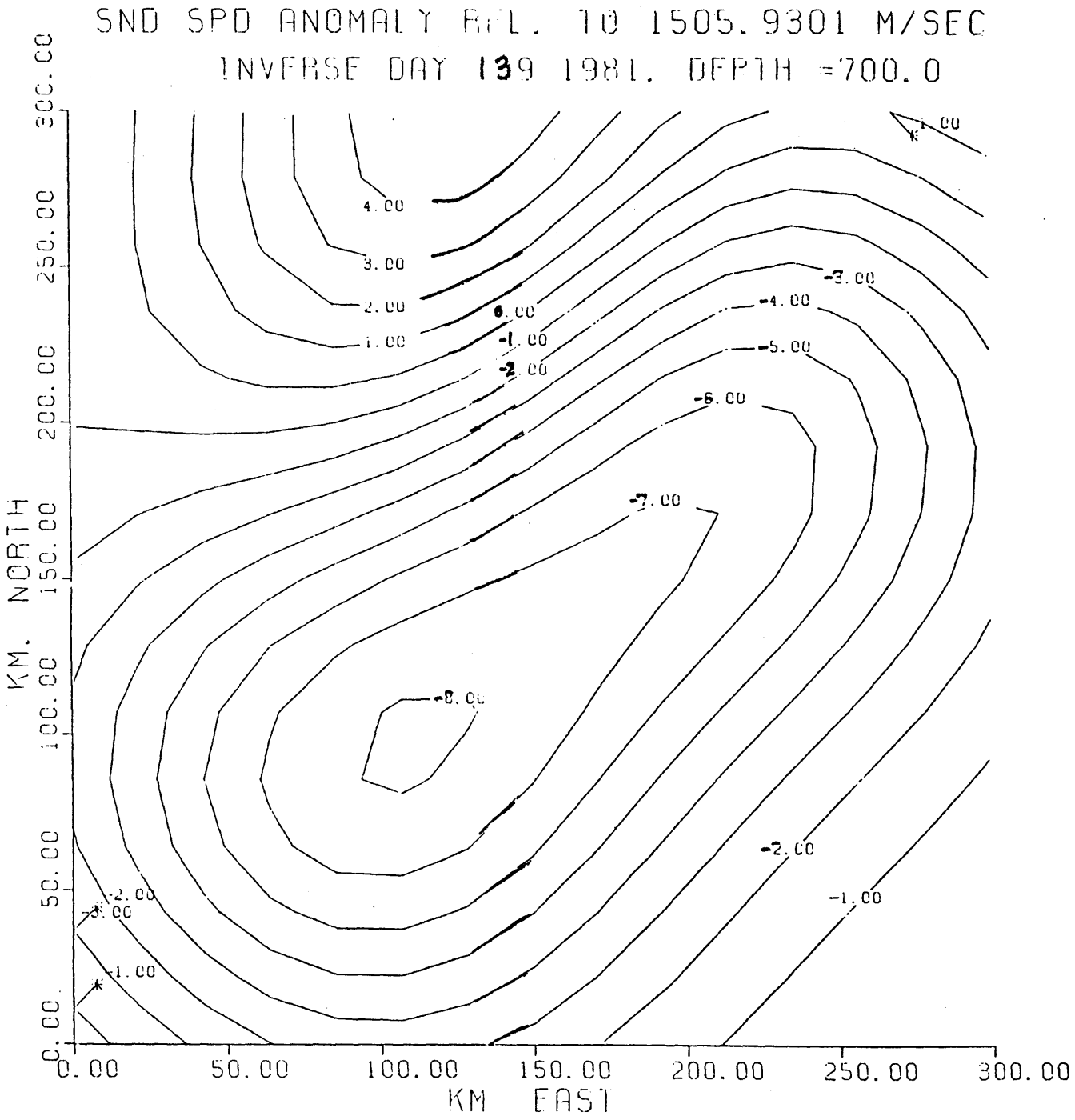

SND SPD ANOMALY MEL. TO 1505.9301 M/SEC
INVERSE DAY 124 1981, DEPTH =700.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
0.00


SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 127 1981, DEPTH =700.0
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY MFL. TO 1505.9301 M/SEC
INVERSE DAY 130 1981, DEPTH =700.0
6.00
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
2.00
3.00
-4.00
-5.00
-1.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 133 1981. DEPTH =700.0
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
-6.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY CFL. TO 1505.9301 M/SEC
INVERSE DAY 136 1981, DEPTH =700.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.0
5.00
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
-6.00
-7.00


SND SPD ANOMALY REL. TO 1505.9301 M/SEC
INVERSE DAY 139 1981. DEPTH =700.0
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
-6.00
-7.00
-8.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM EAST

parameterized in the inversions, but clock corrections obviously do not depend on ray geometry, and faulty corrections do not add random errors, so the clock corrections were not uniformly removed from the data.

The mooring offset data-data covariance was then constructed as in Chapter 7, using the forward problem for mooring parameters with a diagonal "model" covariance matrix made up of the expected variances of the mooring offset parameters. Typical values of expected mooring offset parameters variances as used in the inverses are shown in Table 9.2. These rough estimates were based on previous experience and on records from Temperature-Pressure (T-P) sensors mounted at various depths on the moorings, and were intended to be generous for maximum immunity to errors and freak events. Because the inverses were time independent, the uncertainties in mooring anchor location were lumped with the expected motions even though the anchor positions are constant throughout the experiment.

A significant reduction in the horizontal motion variances can be achieved by separating mooring motion from anchor offset and assigning them appropriate temporal covariance matrices, but that will be covered in later work. An approximation to this procedure was used for

TABLE 9.2: EXPECTED MOORING OFFSETS

Source #	(meters) x	y	z	(seconds) t
1	800	800	10	0.01
2	800	800	80	0.01
3	800	800	30	0.01
4	1100	1100	140	0.40

Receiver #	x	y	z	t
1	500	500	10	0.10
2	500	500	50	0.01
3	500	500	30	0.01
4	500	500	30	0.01
5	500	500	40	0.10

These numbers were input to the estimation framework in order to bound the uncertainties of these parameters

these maps, in which the travel times for each path were averaged throughout the experiment and then fed to the inverse operator to give rough estimates of the mooring anchor positions (Table 9.3). Numerical rays corresponding to those found in the data were then traced for these positions, so that some of the initial uncertainty was removed from the data.

T-P recorders on some of the moorings gave useful estimates of instrument depth offsets, but the inverses were calculated without using this information, except in adjusting the offset parameter variances, as mentioned above. The vertical position uncertainties, like the horizontal offsets, have 2 components. The "rest" depth of an instrument is its depth when the mooring is vertical and straight, and should ideally be the depth that was specified when the mooring was designed. The actual depth is estimated from the local bottom depths, the cable lengths as specified in the mooring plan, and any T-P information available from the mooring. If the T-P recorder was attached at the hydrophone then the uncertainty in "rest" depth would be only about 1 meter,

TABLE 9.3: ORIGINAL AND ESTIMATED MOORING POSITION

TOP = ORIGINAL POSITION BOTTOM = ESTIMATED POSITION

Source #	(KM) x	(KM) y	(M) z	(MSEC) t
1	17.336	284.287	2150.	0.00
	19.047	283.623	2150.	0.00
2	16.216	207.377	1995.	0.00
	16.843	207.139	1980.	0.00
3	17.964	91.735	2120.	0.00
	17.649	91.618	2117.	0.00
4	18.014	16.122	2143.	0.00
	17.657	16.084	2123.	0.00

Receiver #	(KM) x	(KM) y	(M) z	(MSEC) t
1	281.490	286.696	1694.	0.00
	281.068	286.537	1698.	0.00
2	283.357	189.957	1325.	0.00
	282.494	189.887	1370.	0.00
3	284.155	114.344	1708.	0.00
	283.271	115.425	1675.	0.00
4	281.607	19.273	1744.	0.00
	281.285	20.509	1700.	0.00
5	146.190	281.693	1695.	0.00
	147.013	280.661	1616.	0.00

The estimated positions were calculated using an average of travel time throughout the experiment and so may not truly represent the anchor positions.

fixed by the level of calibration and the least significant bit. The rest depth is the minimum depth observed by the T-P sensor, since as the mooring leans, the instrument depth can only increase. If the rest depth were known, then the positivity of the depth perturbations would allow the use of maximum-entropy inversion algorithms, but in the 1981 experiment, the errors were generally greater than the T-P error alone. Most moorings had an uncertain length of cable between the T-P recorder and the hydrophones, and the mooring R2 had no T-P data at all. These uncertainties provide much of the variances listed for the receivers in Table 3 because the receivers tended not to have large vertical excursions.

The other source of variance is, naturally, mooring motion, which acounts for much of the variance listed for the sources. On moorings with working T-P recorders near the instrument, most of the depth changes could be corrected for, down to the level of T-P and cable length errors, but this was not done for the maps in Figure 9.6. The inverse thus produced time series of sound speed in the 300 km X 300 km box, instrument x,y and z coordinate, and clock offset.

The final instrument-related source of travel time variance is the drift of the quartz oscillator. The low-power clocks were compared daily against a rubidium frequency standard, and the measured frequency shifts were recorded on tape and integrated to estimate clock offsets, which are then removed by shifting the time base. Clock corrections were retained for rays to receivers 2,3, and 4 in the data set used for the maps in Figure 9.4, and if the corrections were perfect then no clock error would be expected, and no variance would be needed in the inverse. The variances entered in Table 9.2 are insurance against unexpected problems and/or dropped cycles in the clocks. The clock offsets calculated by the inverse on a given day can be checked against these a priori expectations, and a large mis-match is an indication that re-computation using different limits may be necessary (See Chapter 4).

CHAPTER 10

DISCUSSION AND CONCLUSIONS

10.1 COMPARISONS OF ACOUSTIC AND TRADITIONAL MAPS

In this chapter, I will present a comparison of initial results from the acoustic data taken during the 1981 ocean acoustic tomography experiment with more traditional measurements made more or less concurrently. The inverse produced an independent estimate of the sound speed field for the entire ocean volume within the 300 by 300 km box every 3 days between yearday 52 to 139 of 1981. Data for two of the receivers, numbers 1 and 5, continue until day 172, (Table 8.1), but the time series of maps has not yet been extended completely. NOAA ships made 3 CTD surveys in the area during the time that the moorings were in the water, but only the first two overlap with the acoustic data. There were two environmental moorings deployed as part of the array (Figure 1.4), with current meters and T-P recorders, and the acoustic moorings carried T-P recorders as well. Each observation method, acoustic, CTD, or moored instrument, has particular strengths and weaknesses, which must be taken into account when making the comparison. For example, the CTD surveys observed vertical profiles at about 65 points during a period of nearly 3 weeks, while the acoustics partially sample and average the volume during a single day.

At present, the inverse procedure has been kept simple, estimating sound speed instead of temperature or density; these will be covered in a later paper. The CTD survey has thus been used to calculate sound speed, while the temperature time series from the moorings have been left as temperature. Comparison with sound speed time series can be made on the basis of the curve shapes, using the approximately linear dependence of sound speed on temperature at any given depth.

Figures 2.1 and 10.1 are maps of sound speed anomaly (with respect to the reference $C_o(z)$) calculated from the first 2 NOAA CTD surveys of the region and from one Navy AXBT flight. Unless specified otherwise, all maps of sound speed have been referenced to the basic state. The "traditional" data has been mapped at 700, 350, 1500, and 2000 meters depth, in order to provide a wide range of depths at which to compare the various observation techniques. 700 meters has the maximum energy, and provides the best test of resolution, while the deeper levels are quieter, and the shallow level was picked because it was the deepest the AXBT's could penetrate.

Figure (9.4) shows maps made from corrected, day differential times, while figure (9.6) shows maps made from uncorrected data, with mooring motion, anchor position, and clock offset as part of the unknowns. The day differential

FIGURE 10.1 A SOUND SPEED ANOMALY FIELD AT 350 METERS DEPTH. CALCULATED FROM FIRST NOAA CTD SURVEY, 1981 DAYS 66-85. COUNTOURS ARE M/SEC DIFFERENCE FROM THE AVERAGE SOUND SPEED PROFILE. CONTOUR INTERVAL IS 1 M/SEC.

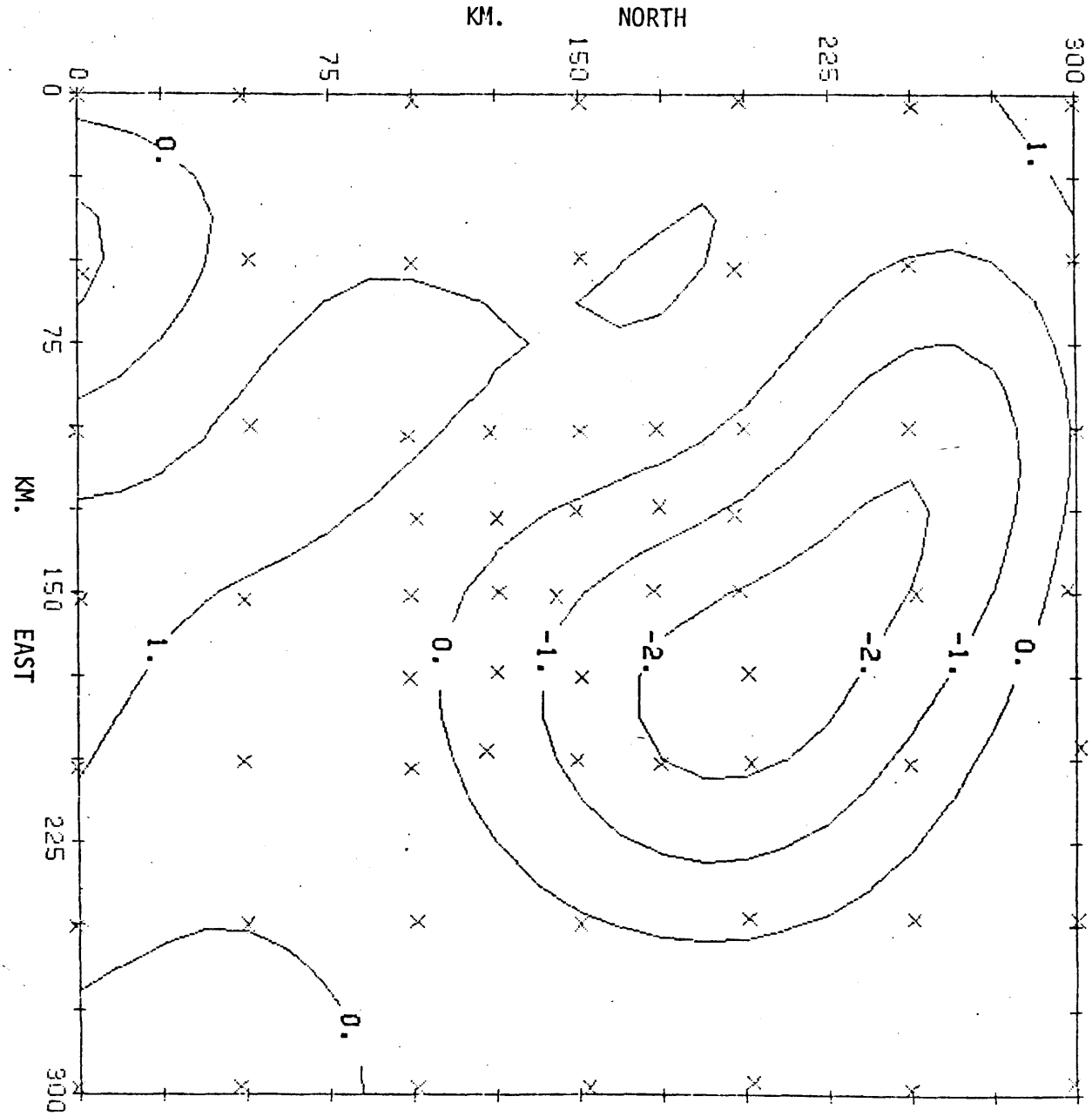


FIGURE 10.1 B SOUND SPEED ANOMALY FIELD AT 350 METERS DEPTH. CALCULATED FROM NAVY AXBT SURVEY, 1981 DAYS 106-7. COUNTOURS ARE M/SEC DIFFERENCE FROM THE AVERAGE SOUND SPEED PROFILE. CONTOUR INTERVAL IS 1 M/SEC.

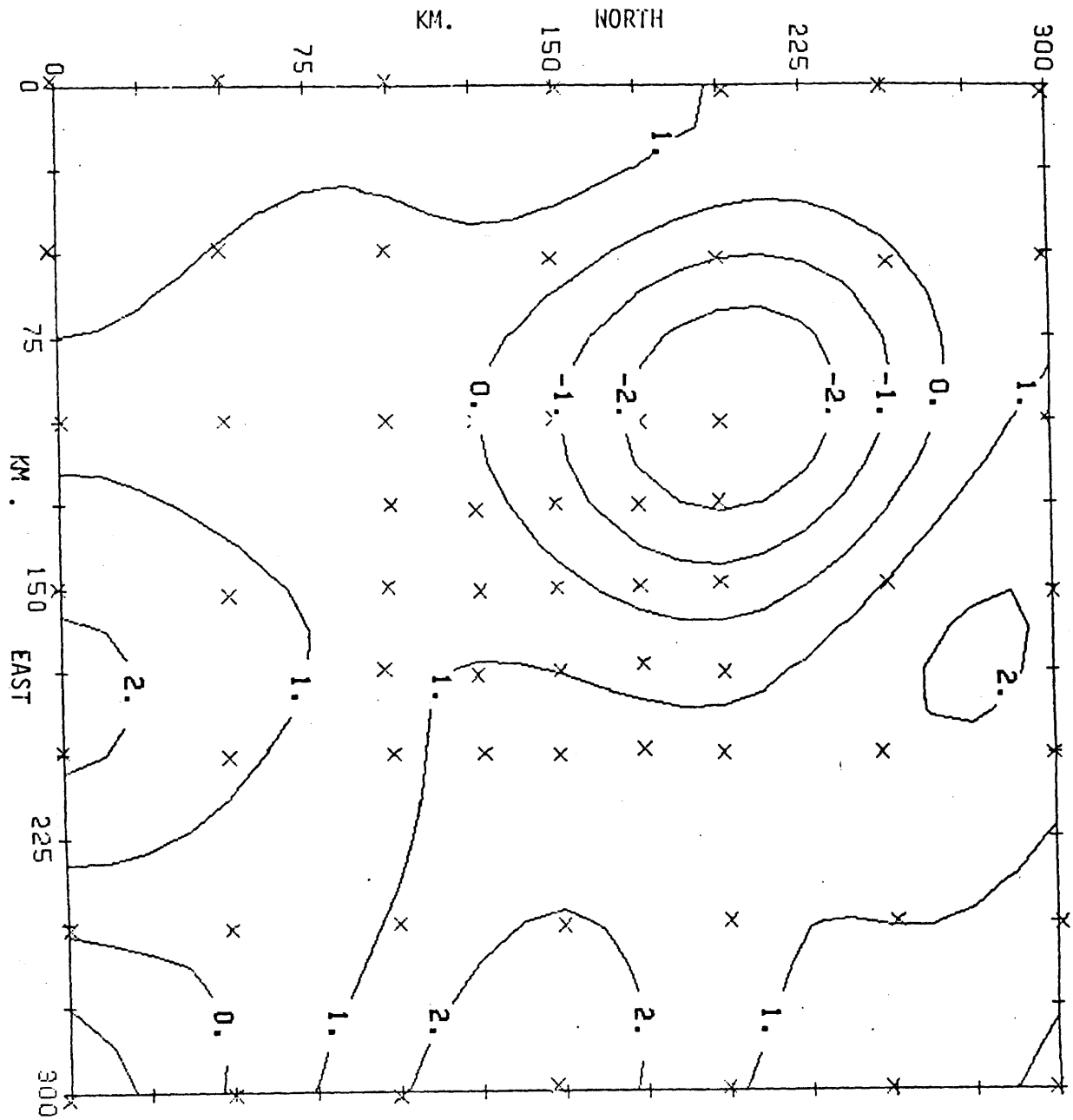


FIGURE 10.1 C SOUND SPEED ANOMALY FIELD AT 350 METERS DEPTH.
CALCULATED FROM 2ND NOAA CTD SURVEY, 1981 DAYS 120-139.

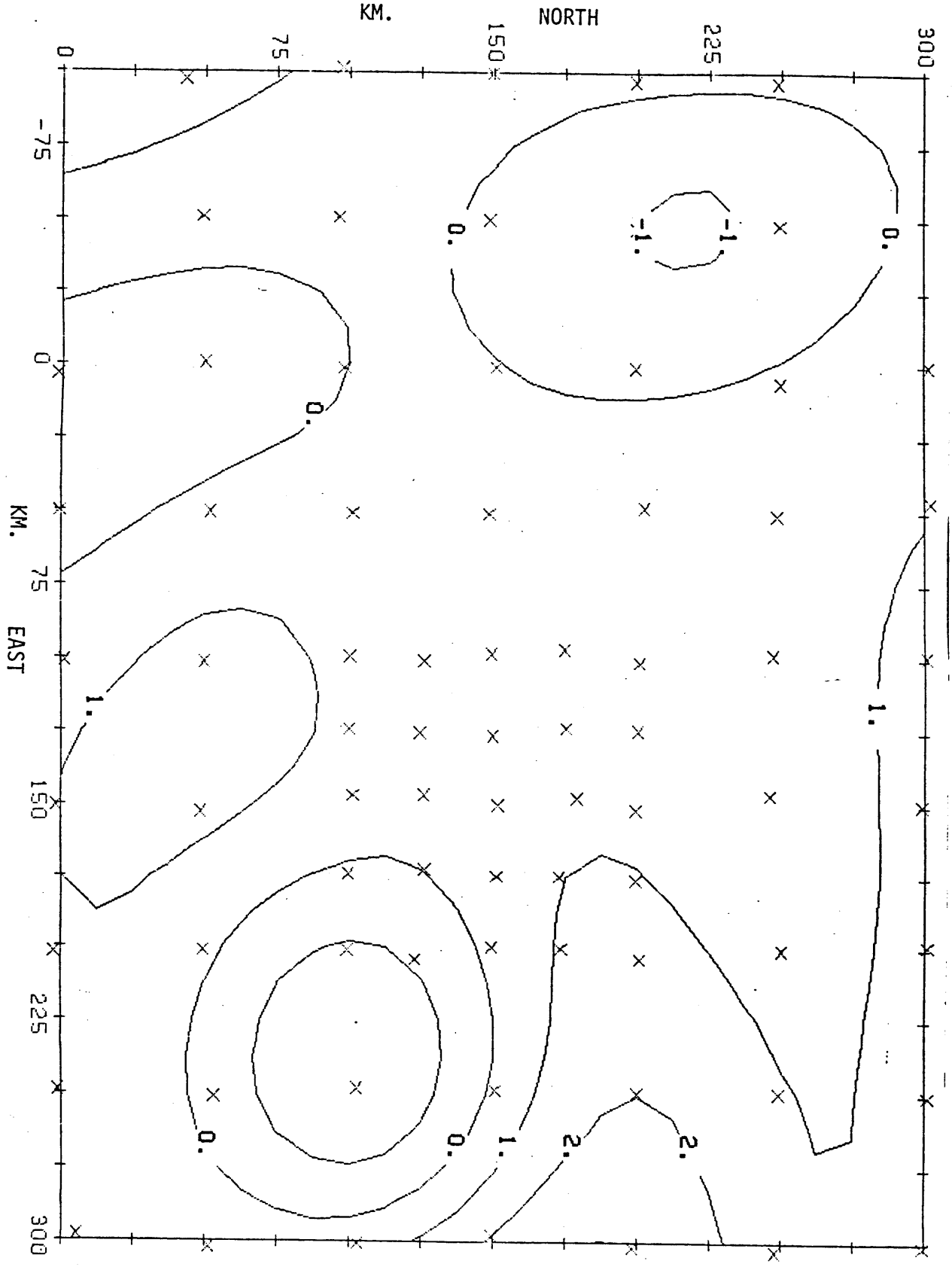


FIGURE 10.1 D SOUND SPEED ANOMALY FIELD AT 700 METERS DEPTH. CALCULATED FROM FIRST NOAA CTD SURVEY, 1981 DAYS 66-85. COUNTOURS ARE M/SEC DIFFERENCE FROM THE AVERAGE SOUND SPEED PROFILE. CONTOUR INTERVAL IS 1 M/SEC. (DUPLICATE OF FIGURE 2.1)

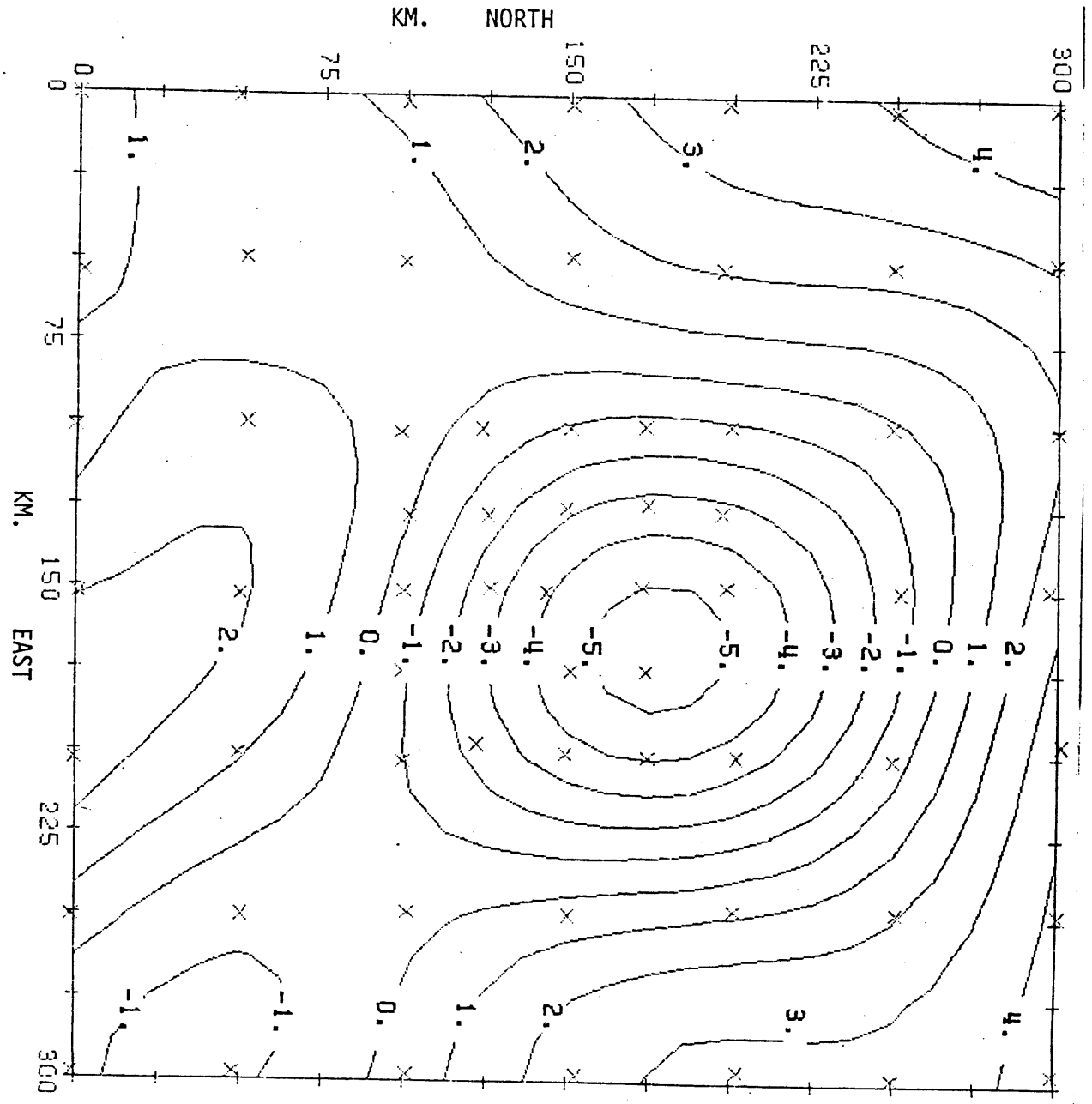


FIGURE 10.1 E SOUND SPEED ANOMALY AT 700 M. 2ND NOAA CTD SURVEY.

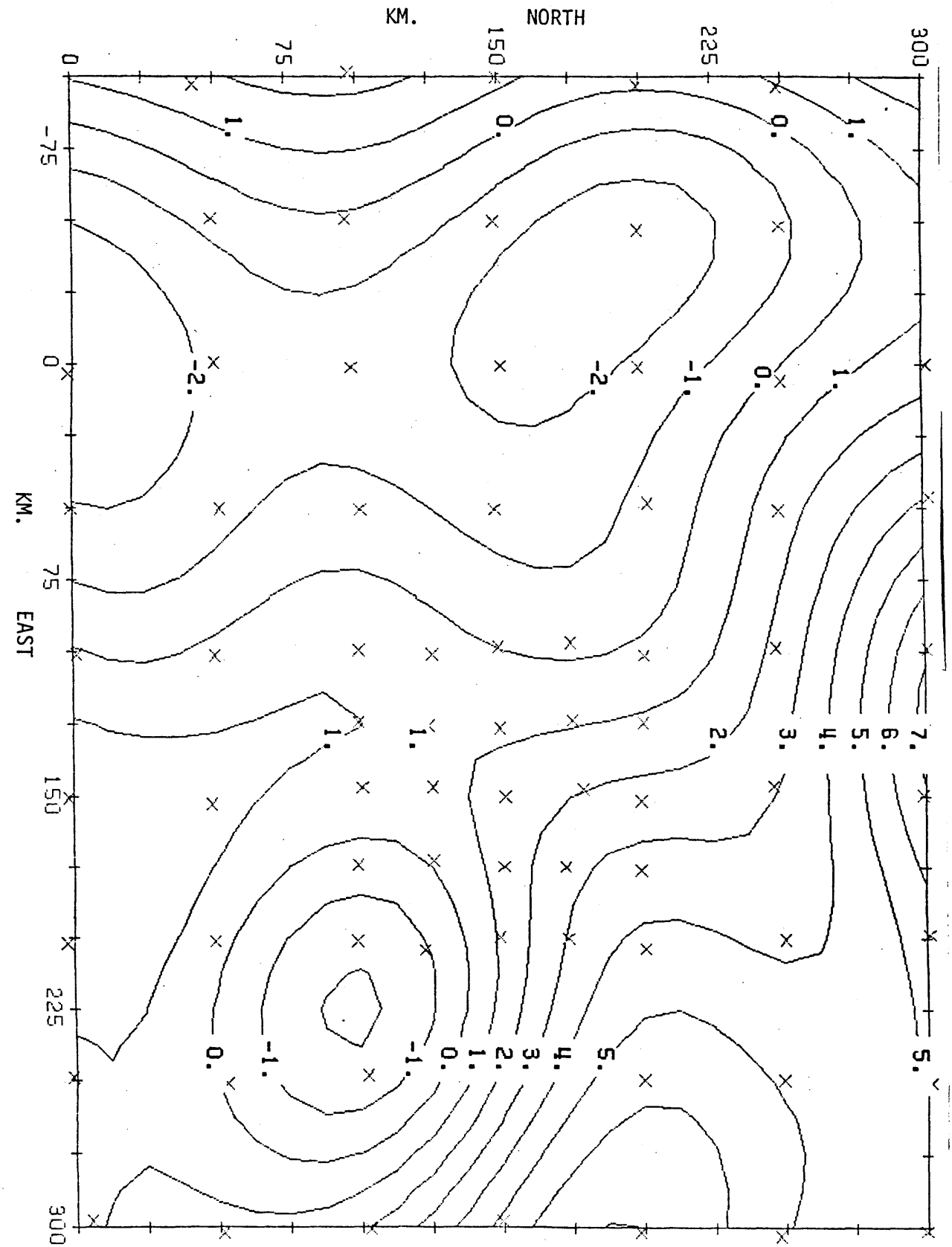


FIGURE 10.1 F SOUND SPEED ANOMALY AT 1500 M. 1ST NOAA CTD SURVEY.

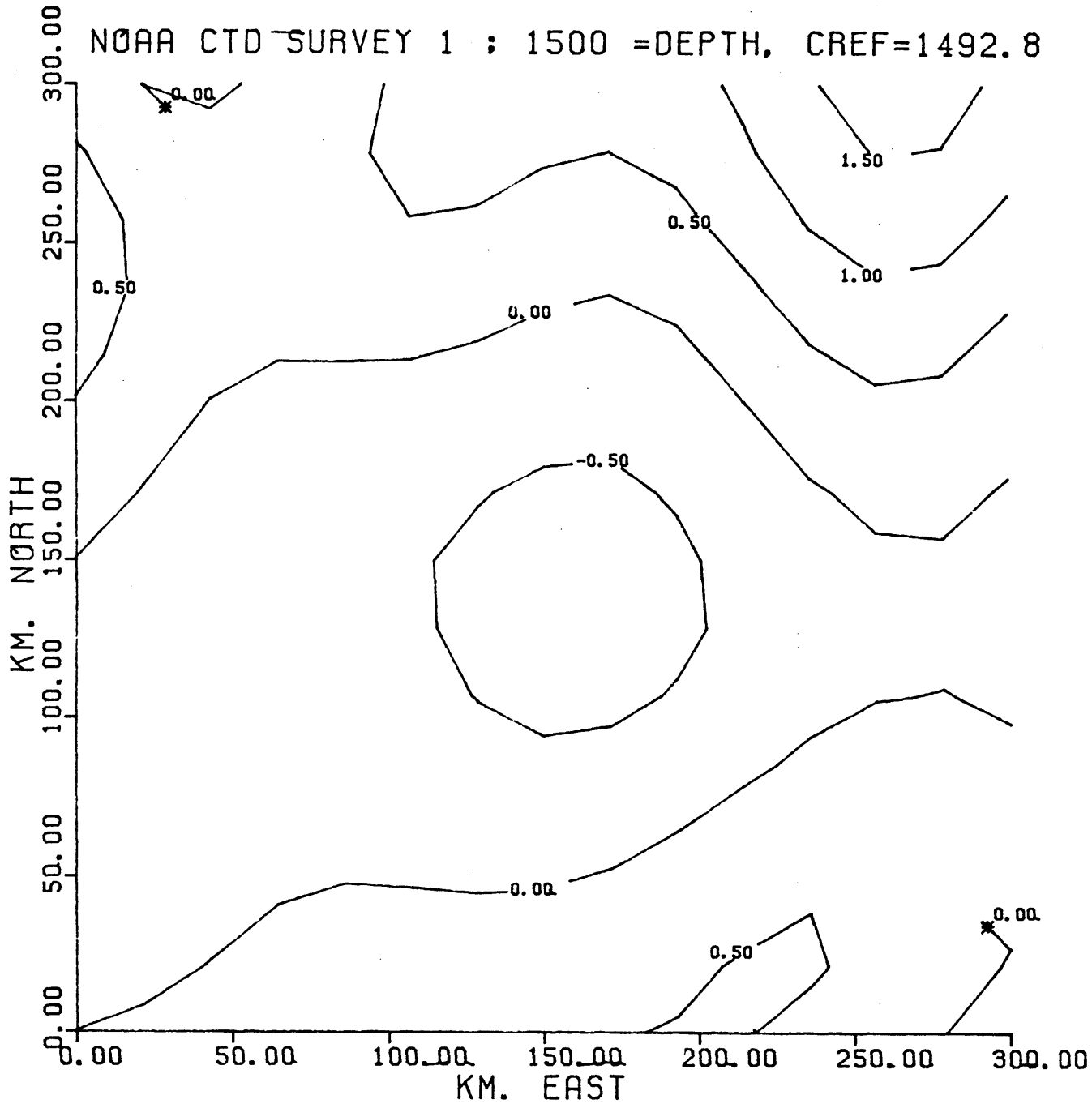


FIGURE 10.1 G SOUND SPEED ANOMALY AT 1500 M. 2ND NOAA CTD SURVEY.

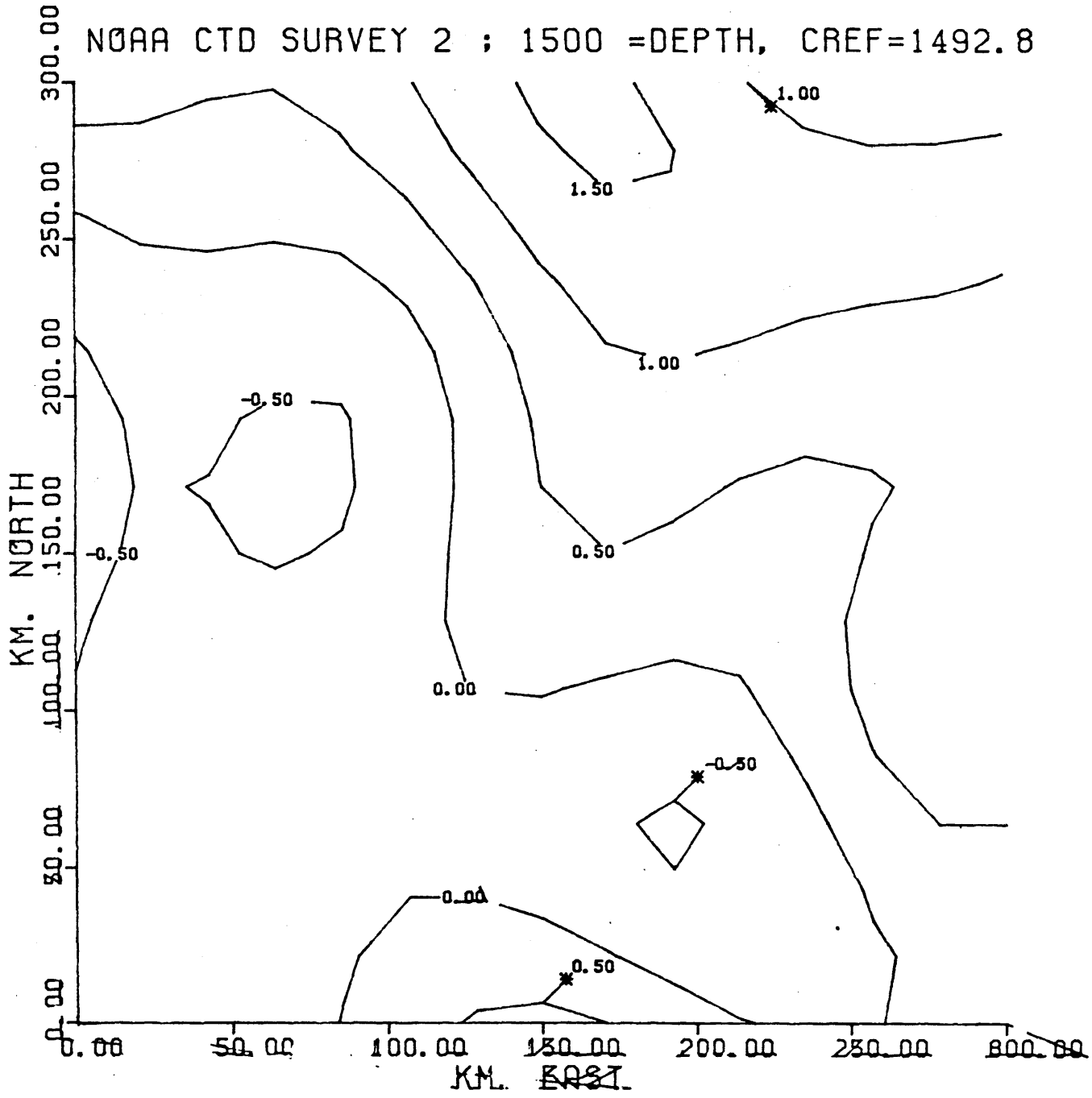


FIGURE 10.1 H SOUND SPEED ANOMALY AT 2000 M. 1ST NOAA CTD SURVEY.

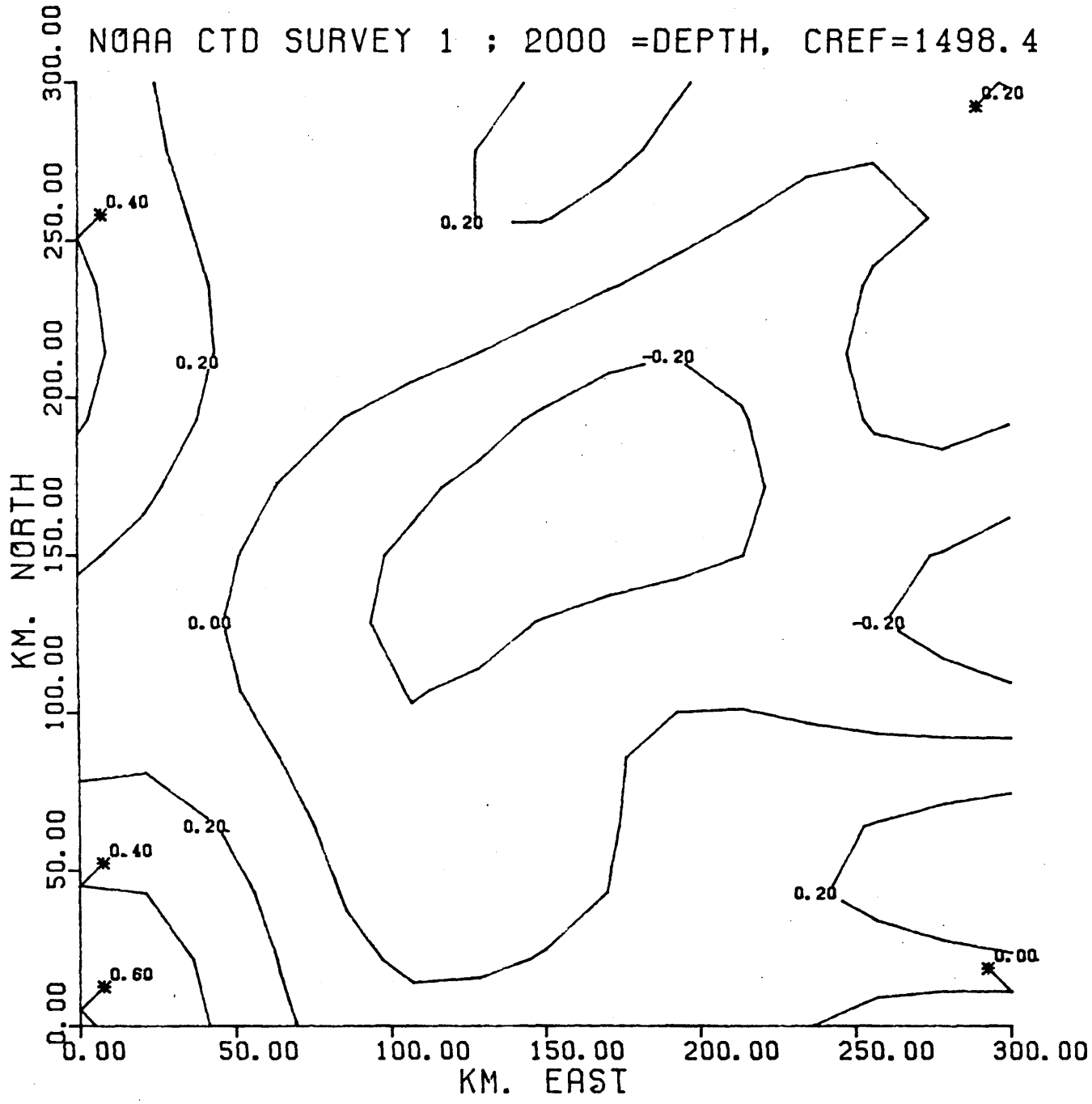
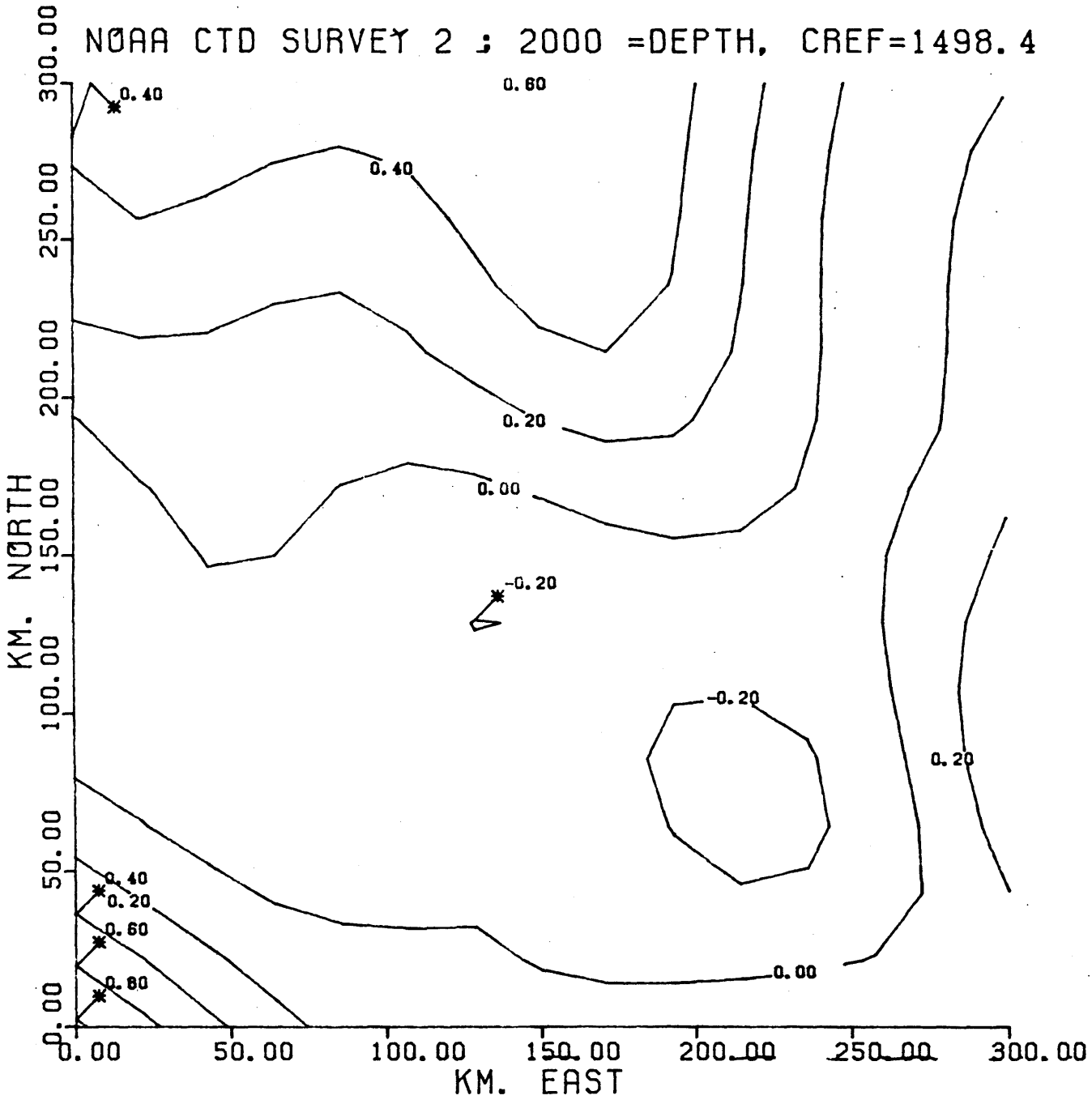


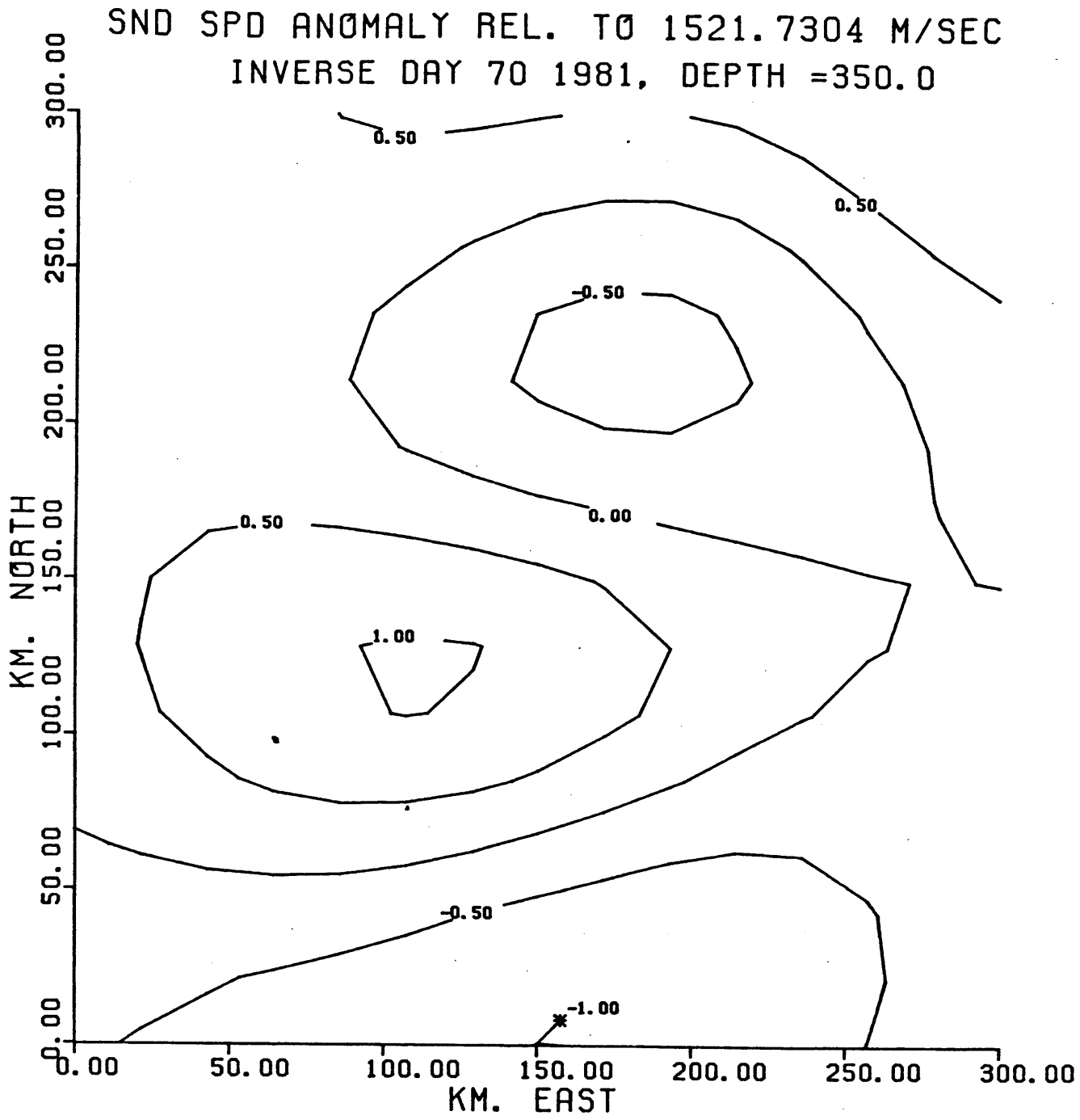
FIGURE 10.1 I SOUND SPEED ANOMALY AT 2000 M. 2ND NOAA CTD SURVEY.

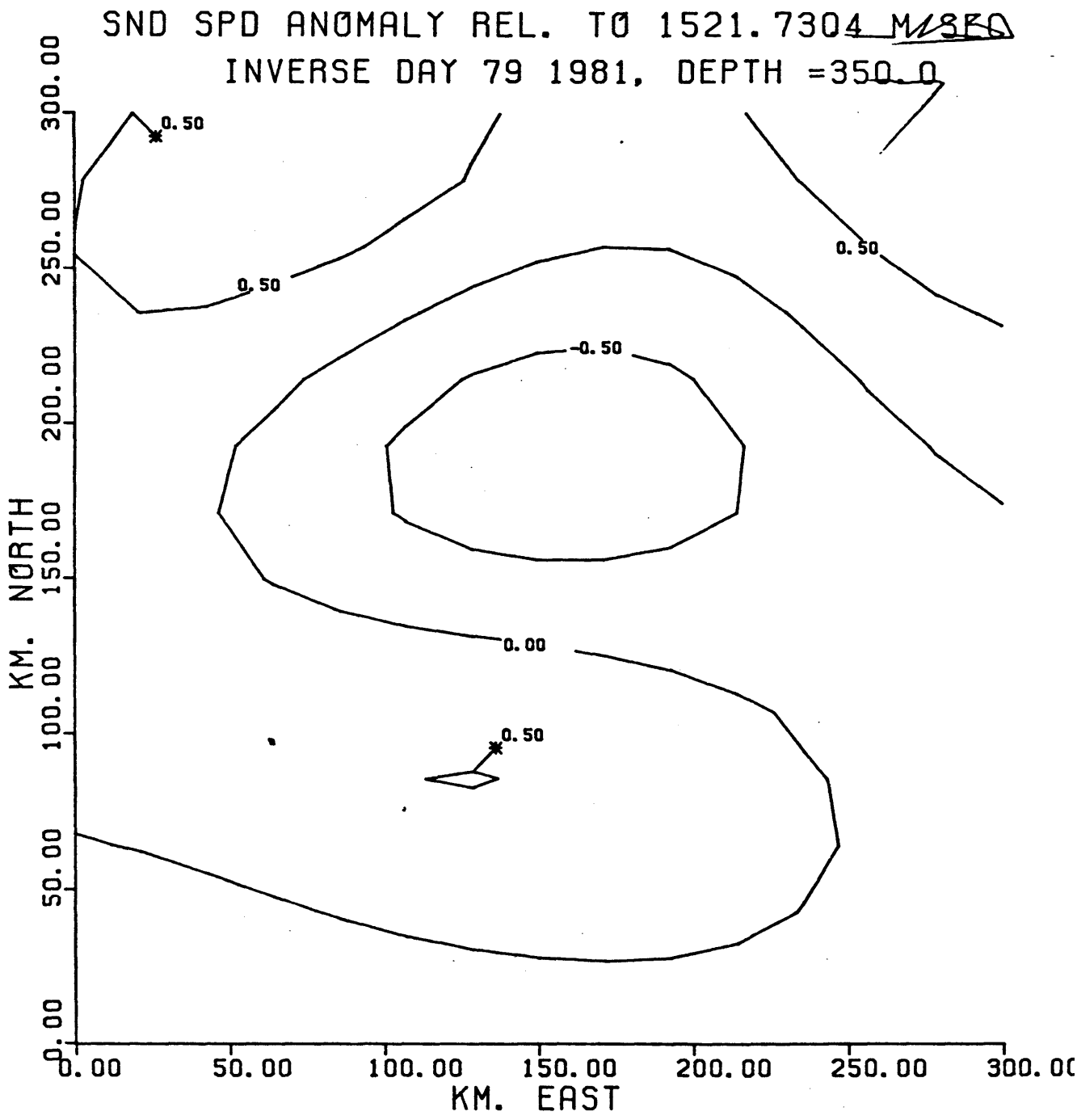


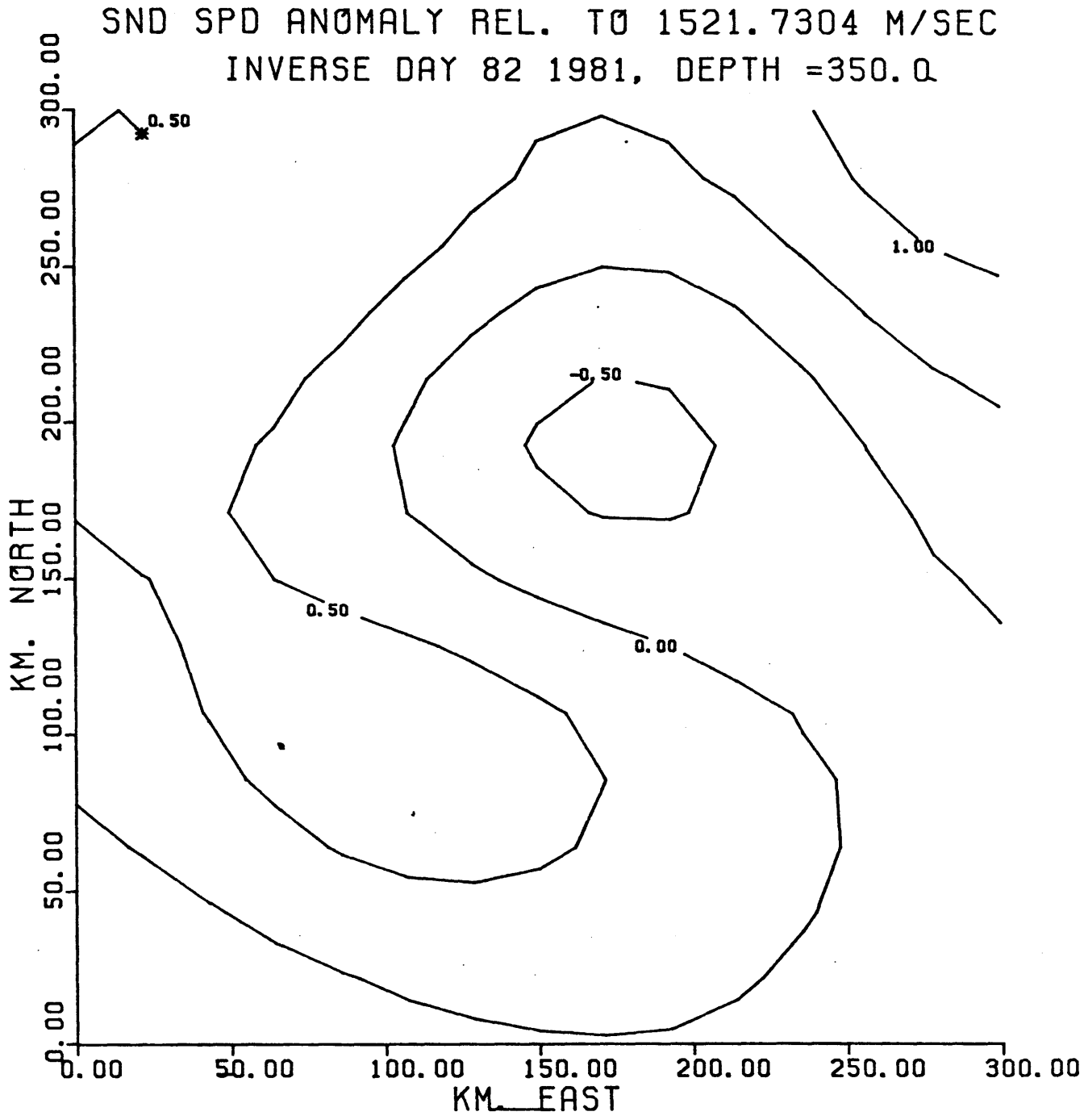
inversions require the first CTD survey as an initialization, and have not been used beyond day 106, so they cannot be directly compared with either CTD survey. The corrections are only complete on a few days, and so the maps cannot be displayed as a time series. Because of the heightened error level in the day differential data resulting from the subtractions, the resolution of these maps is low, and the initialization using the CTD survey tends to dominate the map. Finally, the simple corrections for line-of-sight range changes introduce errors of order 5 msec. For these reasons, it is better to compare the traditional data with the estimates of sound speed made using uncorrected data.

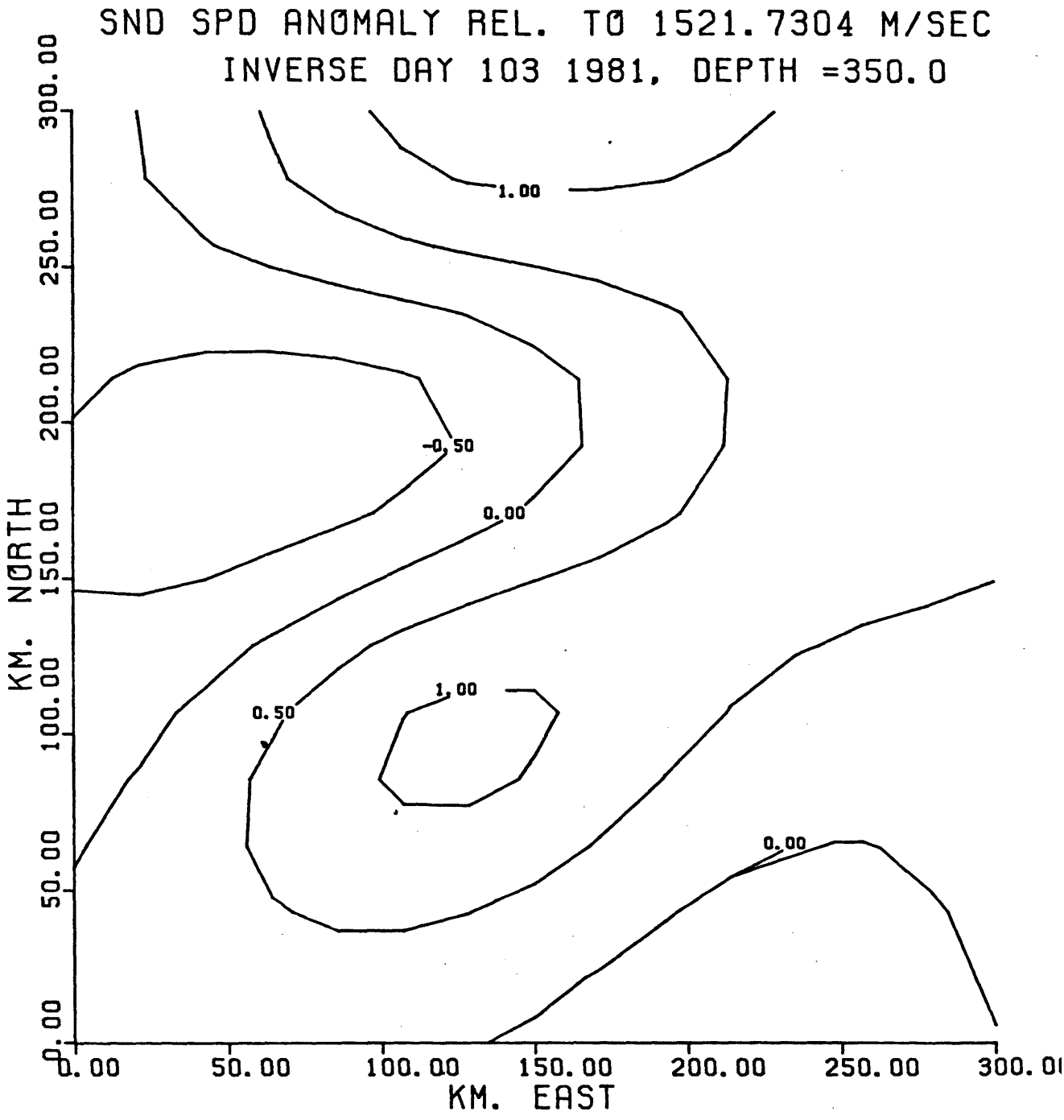
Figure 9.6 shows time series of sound speed anomaly field estimates at 700 meters depth, Figures 10.2, 10.3, and 10.4 show maps for 350, 1500, and 2000 meters, respectively. The continuous nature of the inverse means that maps could be produced for any level, but that might become somewhat tedious. Only a few of the rays used at present penetrate to within 300 meters of the surface, so the resolving power of the estimator decreases with decreasing depth (see Figure 10.5). The perturbations due to mesoscale dynamics presumably have structures similar to the calculated first and second baroclinic modes, (see

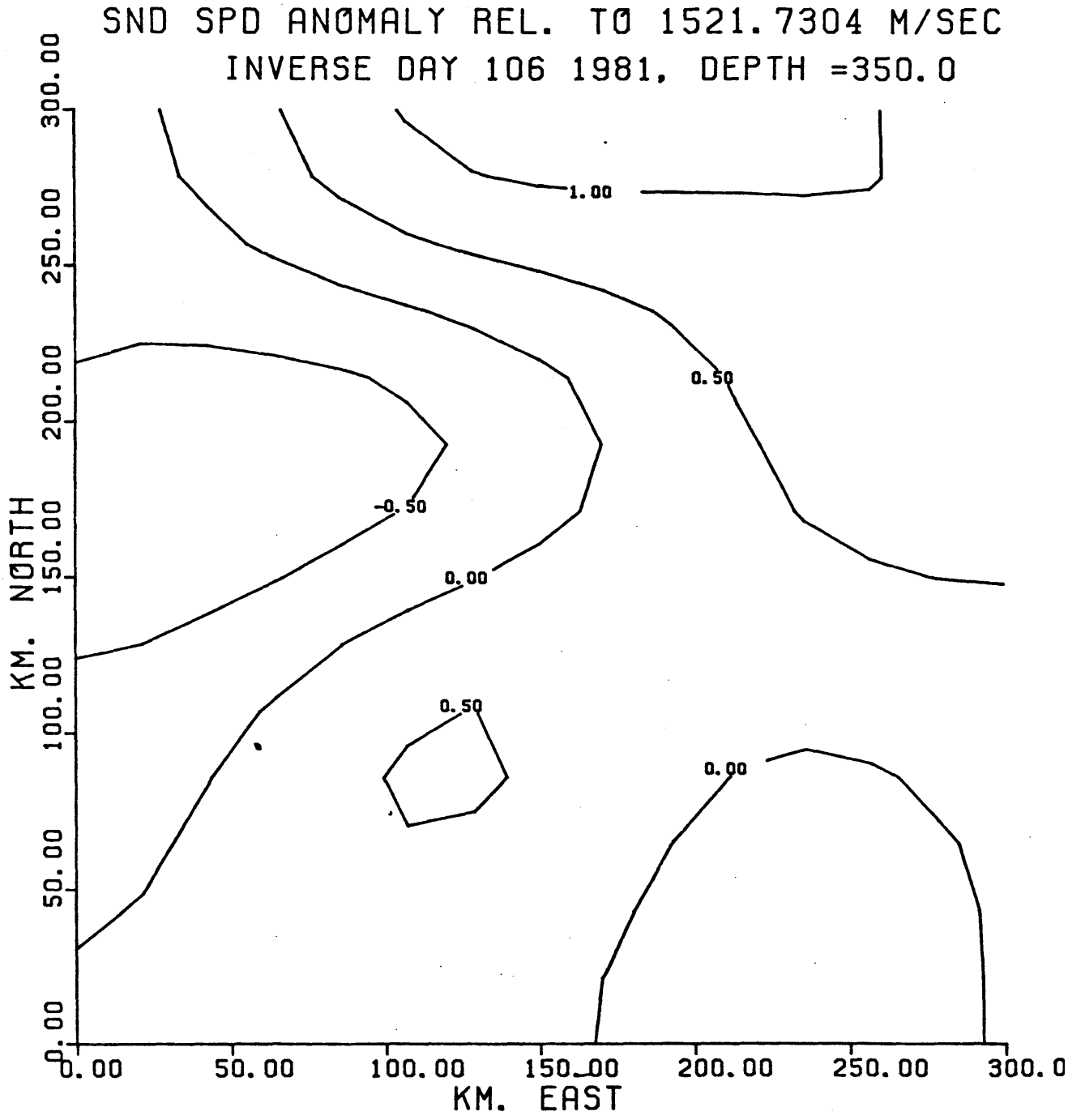
FIGURE 10.2 A-F MAPS OF SOUND SPEED ANOMALY AT 350 METERS ESTIMATED BY THE ACOUSTIC INVERSE. C.I. = 0.5 M/SEC.

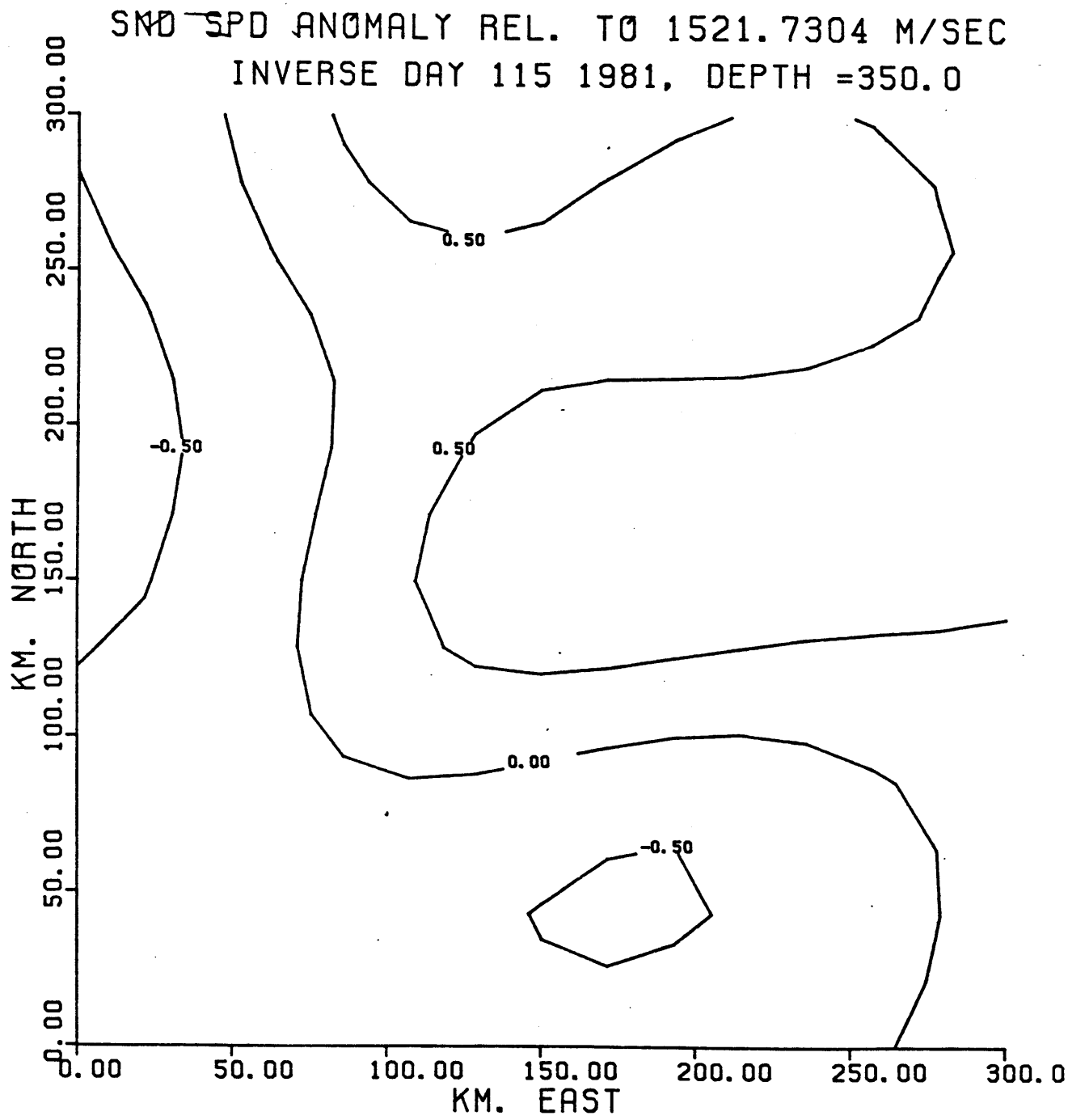



SND SPD ANOMALY REL. TO 1521.7304 M/SEC
INVERSE DAY 79 1981, DEPTH =350.0
0.50
0.50
0.50
-0.50
0.00
0.50
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1521.7304 M/SEC
INVERSE DAY 82 1981, DEPTH =350.0
0.50
1.00
-0.50
0.50
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1521.7304 M/SEC
INVERSE DAY 103 1981, DEPTH =350.0
1.00
-0.50
0.00
1.00
0.50
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1521.7304 M/SEC
INVERSE DAY 106 1981, DEPTH =350.0
1.00
0.50
-0.50
0.00
0.50
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST


SMD SPD ANOMALY REL. TO 1521.7304 M/SEC
INVERSE DAY 115 1981, DEPTH =350.0
0.50
-0.50
0.50
0.00
-0.50
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.0
KM. EAST

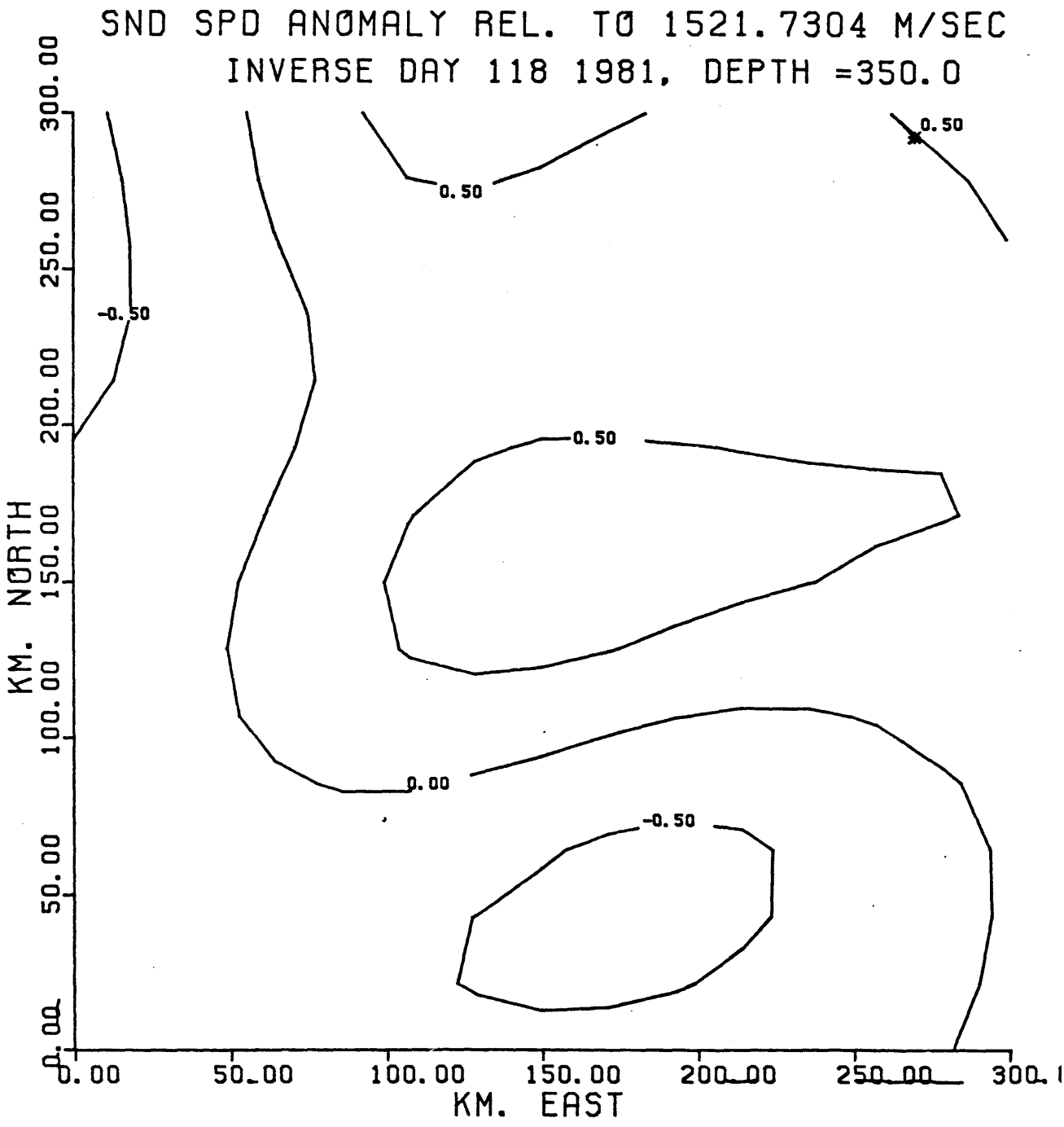
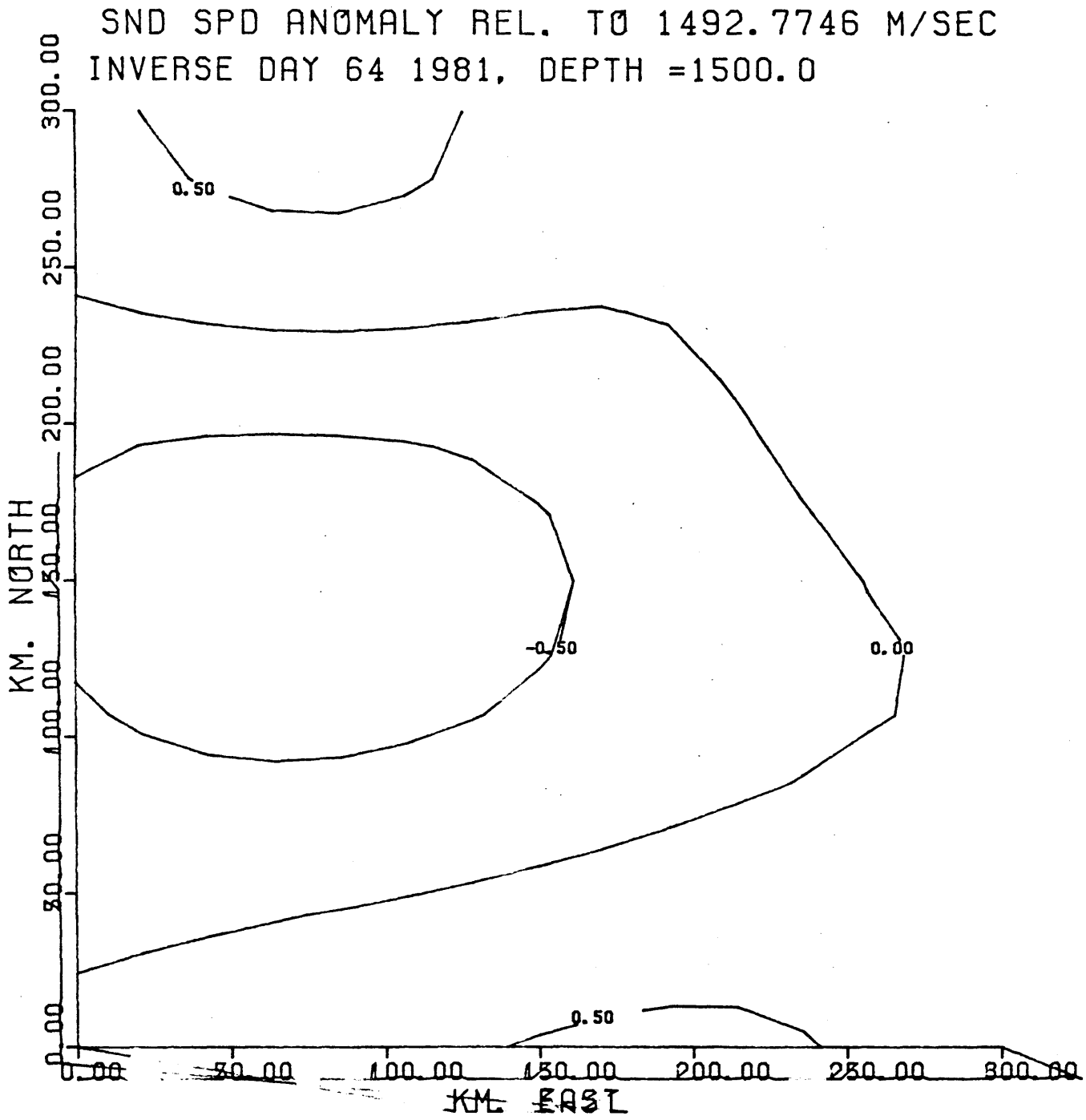
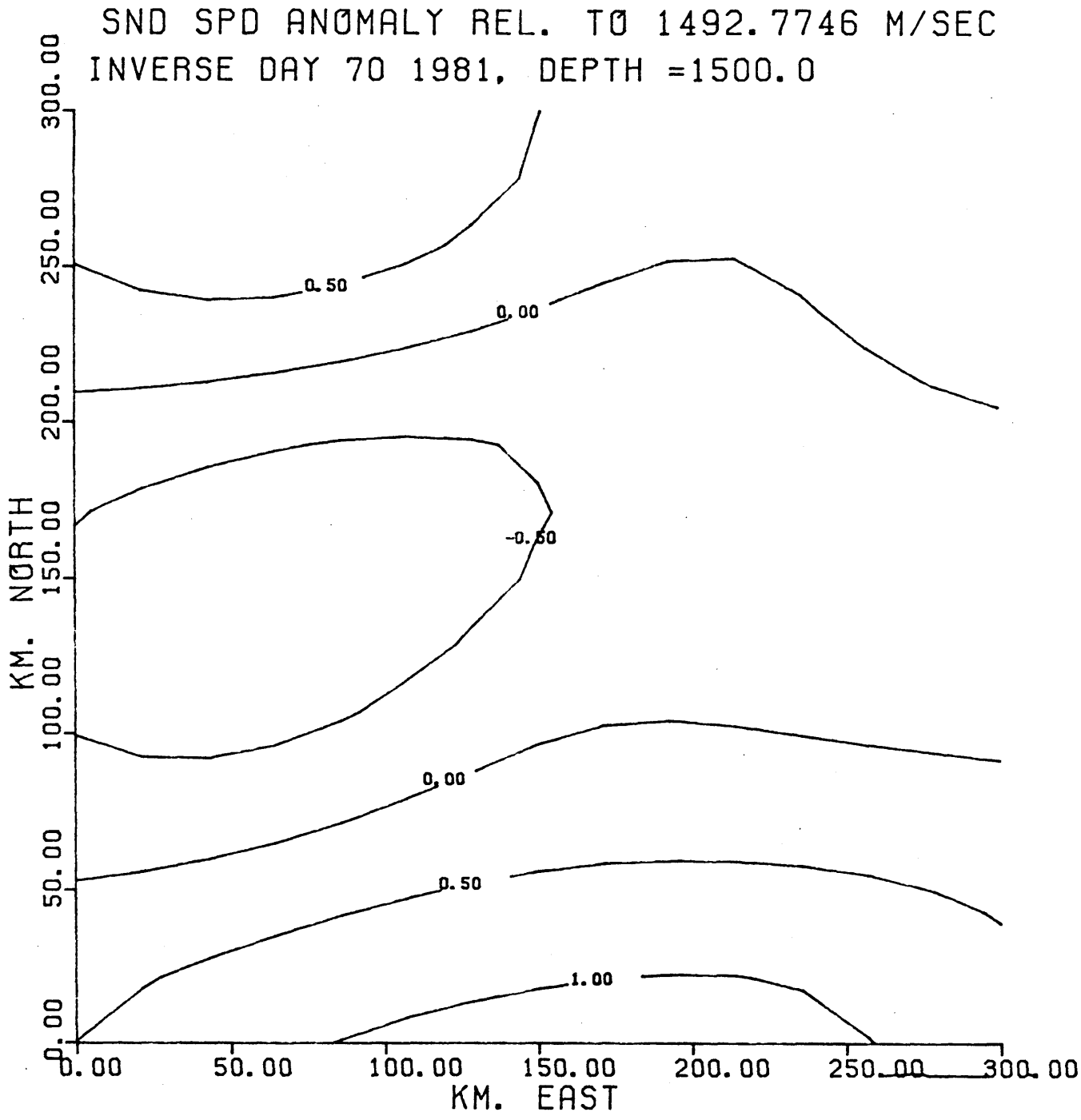
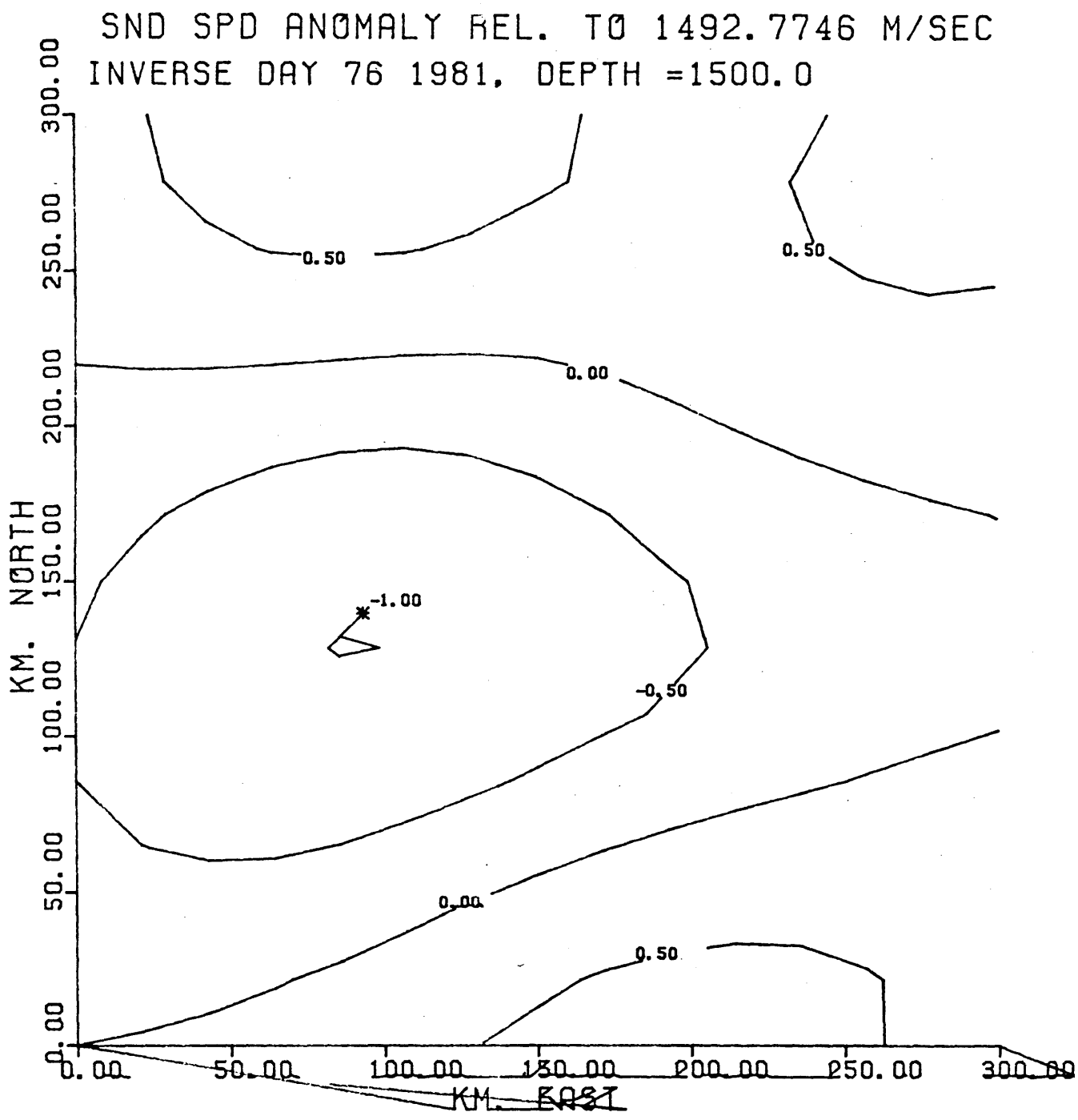
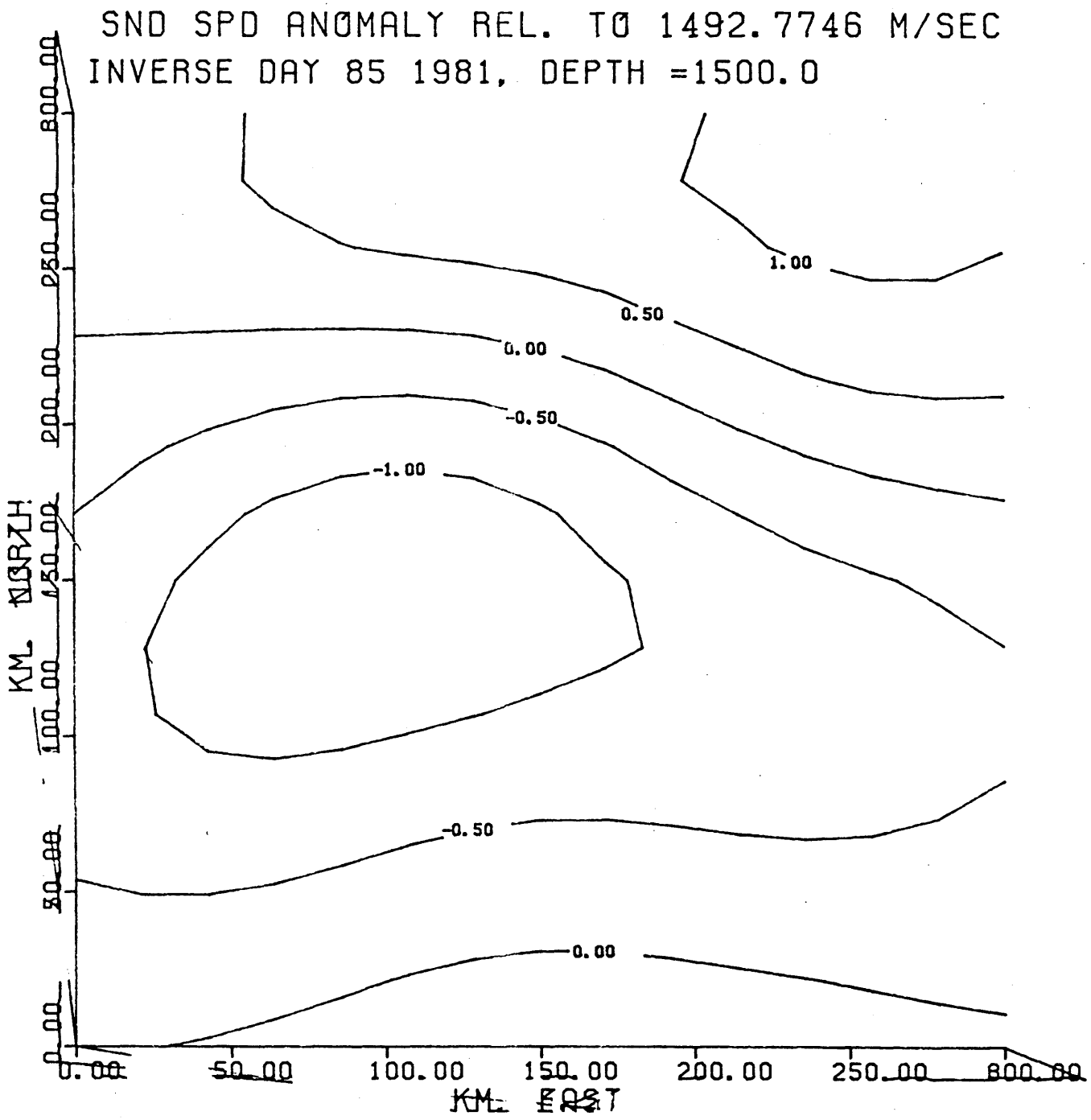

SND SPD ANOMALY REL. TO 1521.7304 M/SEC
INVERSE DAY 118 1981, DEPTH =350.0
0.50
0.50
0.50
-0.50
0.00
-0.50
KM. NORTH
300.00
250.00
200.00
150.00
100.00
50.00
0.00
0.00
50.00
100.00
150.00
200.00
250.00
300.
KM. EAST

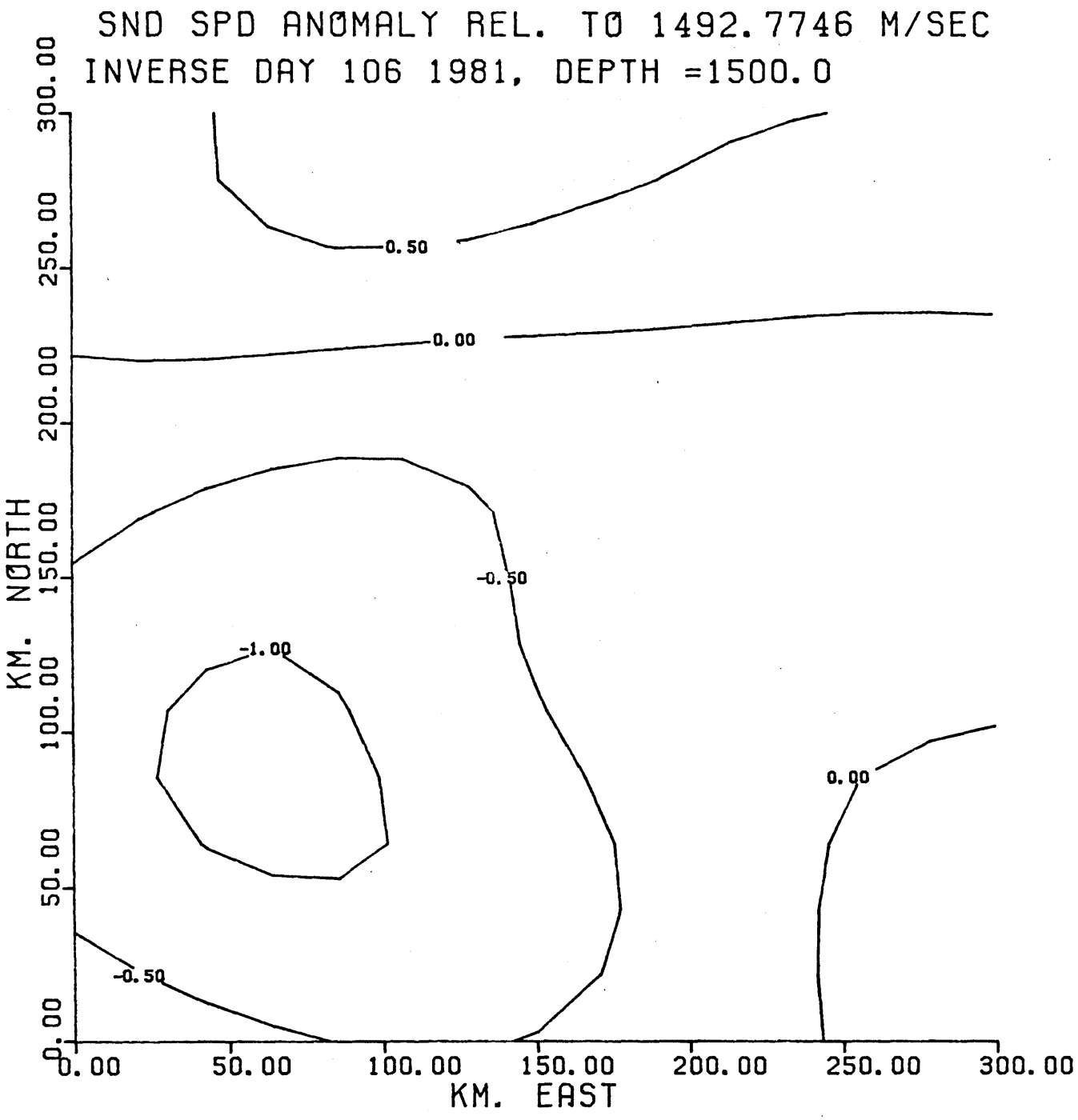
FIGURE 10.3 A-b MAPS OF SOUND SPEED ANOMALY AT 1500 METERS ESTIMATED BY THE ACOUSTIC INVERSE. C.I. = 0.5 M/SEC.

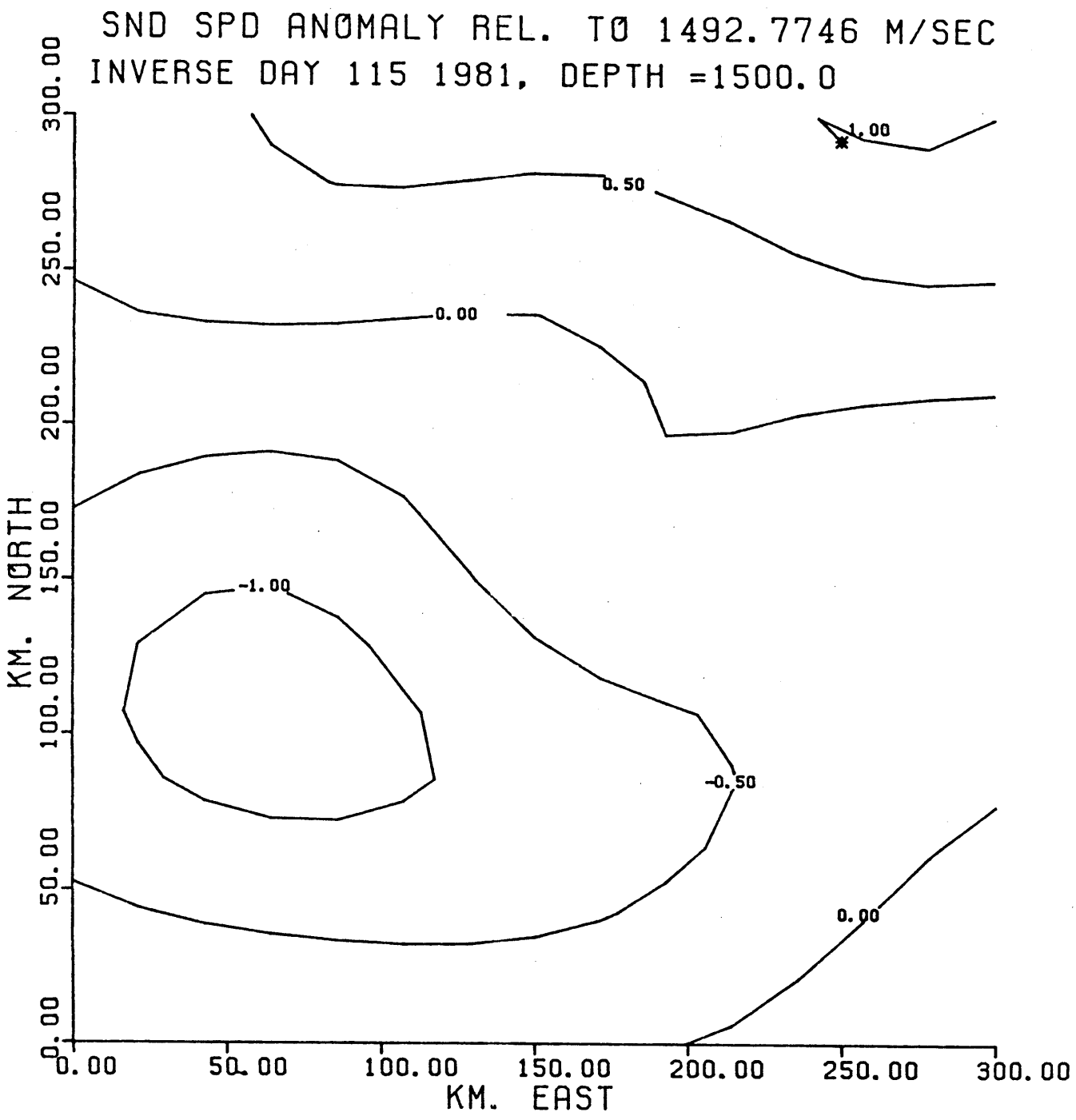



SND SPD ANOMALY REL. TO 1492.7746 M/SEC
INVERSE DAY 70 1981, DEPTH =1500.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.50
0.00
-0.50
0.00
0.50
1.00


SND SPD ANOMALY REL. TO 1492.7746 M/SEC
INVERSE DAY 76 1981, DEPTH =1500.0
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.50
0.50
0.00
-1.00
-0.50
0.00
0.50


SND SPD ANOMALY REL. TO 1492.7746 M/SEC
INVERSE DAY 85 1981, DEPTH =1500.0
1.00
0.50
0.00
-0.50
-1.00
-0.50
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
800.00
KM. EAST


SND SPD ANOMALY REL. TO 1492.7746 M/SEC
INVERSE DAY 106 1981, DEPTH =1500.0
0.50
0.00
-0.50
-1.00
0.00
-0.50
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1492.7746 M/SEC
INVERSE DAY 115 1981, DEPTH =1500.0
1.00
0.50
0.00
-1.00
-0.50
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00

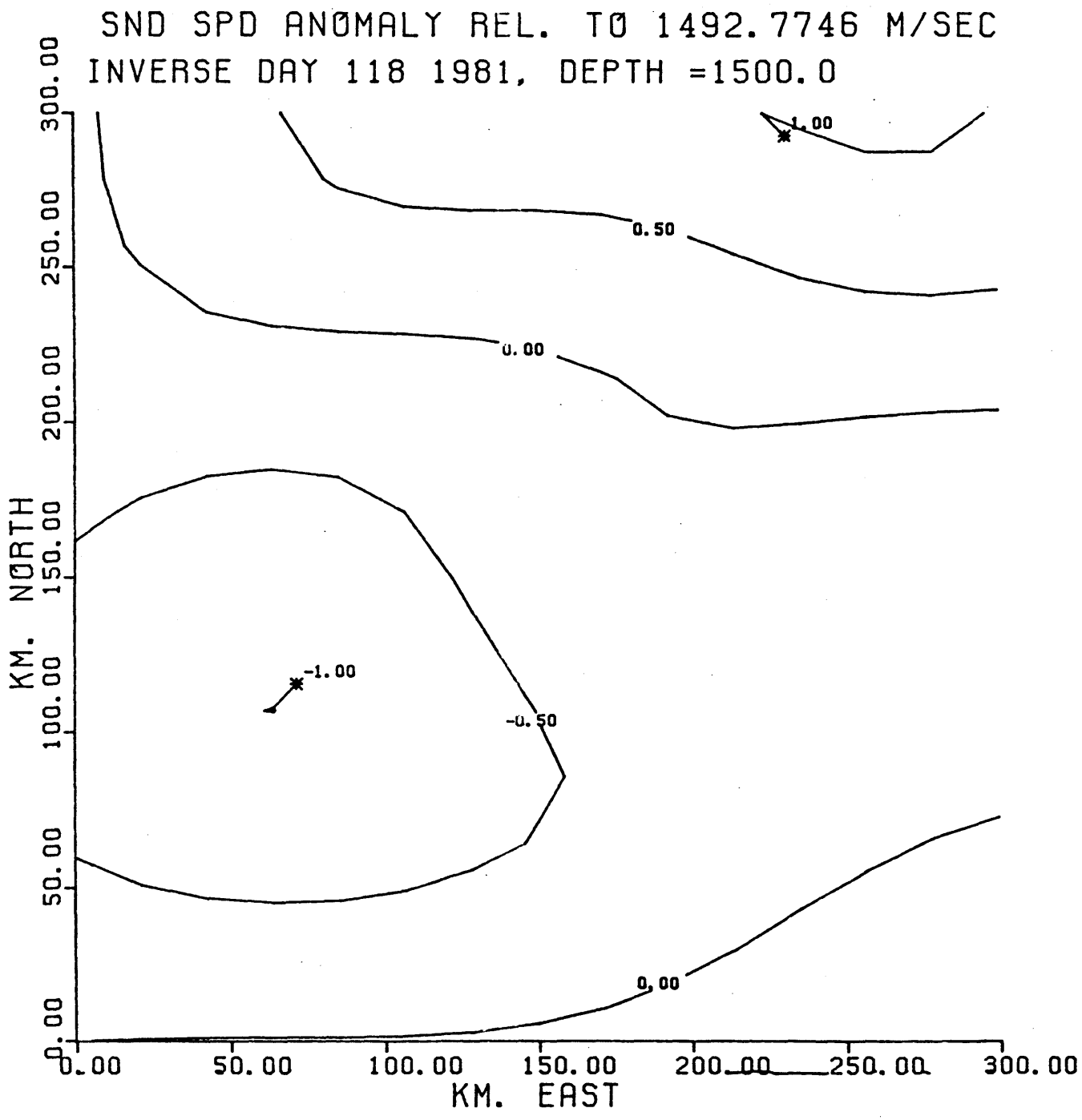
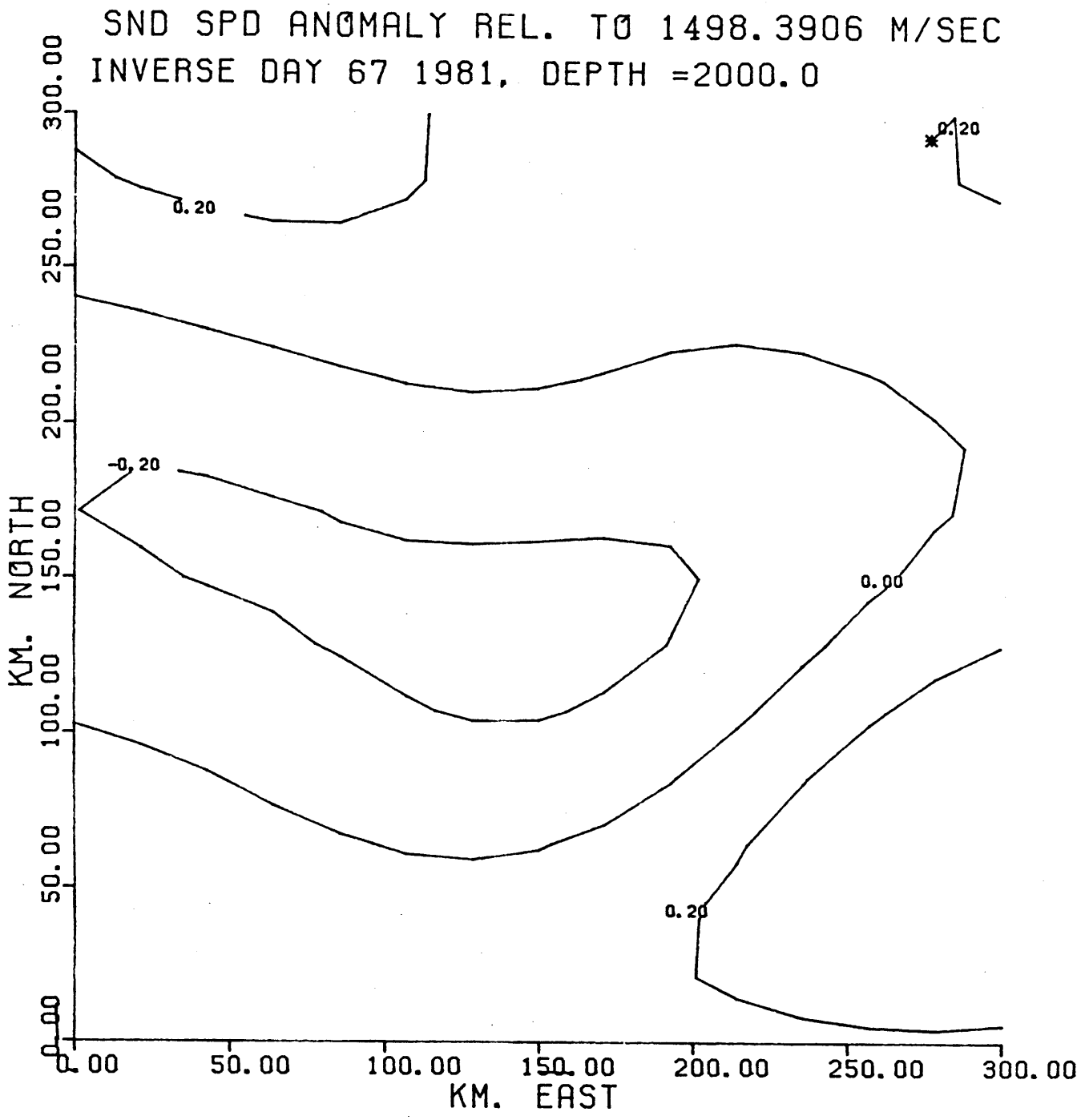
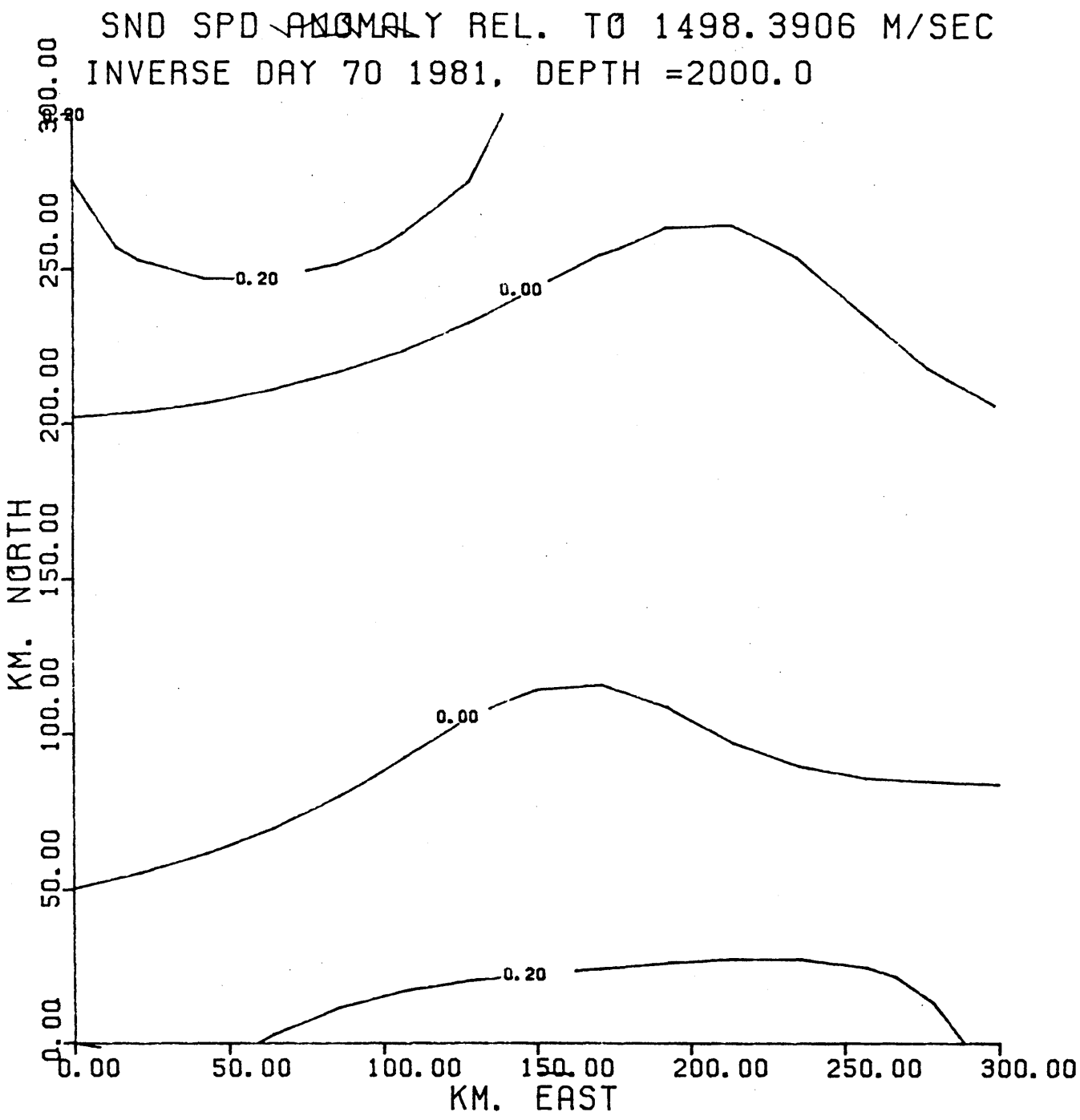
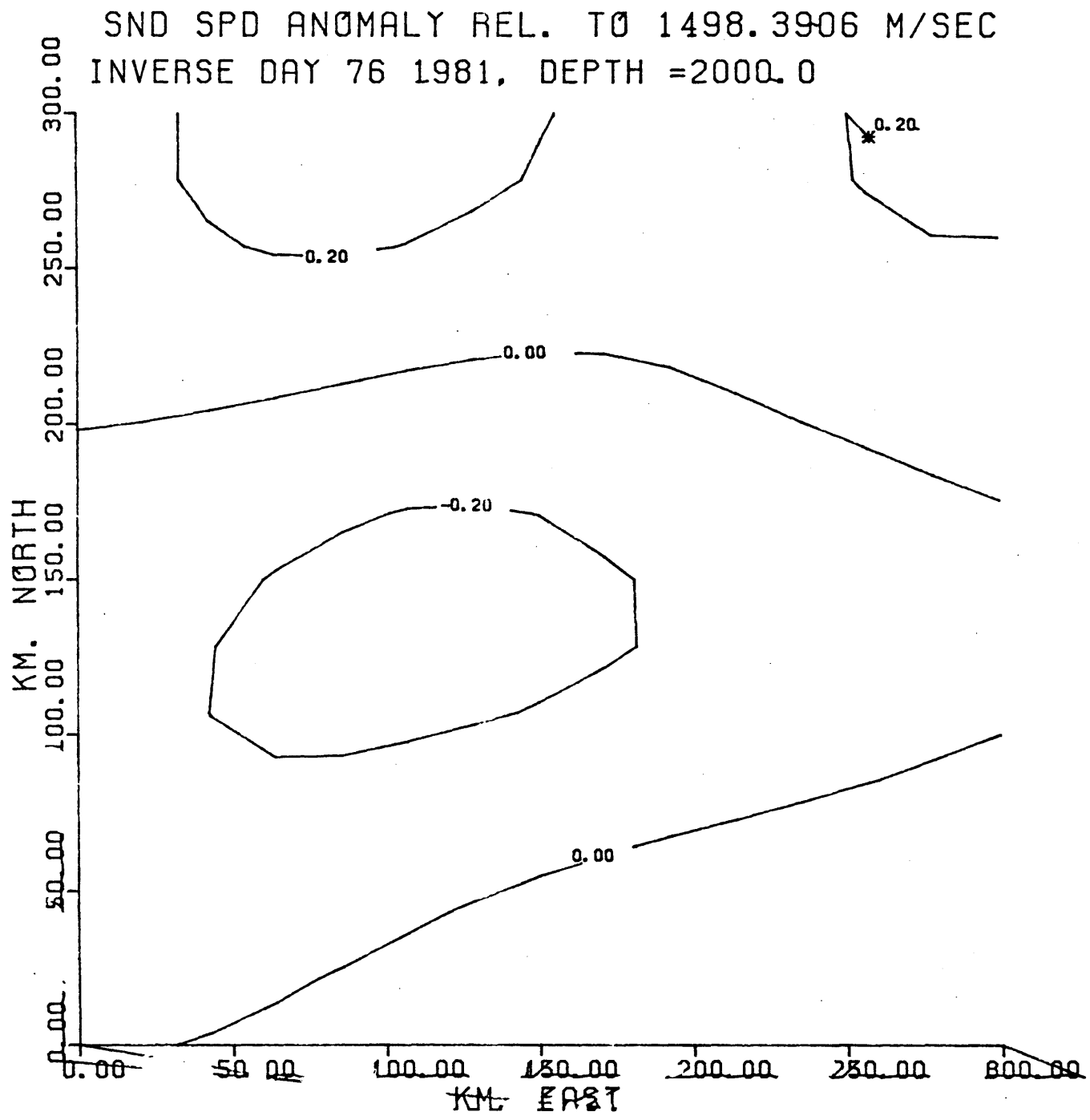
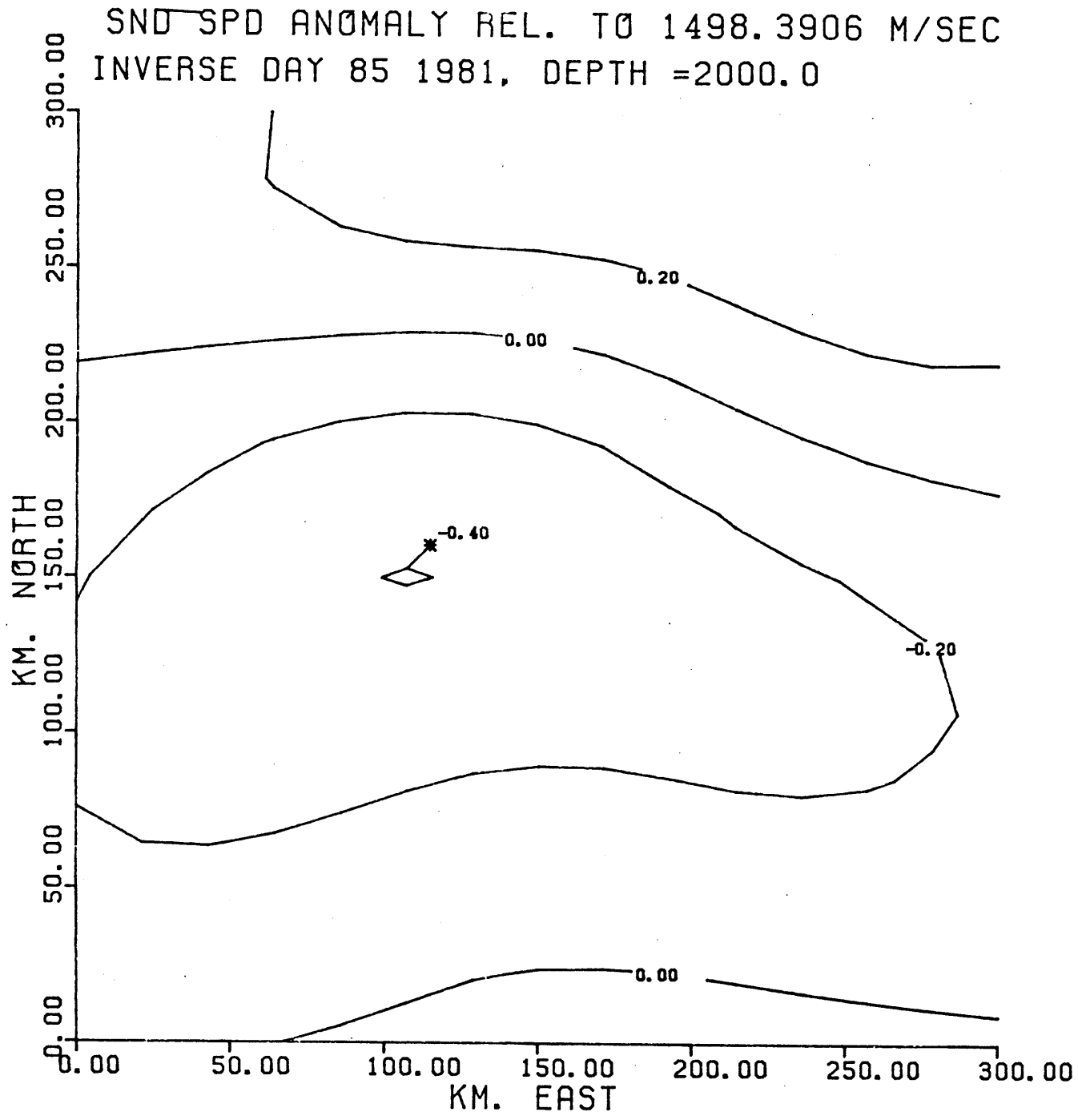

SND SPD ANOMALY REL. TO 1492.7746 M/SEC
INVERSE DAY 118 1981, DEPTH =1500.0
1.00
0.50
0.00
-1.00
-0.50
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST

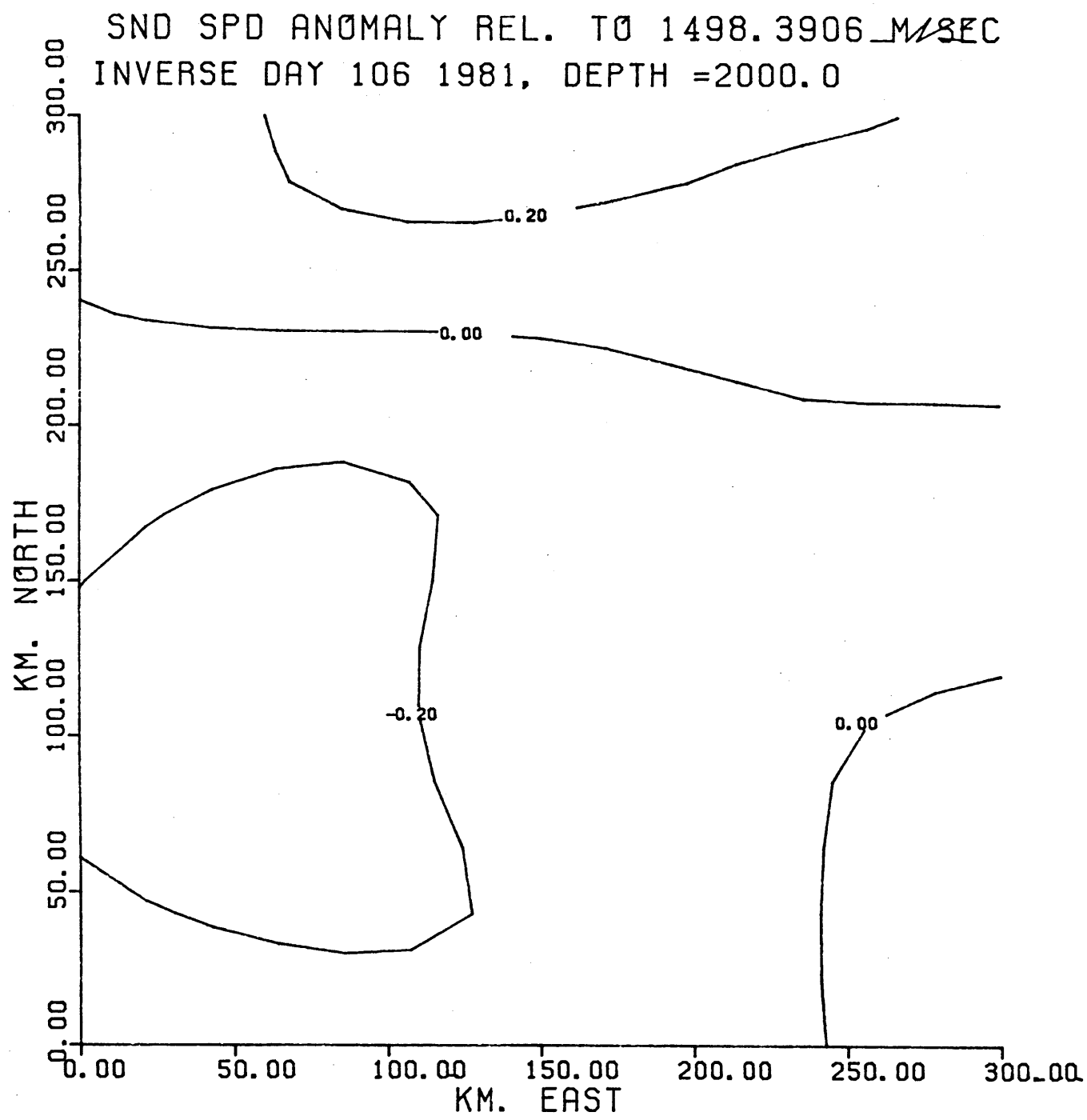
FIGURE 10.4 A-6 MAPS OF SOUND SPEED ANOMALY AT 2000 METERS ESTIMATED BY THE ACOUSTIC INVERSE. C.I. = 0.2 M/SEC.

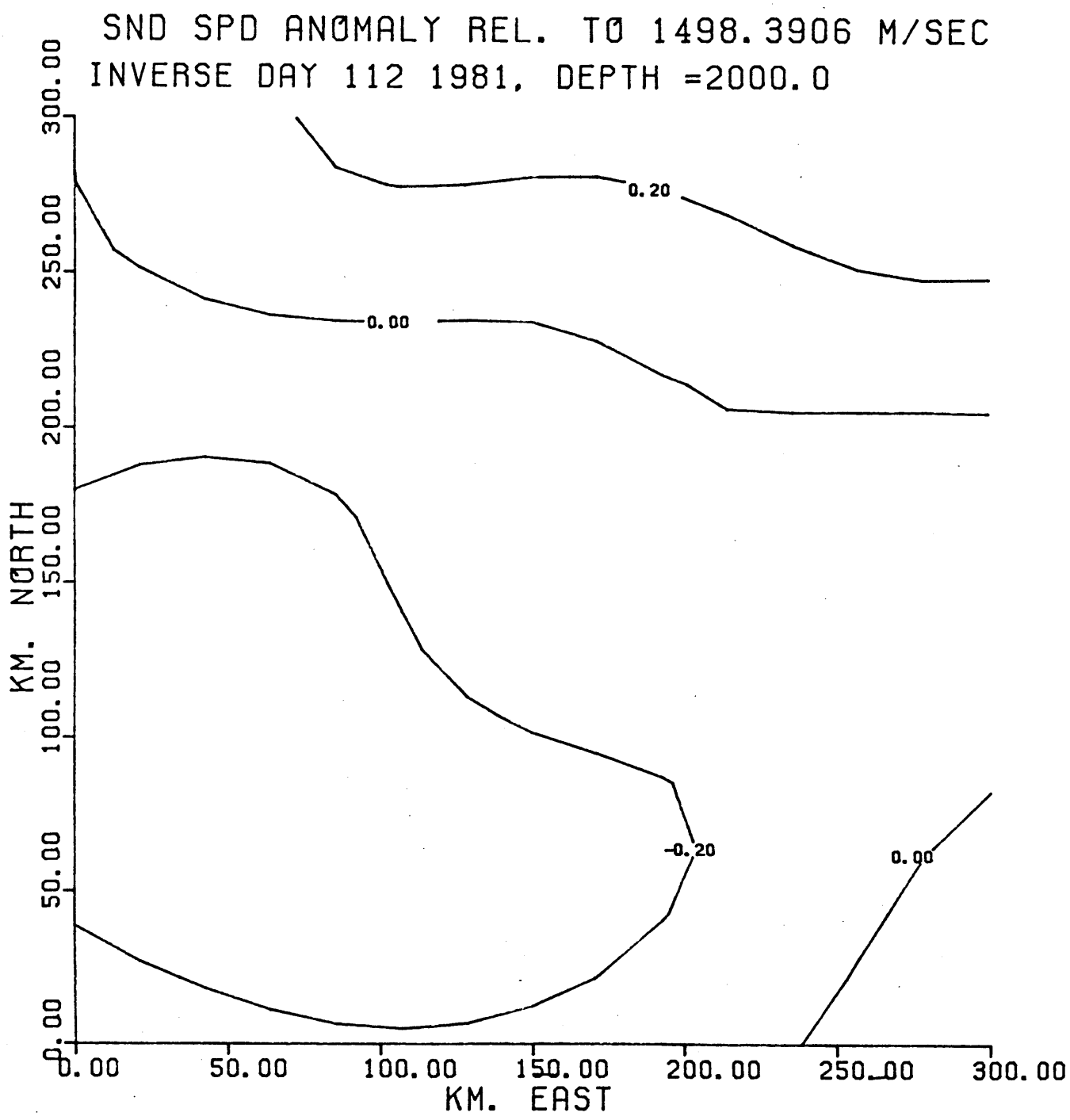


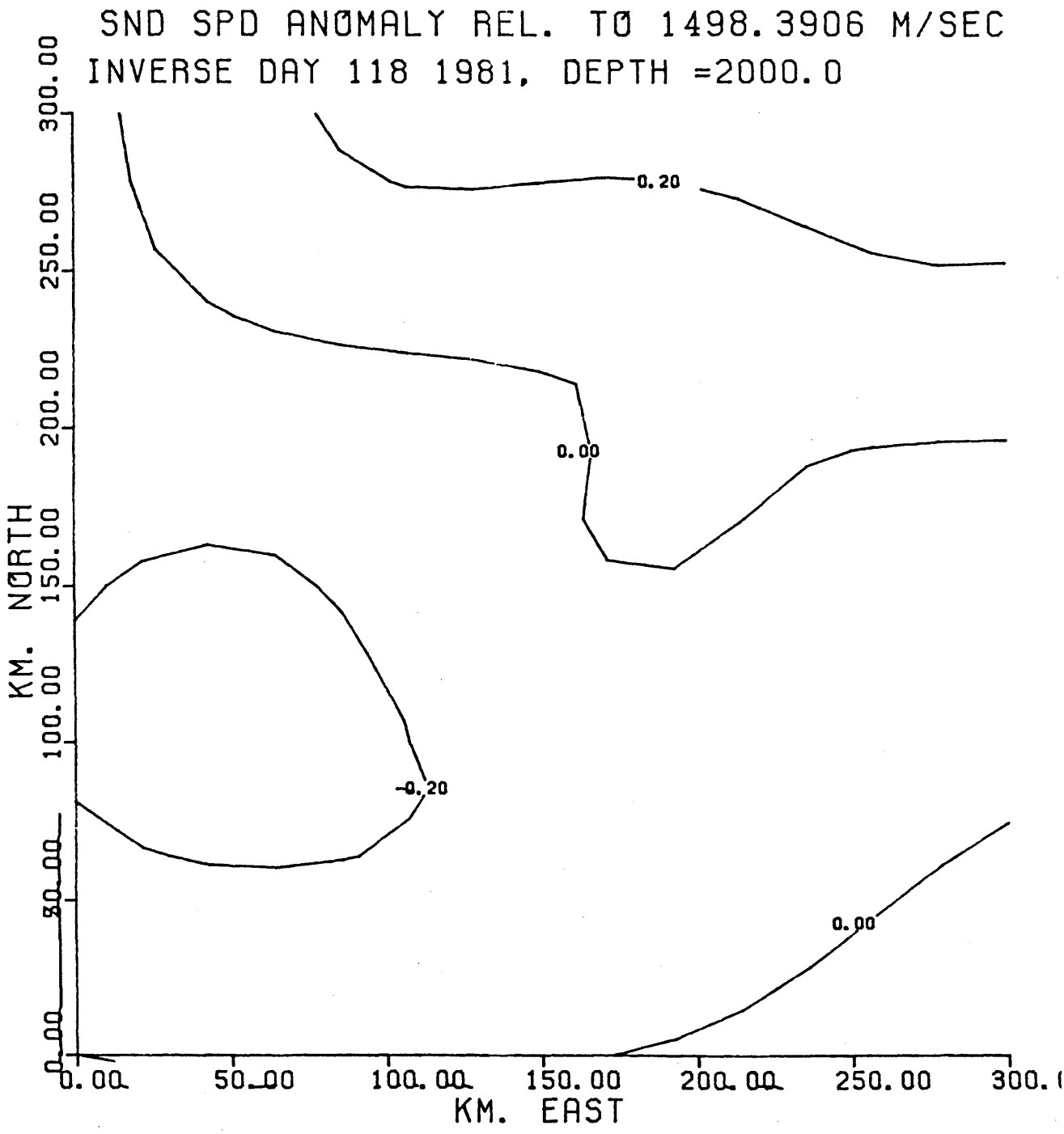

SND SPD ANOMALY REL. TO 1498.3906 M/SEC
INVERSE DAY 70 1981, DEPTH =2000.0
0.20
0.00
0.00
0.20
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00


SND SPD ANOMALY REL. TO 1498.3906 M/SEC
INVERSE DAY 76 1981, DEPTH =2000.0
0.20
0.20
0.00
-0.20
0.00
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
800.00


SND SPD ANOMALY REL. TO 1498.3906 M/SEC
INVERSE DAY 85 1981. DEPTH =2000.0
0.20
0.00
-0.40
-0.20
0.00
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00


SND SPD ANOMALY REL. TO 1498.3906 M/SEC
INVERSE DAY 106 1981, DEPTH =2000.0
0.20
0.00
-0.20
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1498.3906 M/SEC
INVERSE DAY 112 1981, DEPTH =2000.0
0.20
0.00
-0.20
0.00
KM. NORTH
0.00
50.00
100.00
150.00
200.00
250.00
300.00
0.00
50.00
100.00
150.00
200.00
250.00
300.00
KM. EAST


SND SPD ANOMALY REL. TO 1498.3906 M/SEC
INVERSE DAY 118 1981, DEPTH =2000.0
0.20
0.00
-0.20
0.00
KM. NORTH
KM. EAST
0.00
50.00
100.00
150.00
200.00
250.00
300.00

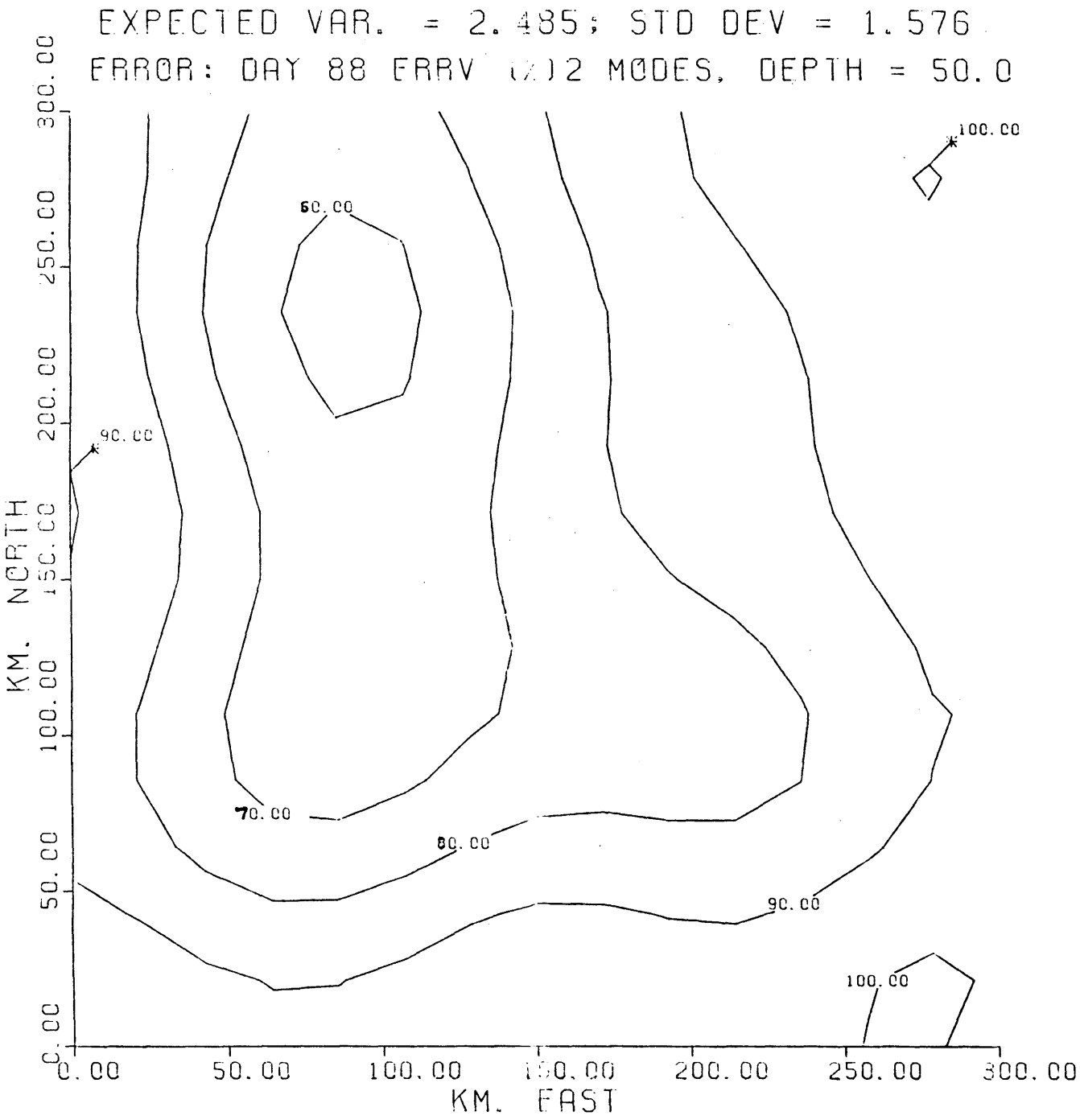


FIGURE 10.5 A: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS PERCENTAGE OF TOTAL VARIANCE. AT THIS LEVEL (50 METERS DEEP) THE MODEL PREDICTS 1.68 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

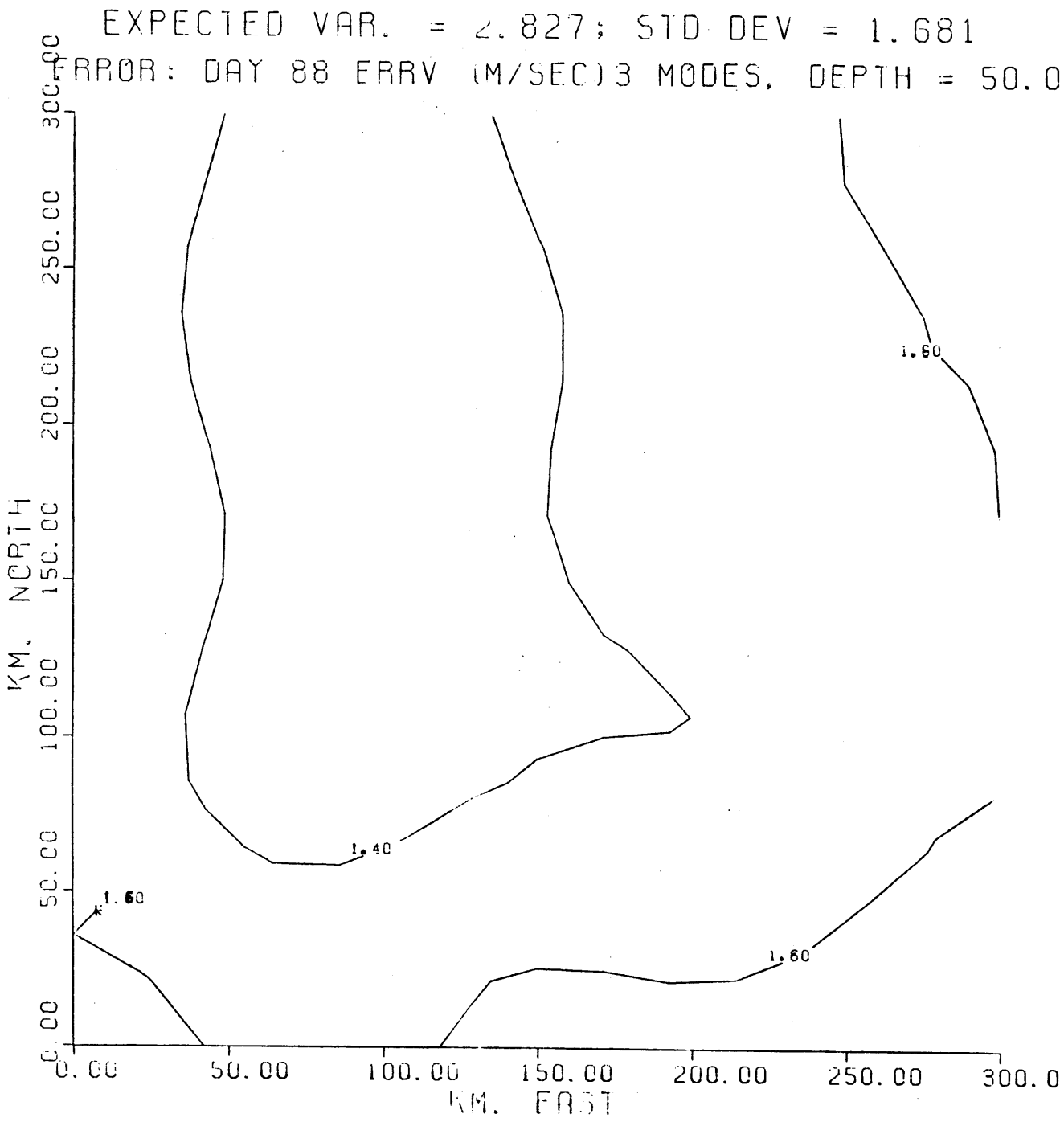


FIGURE 10.5 B: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS SQUARE ROOT OF TOTAL VARIANCE. AT THIS LEVEL (50 METERS DEEP) THE MODEL PREDICTS 1.68 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

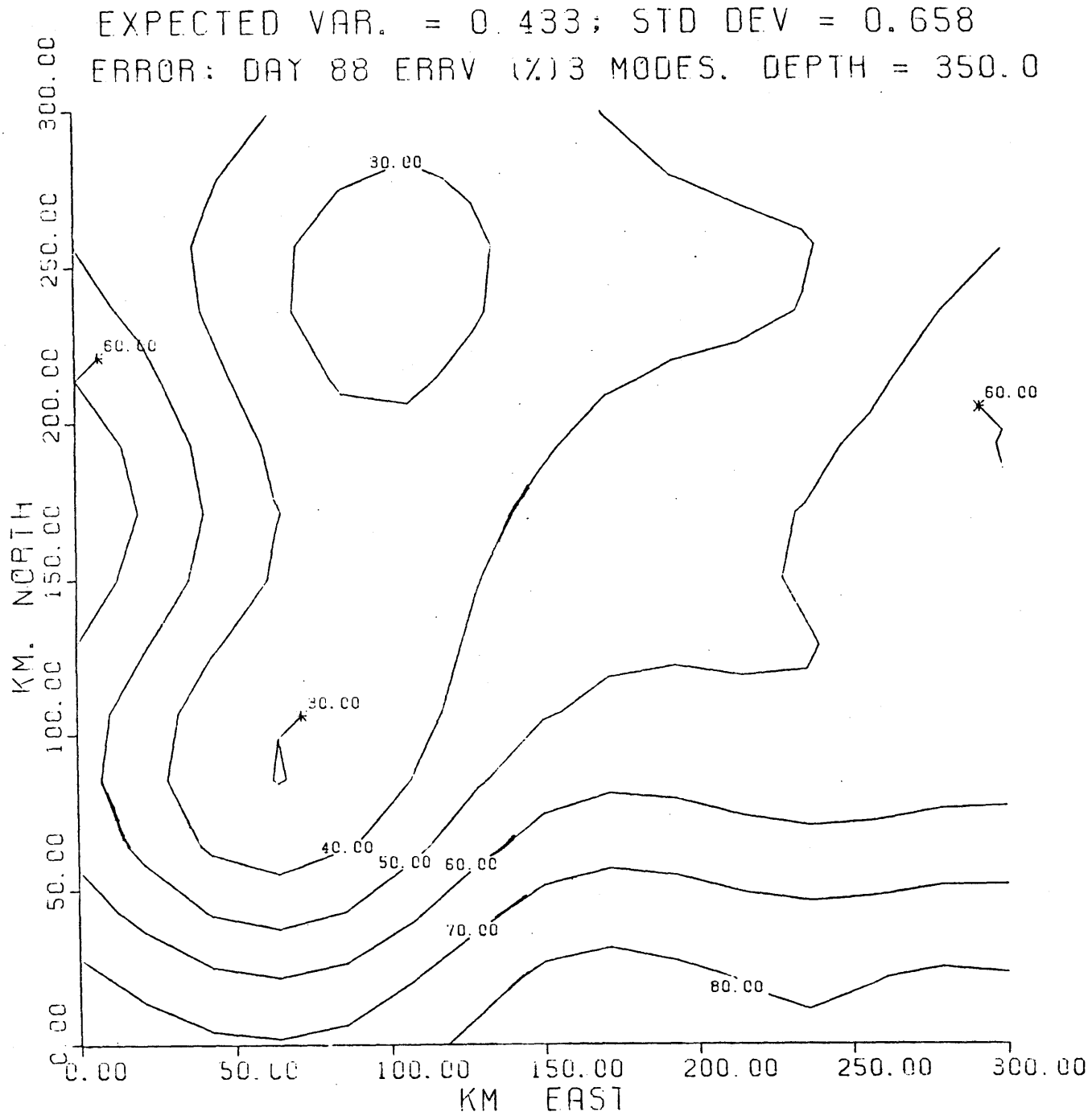


FIGURE 10.5 C: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS PERCENTAGE OF TOTAL VARIANCE. AT THIS LEVEL (350 METERS DEEP) THE MODEL PREDICTS 0.66 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

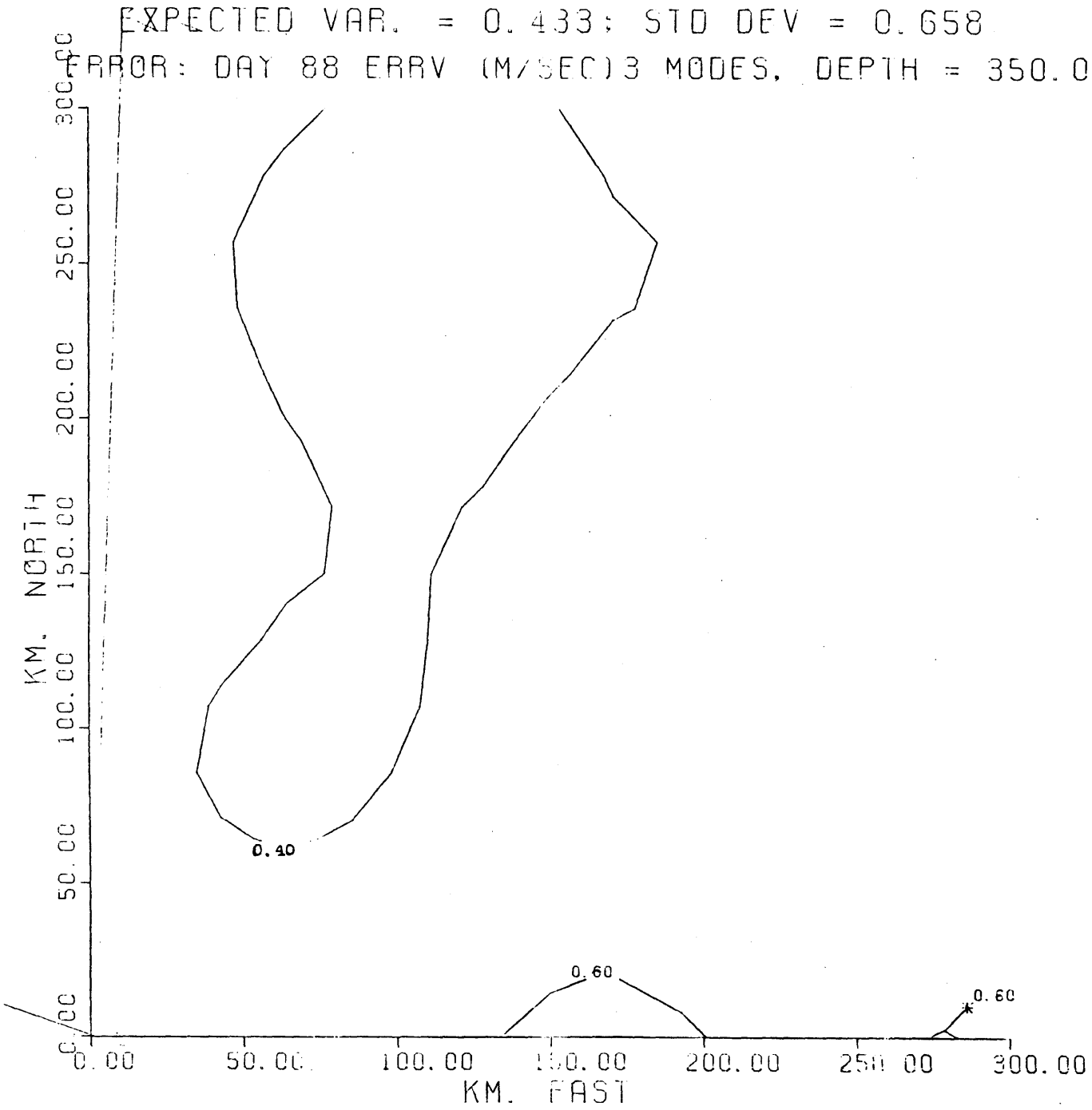


FIGURE 10.5 D: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS SQR. ROOT OF TOTAL VARIANCE. AT THIS LEVEL (350 METERS DEEP) THE MODEL PREDICTS 0.66 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

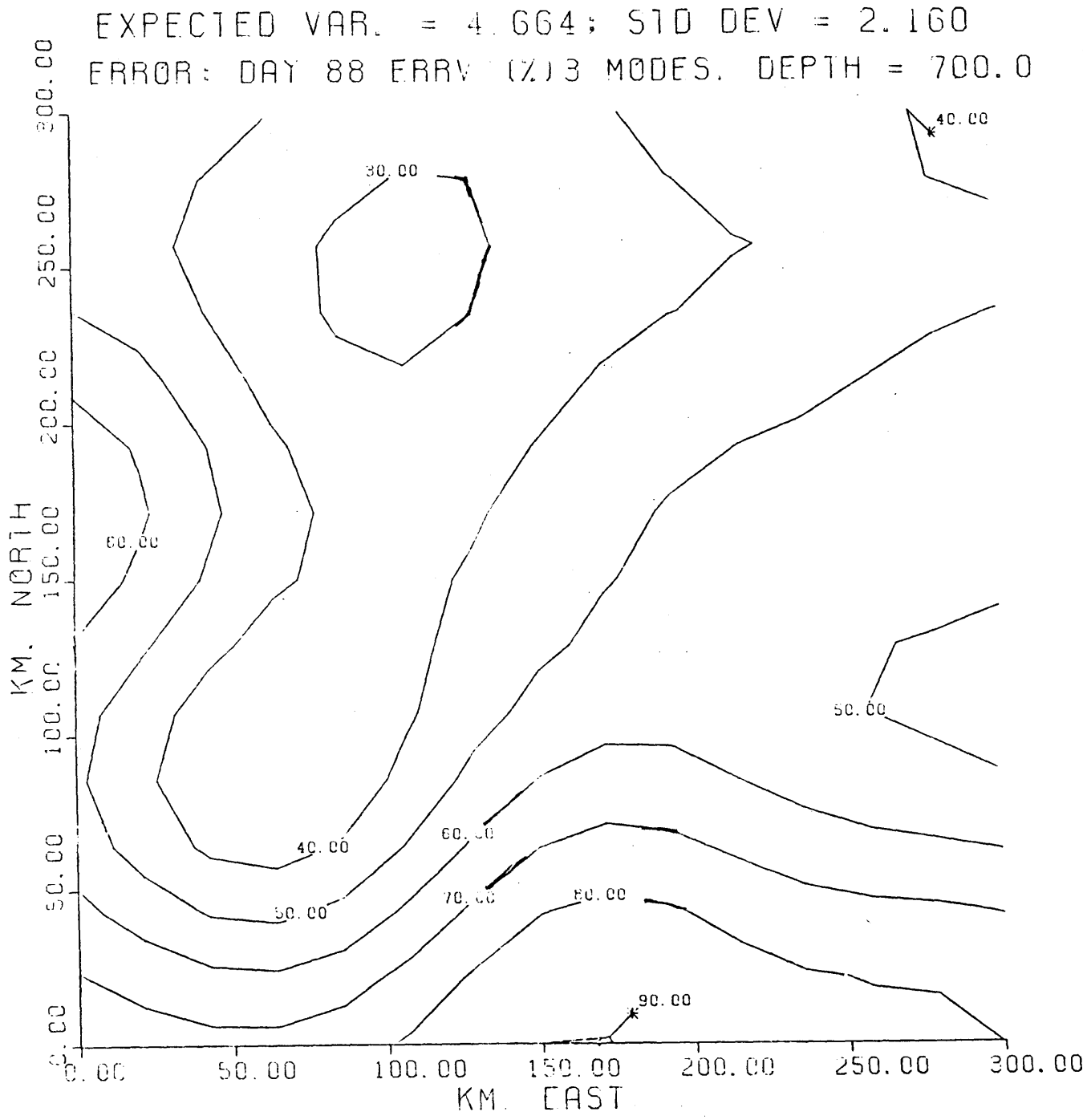


FIGURE 10.5 E: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS PERCENTAGE OF TOTAL VARIANCE. AT THIS LEVEL (700 METERS DEEP) THE MODEL PREDICTS 2.16 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

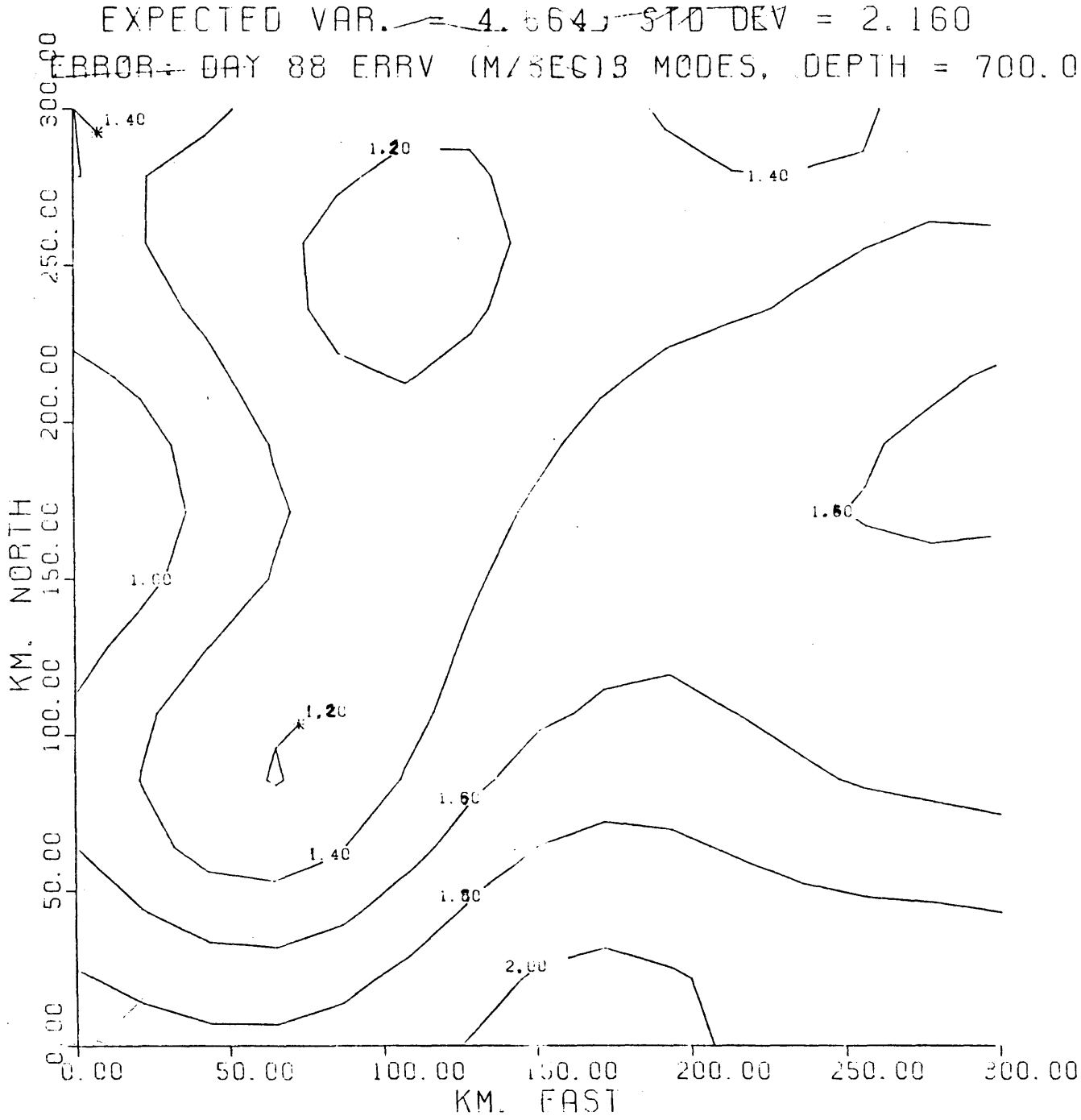


FIGURE 10.5 F: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS SQR. ROOT OF TOTAL VARIANCE. AT THIS LEVEL (700 METERS DEEP) THE MODEL PREDICTS 2.16 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

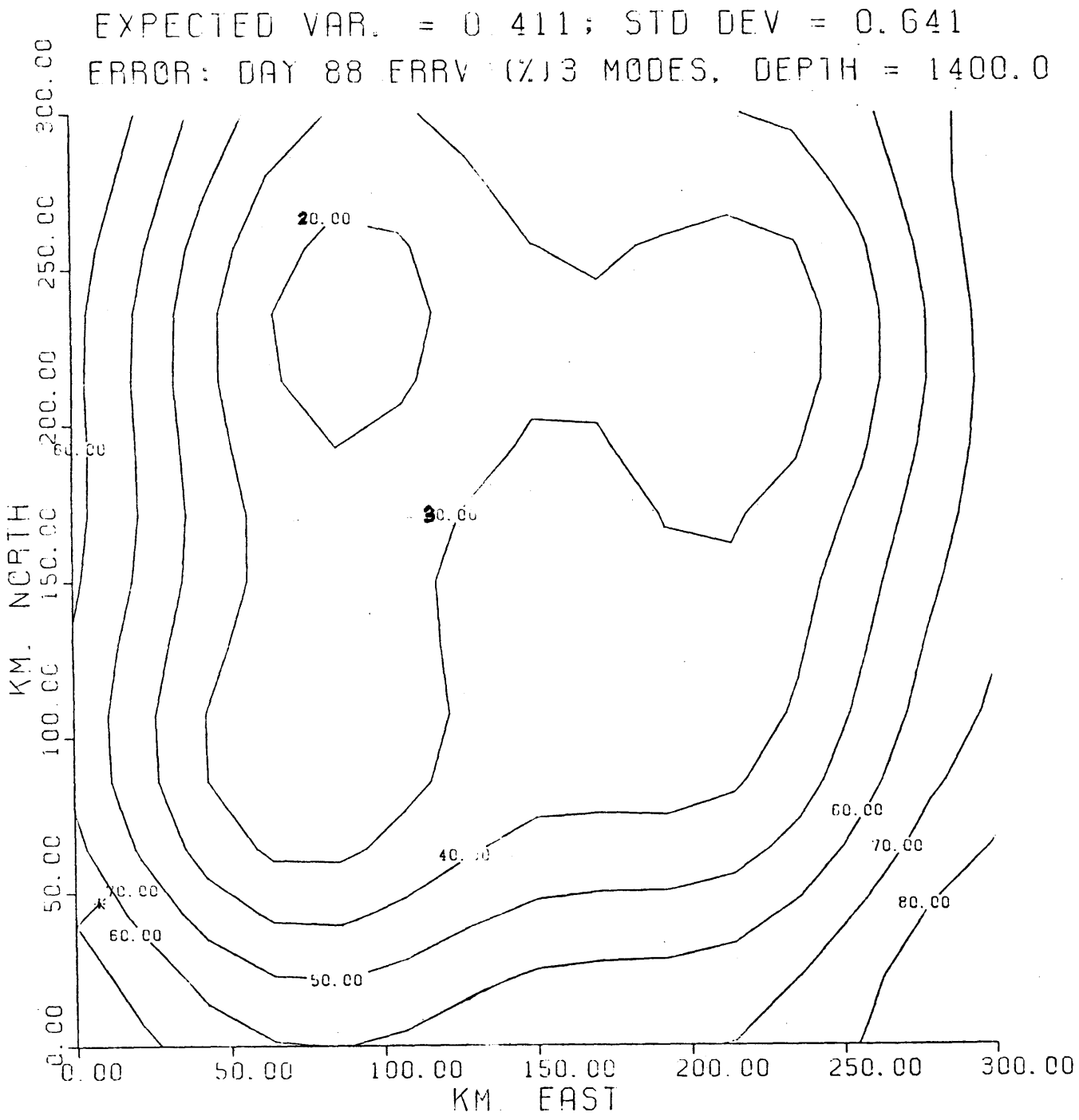


FIGURE 10.5 G: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS PERCENTAGE OF TOTAL VARIANCE. AT THIS LEVEL (1400 METERS DEEP) THE MODEL PREDICTS 0.64 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

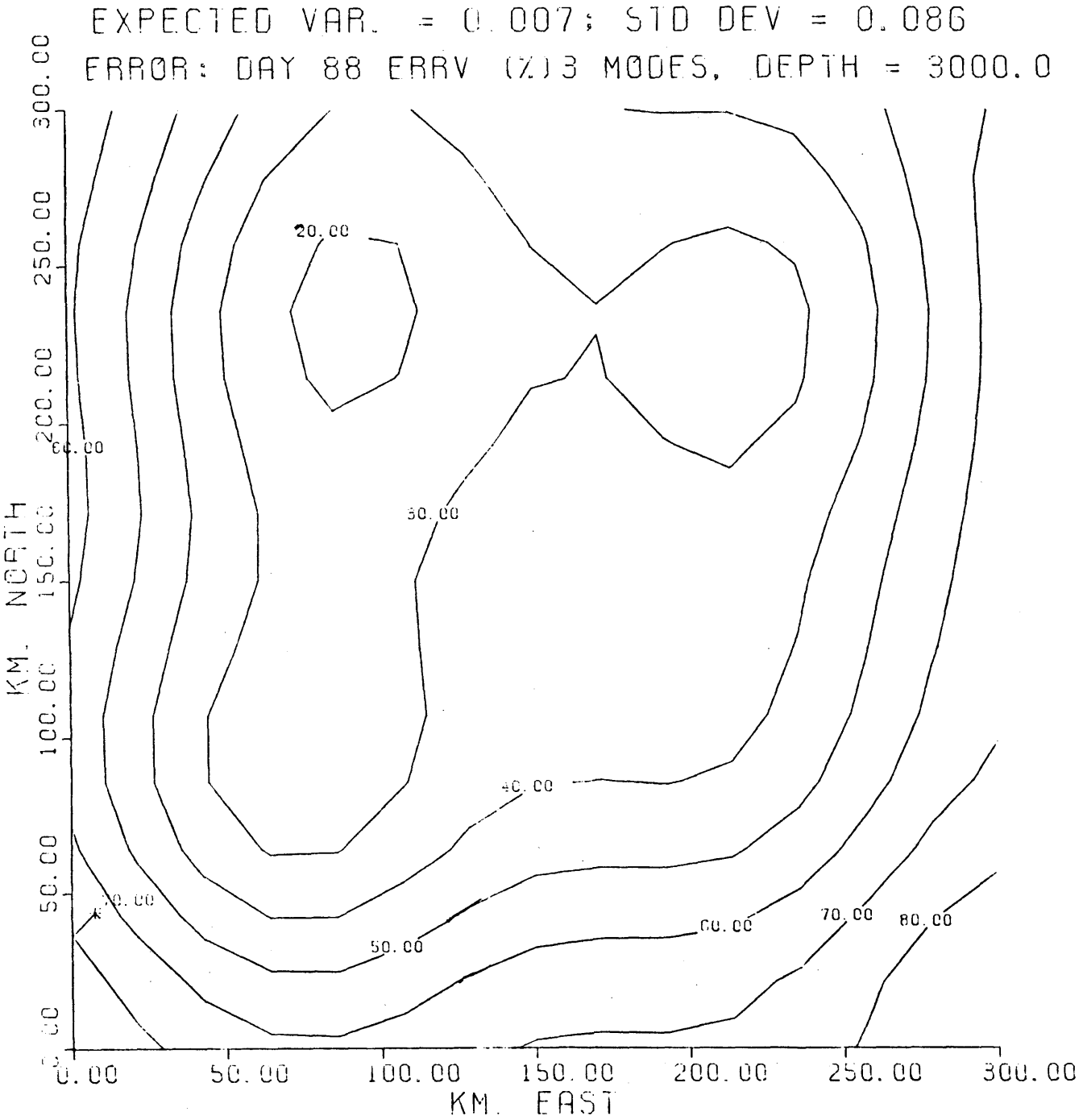


FIGURE 10.5 H: MAP OF EXPECTED ERROR VARIANCE EXPRESSED AS PERCENTAGE OF TOTAL VARIANCE. AT THIS LEVEL (3000 METERS DEEP) THE MODEL PREDICTS 0.09 M/SEC STANDARD DEVIATION FOR THE SOUND SPEED FIELD.

Figure 9.2), so that data at one depth can be used to estimate the amplitude of the mode at another depth where the rays do not sample. An important component of the perturbations is surface intensified, (Figure 9.1B), and is difficult to resolve without using many rays which pass close to the surface.

The ability of the inverse to resolve a given mode is related to the strength of the travel time anomalies that the mode is expected to produce. For example, Table 9.1 lists the expected travel time anomaly variances for several typical rays, broken down by modes. These calculations are produced as part of the data-data covariance matrix construction. The first EOF closely resembles the first baroclinic mode, and is expected to generate strong travel time signals, above the 5 msec noise level. The third EOF somewhat resembles the second baroclinic mode, and is more marginal compared to the noise level, while the second EOF, which accounts for much of the expected variation near the surface, produces a travel time signal which may be lost in the observation noise, so that more near-surface rays are needed before the upper layers can be mapped precisely.

The error maps displayed for the layers (Figure 10.5) summarize the ability of the inverse to resolve the expected variance at each level. Chapters 5 and 6 discussed how the the inverse procedure calculates the expected variance of its estimates of sound speed anomalies everywhere throughout the volume of interest. The error variance is due both to noise in the data and to poor sampling (as when no rays penetrate to the surface). The expected error variance can be expressed as a percentage of the total expected variance, which masks the dependence on the absolute energy level chosen by the parameters listed in Table 9.1. These maps are meant to resemble the error maps which have been included with objective analyses used in oceanography (Bretherton, Davis, and Fandry, 1976).

At locations outside the array of instruments, where no data are available, the stochastic inverse tends to leave the a priori mean undisturbed, producing zero as an anomaly estimate, while the error map shows 100 % of the variance to be unresolved. Because the field is spatially correlated, the resolution does not immediately drop to zero, but the maps are not very reliable around the edges. This impairs comparisons with the southernmost environmental mooring (Figure 1.4), and so time series comparisons have only been made for the central environmental mooring and three of the acoustic moorings. The error maps can also be displayed as error bars, if desired (Figure 10.5), where the numbers are now the expected standard deviations of the estimates in m/sec. Some of the maps have also been made showing the standard deviation of the error, to facilitate quantitative comparisons with the traditional data. These error bars can also be used to quantify the point-by-point time series comparisons presented in Figure 10.6.

The agreement between the acoustics and the CTD survey is generally good, to within the error levels as specified by the maps, except for a few days late in the record, where a strong negative anomaly appears to emanate from source 4, and for a few days near day 100, where a positive anomaly appears near the center of the array. One possible explanation for these "anomalous anomalies" is extreme mooring motion.

The inverse has mooring motion and clock offsets parameterized as part of the forward problem, but the dependences are linearized, just as the dependence of travel time on the sound speed anomalies is linearized around a basic state. For clock error, the linearity is exact, but both horizontal and vertical mooring position changes have been treated by assuming a straight ray (locally) and a constant sound speed. The horizontal

FIGURES 10.6 A,B,C:

THESE FIGURES SHOW THREE COMPARISONS BETWEEN TIME SERIES OF SOUND SPEED CALCULATED FROM THE TOMOGRAPHY SYSTEM (PLOTTED AS SQUARES) AND TIME SERIES OF SOUND SPEED FROM TEMPERATURE-PRESSURE RECORDERS LOCATED ON MOORINGS IN THE ARRAY (PLOTTED AS TRIANGLES). THE TWO CURVES HAVE BEEN OFFSET SLIGHTLY TO AVOID CONFUSING ERRORS ASSOCIATED WITH TEMP-SOUND SPEED CONVERSION, AND SO ONLY THE SHAPES (SLOPES AND EXCURSIONS) OF THE CURVES SHOULD BE COMPARED.

FIGURE 10.6 A: TIME SERIES OF SOUND SPEED FOR LOCATION OF T-P RECORDER ON S3 MOORING AT ABOUT 867 METERS DEPTH

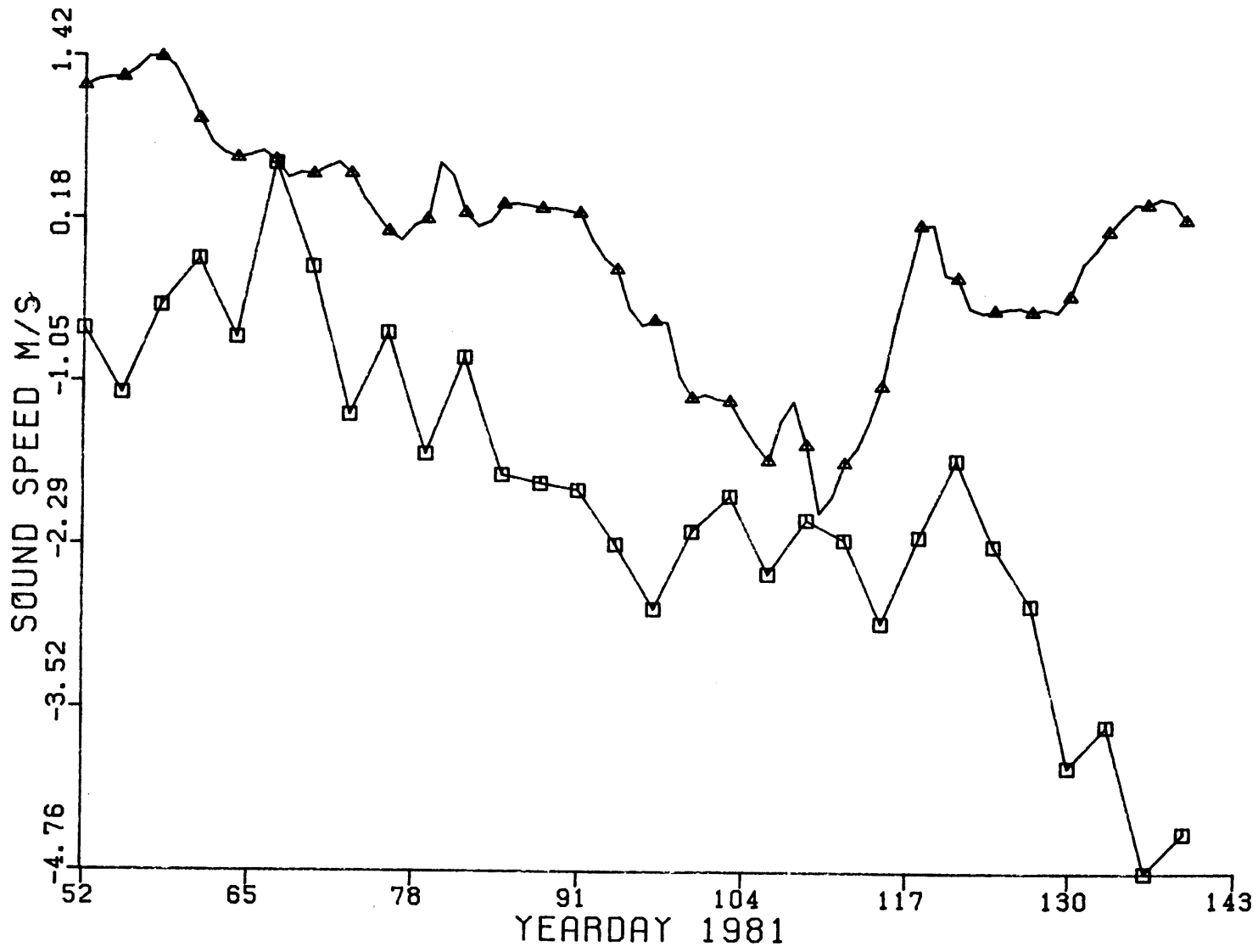


FIGURE 10.6 B: TIME SERIES OF SOUND SPEED FOR LOCATION OF T-P RECORDER ON E1 MOORING (AT CENTER OF ARRAY). DEPTH IS ABOUT 626 METERS.

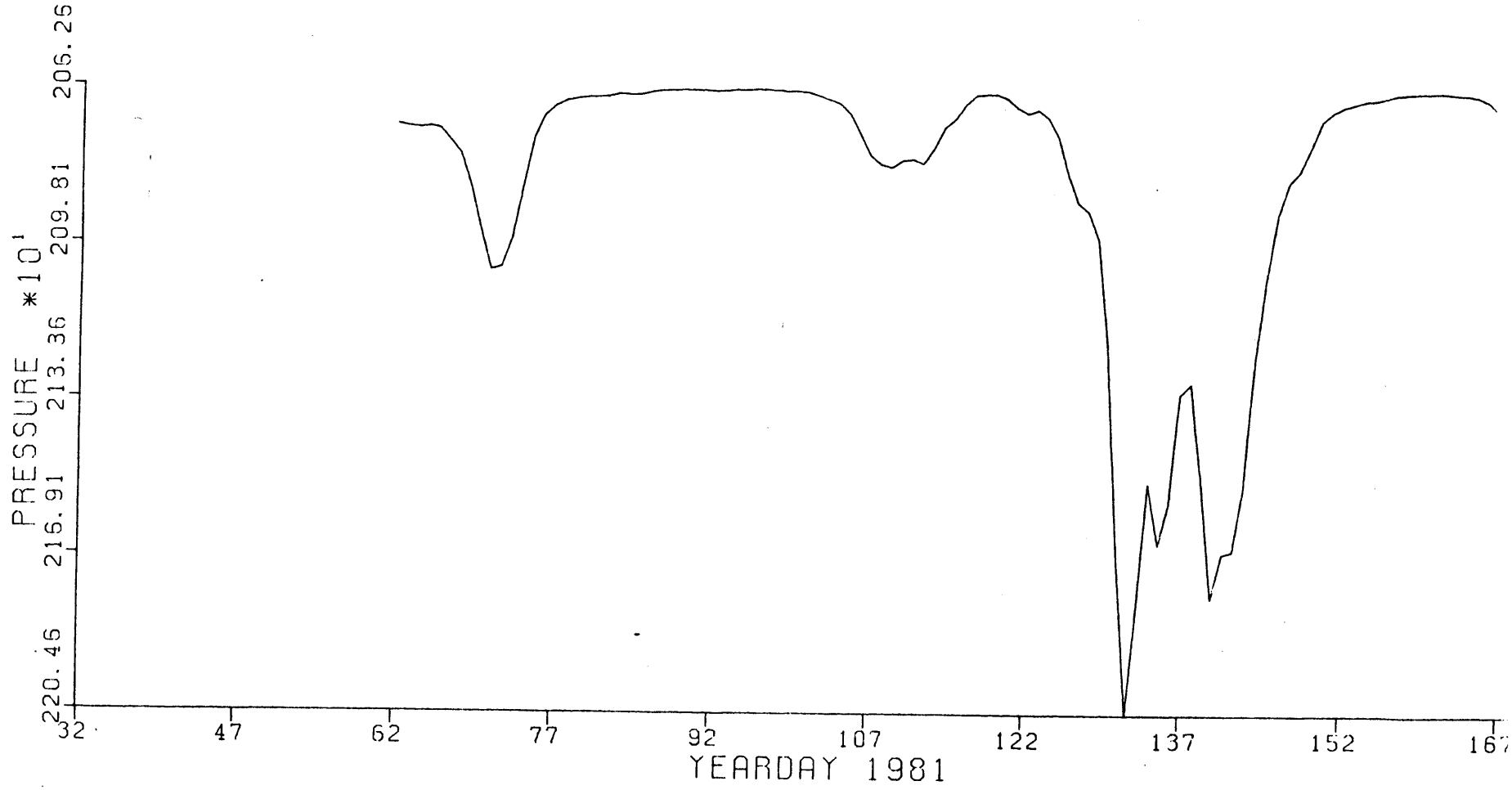
FIGURE 10.6 C: TIME SERIES OF SOUND SPEED FOR LOCATION OF T-P RECORDER ON R3 MOORING AT ABOUT 625 METERS DEPTH

linearization holds for displacements of up to 2 km., but the vertical displacements are considerably less robust. I estimate that depth changes of more than about 50 meters will produce significant (Order 5 msec) errors in the linearization, both through local inaccuracies and through changes in the overall ray path.

The inverse procedure returns estimated locations of the instruments as well as the sound speed maps, so large estimated displacements signal that the linearization may be questionable. At this point, it is also possible to take advantage of the physical structure of the mooring, since the x, y,and z displacements were originally assumed to be independent. A large horizontal displacement of the mooring should be accompanied by a deepening of the instrument, while the instrument should never go shallower than the "rest" depth defined above for the undisturbed mooring. These two constraints may perhaps be included in later inversions, but at the present they permit consistency checks on the estimates. A simpler check of consistency is to compare the acoustic estimates of instrument displacements with T-P records.

Figure 10.8 is the depth variation of source 2 calculated from the acoustics and Figure 10.7 is pressure from a T-P recorder on the mooring near the source. The

FIGURE 10.7 : TIME SERIES OF PRESSURE IN DECIBARS FOR A T-P RECORDER ON THE SAME MOORING AS SOURCE 2. NOTE THAT PRESSURE IS INCREASING DOWNWARD. THE DEPTH INCREASES RESULT FROM MOORING LEAN, SO THAT SOURCE 2 SHOULD ALSO HAVE INCREASED DEPTHS AT THESE TIMES.



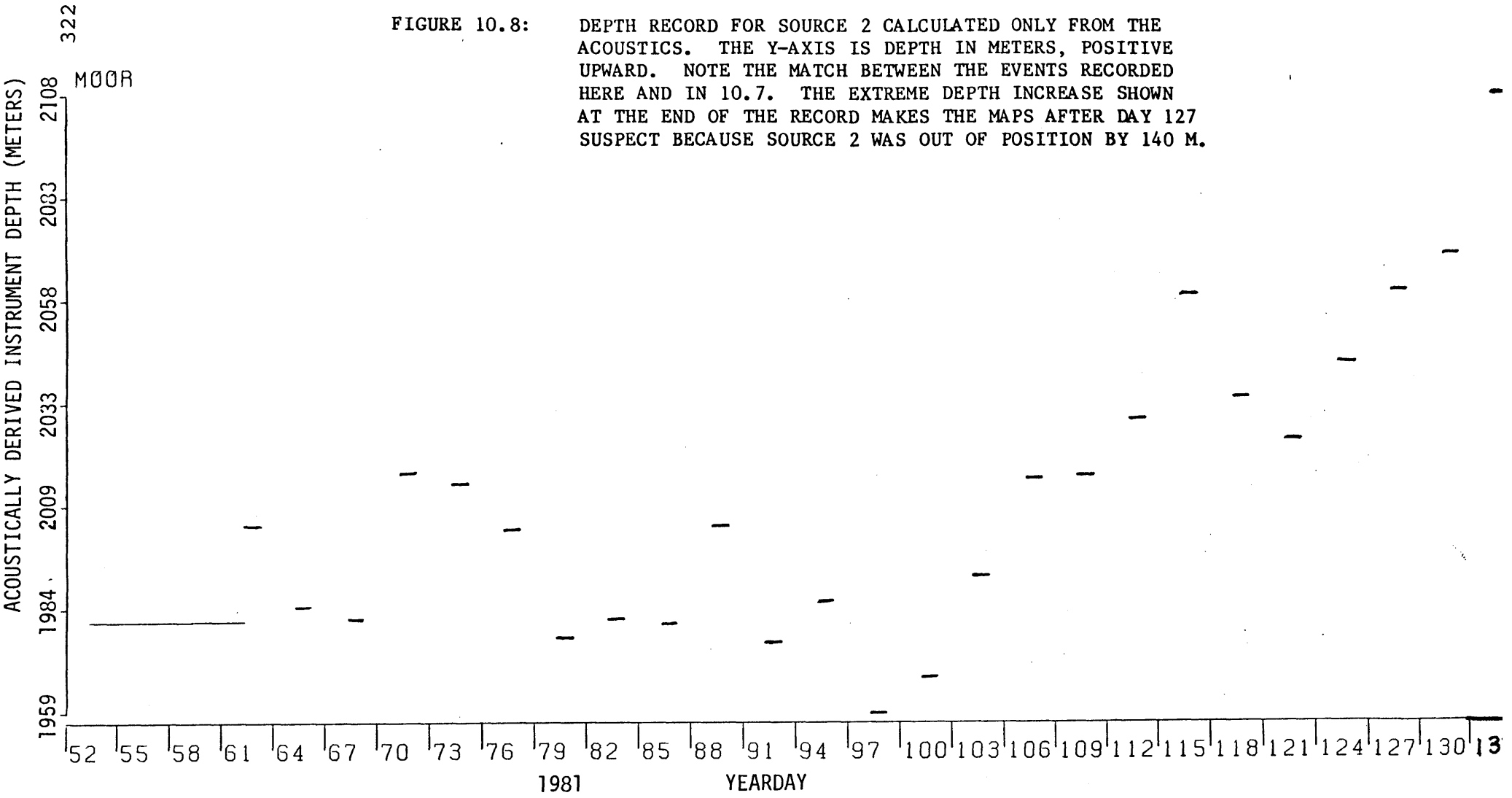


FIGURE 10.8: DEPTH RECORD FOR SOURCE 2 CALCULATED ONLY FROM THE ACOUSTICS. THE Y-AXIS IS DEPTH IN METERS, POSITIVE UPWARD. NOTE THE MATCH BETWEEN THE EVENTS RECORDED HERE AND IN 10.7. THE EXTREME DEPTH INCREASE SHOWN AT THE END OF THE RECORD MAKES THE MAPS AFTER DAY 127 SUSPECT BECAUSE SOURCE 2 WAS OUT OF POSITION BY 140 M.

two time series compare quite well, and both show the extreme depth excursion of source 2 beginning on about day 122. The receivers move only weakly (O(10 m.), but sources 2 and 4 are particularly active. Source 2 (S2) is about 40 m. below its "rest" position during the days 67-77, and is about 140 m. deeper beginning on about day 122. Source 4 is 120 to 170 meters deeper between day 60 and day 77, and goes completely off scale (deeper than 170 meters) after day 136. The inverse results during these periods is thus suspect. Once again, in later inversions these T-P data should be included as part of the total data set, but in the present "proof" stage they provide another point of comparison for evaluating the inverse system.

The system could be re-linearized around the new positions, but that was not done for these simple demonstrations, nor were the data weighted variably for error and expected mooring offsets. The inversions presented here represent very little "tweaking" or tuning of parameters, in the hope that the relatively simple procedure would increase credibility. No mooring motion corrections were used in the data set, and the positions and depths of the instruments were determined by the inversions themselves. The weighting parameters were based at least partly on the residual uncertainties from anchor

position and "rest" depth determinations, although sources 2 and 4 were given large variances on the basis of the T-P records. In the next version of the data processing, the data from the mooring tracking will be used, where it exists, providing both an a priori estimate of instrument location and an estimate of the remaining uncertainty day by day. At the very least, the large variances for the instrument depths can be reduced, and the linearization can be re-done on each day, using the a priori position estimates.

10.2 IMPROVEMENTS TO THE 1981 MAPS

One of the most striking features of the maps from the inverse system (Figures 9.6, 10.2, 10.3, 10.4) is the continuity from day to day. This is expected on the basis of the time scales (O(50 days)) of the mesoscale motions, and it is very tempting to incorporate these expectations into the inverse methods. At present, the maps on a given day are independent of all the other days, even though the mesoscale features change very little over three days, so the similarity between successive maps provides a consistency check on the inversions. These consistency checks can of course be converted to constraints on the inversions to improve the performance of the system. The simplest modification would be to average the travel time data over a period of 6 to 12 days, reducing the random errors but complicating the mooring position problem somewhat.

Simple averaging is only a stopgap measure, and it is preferable to impose short-term continuity as a constraint, either explicitly, producing additional "data", or implicitly, by requiring the model to satisfy the constraint directly. The implicit approach is more elegant, and is frequently far simpler. Throughout the discussion in Chapters 4-7, the covariances were allowed to

be time-dependent, but the covariances used in the processing to date have been time-independent. A persistence constraint could be enforced by specifying a covariance which decayed only slowly over time, while schematic mesoscale dynamics could be introduced by incorporating a "group velocity" into the covariance, so that features would be expected to drift westward at a few km./day. The latter approach has been used for the POLYMODE XBT maps (Carter and Robinson, 1983), to compensate for gaps in a spotty data set. The application to the 1981 tomography maps would be far less critical, due to the relatively short (3 day) time between measurements, so that even the short-term persistence hypothesis would be expected to yield increased resolution without introducing much error due to the assumptions.

The mesoscale dynamics could be enforced more rigorously by requiring the unknown sound speed field to be made up of a superposition of solutions to the linearized potential vorticity equation (Chapter 3). A planetary wave expansion limits the results of the inverse to have specific forms, and so abandons much of the generality originally introduced by adopting the stochastic inverse form. If data exist which allow these forms to be

specified in advance, great improvements in the resolution of the inverse can be expected. For example, in the 1981 experiment, the 3 CTD surveys could be used to build a basis set of waves for the observed anomalies, so that the acoustic data would only be required to establish magnitudes and phases. As always, the increased resolution comes at the cost of becoming blind to phenomena which violate the a priori constraints, although residual levels could be monitored as a check on the consistency of the model.

Including the hydrographic, current meter, and T-P data directly into the inverse is also straightforward, and continues the theme of converting consistency checks into increased resolving power. Once the concept of tomography is legitimized, the data from the experiment should be used to produce the best possible description of the physical oceanography of the region. It would certainly be illogical, given this goal, to exclude any part of the data from the estimation process. The only complication incurred in combining disparate data is that absolute error levels must be established for each of the data sources to control relative weighting.

Many more sophisticated improvements for the inverse are also possible, and several have been mentioned earlier in this thesis. The ocean currents produce travel time

anomalies, but these have been neglected in the maps produced to date. This simplified the calculations of the estimators but also introduced O(2 msec) of quasi-random error, which distorts the results and lowers the resolution limits. The unknown "barotropic" velocity mode should be about as well resolved as the second EOF in the examples presented in Table 9.1, based on comparing the expected travel time anomalies due to velocities to the truly random error level. The inverses may not produce detailed current maps, but it is important to parameterize all sources of variance, to avoid having to add to the basic random error incurred by the limits of the pulse arrival time precision.

As suggested by Figure 9.5, it is this level of irreducible random error which provides the ultimate limits on resolution, since the inverse cannot be allowed to be sensitive to anomalies at or below the level of the error. For example, if the random error standard deviation is 10 msec, then it does little good to add in rays which have expected travel time anomalies less than this amount, or which seem identical to similar rays when looked at subject to this blurring. The addition of constraints to the inverse can improve the resolution by effectively narrowing the "bandwidth" of interest, i.e. restricting the possible

forms of the solution. The total noise power in the restricted range of forms will be less than for the unrestricted range, so that the inverse gains some noise-immunity, and so can be allowed to be sensitive to smaller travel time anomalies, and thus gain resolution. All of the improvements discussed above work in this way, and are designed to combat the relatively large (5 msec) basic random error inherent in the data from processing and transmission channel noise. When the travel time anomalies were expected to be O(200 msec), 5 or 10 msec of error was not a problem, but when the expected "signal" is 40 msec, then a 5 msec noise level greatly restricts the possibilities of even the "ultimate" inversions. For this reason, the modifications to the original data processing outlined in Chapter 8 are of critical importance. Every millisecond reduction in the random error will pay large returns in increased resolving power.

This can be seen graphically in Figures 10.10 and 10.12, which show results produced by the present inverse when fed simulated travel time data for an ocean filled up with planetary waves (Figures 10.9 and 10.11). With no modifications to the inverse except for reduced random error, the resolution of the 1981 tomographic array can be increased radically, and the maps become relatively immune

FIGURE 10.9: TEST "OCEAN" CONSTRUCTED FROM 300-KM WAVELENGTH ROSSBY WAVES. CONTOURS ARE OF SOUND SPEED ANOMALY RELATIVE TO AVERAGE.

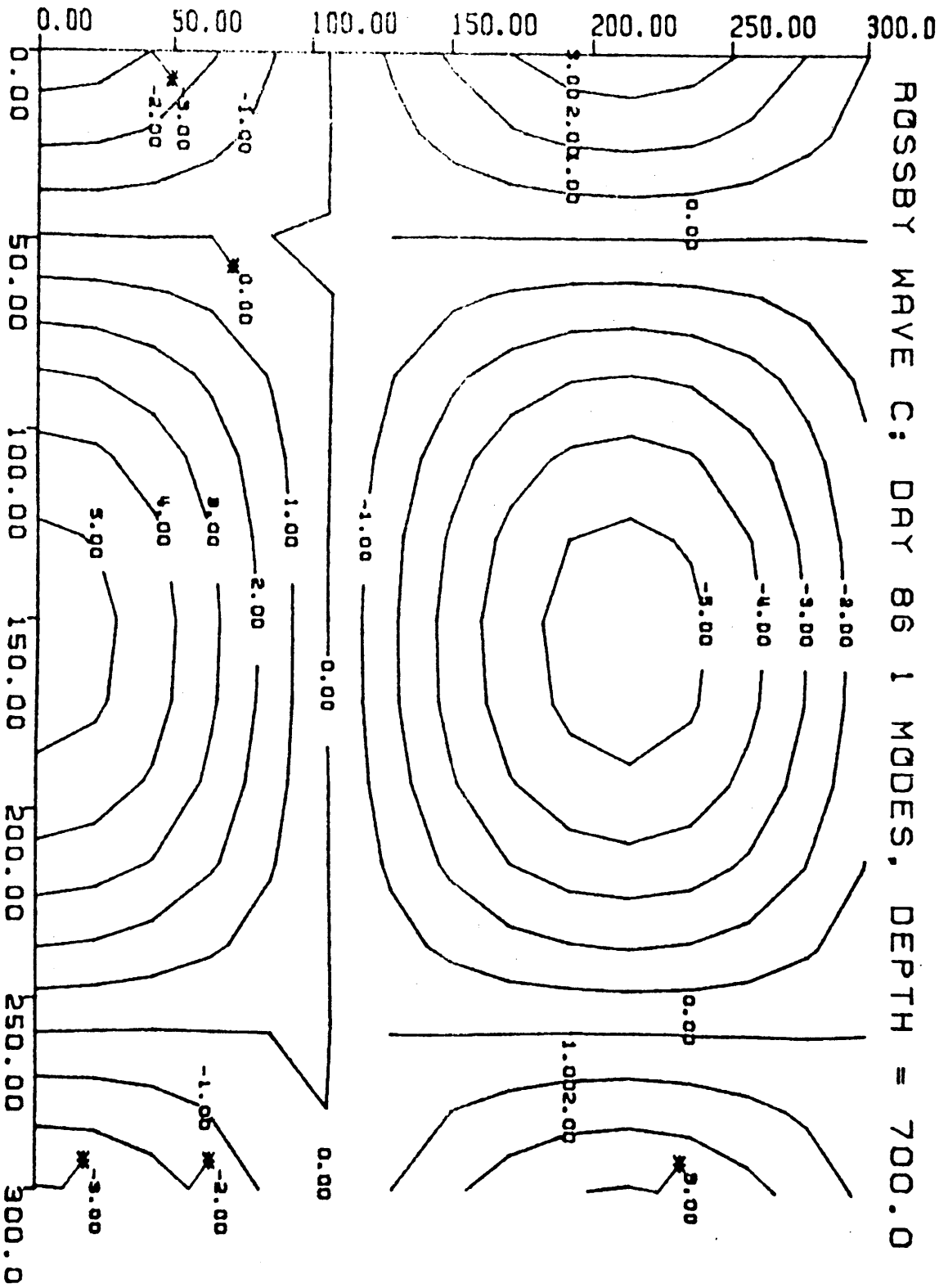


FIGURE 10.10 A: ESTIMATE OF FIELD SHOWN AS FIGURE 10.9 USING DATA CONSTRUCTED BY TRACING RAYS IN THE 3-DIMENSIONAL "OCEAN" REFERRED TO BY FIGURE 10.9. ERROR = 5 MSEC., CORRECTED DATA.

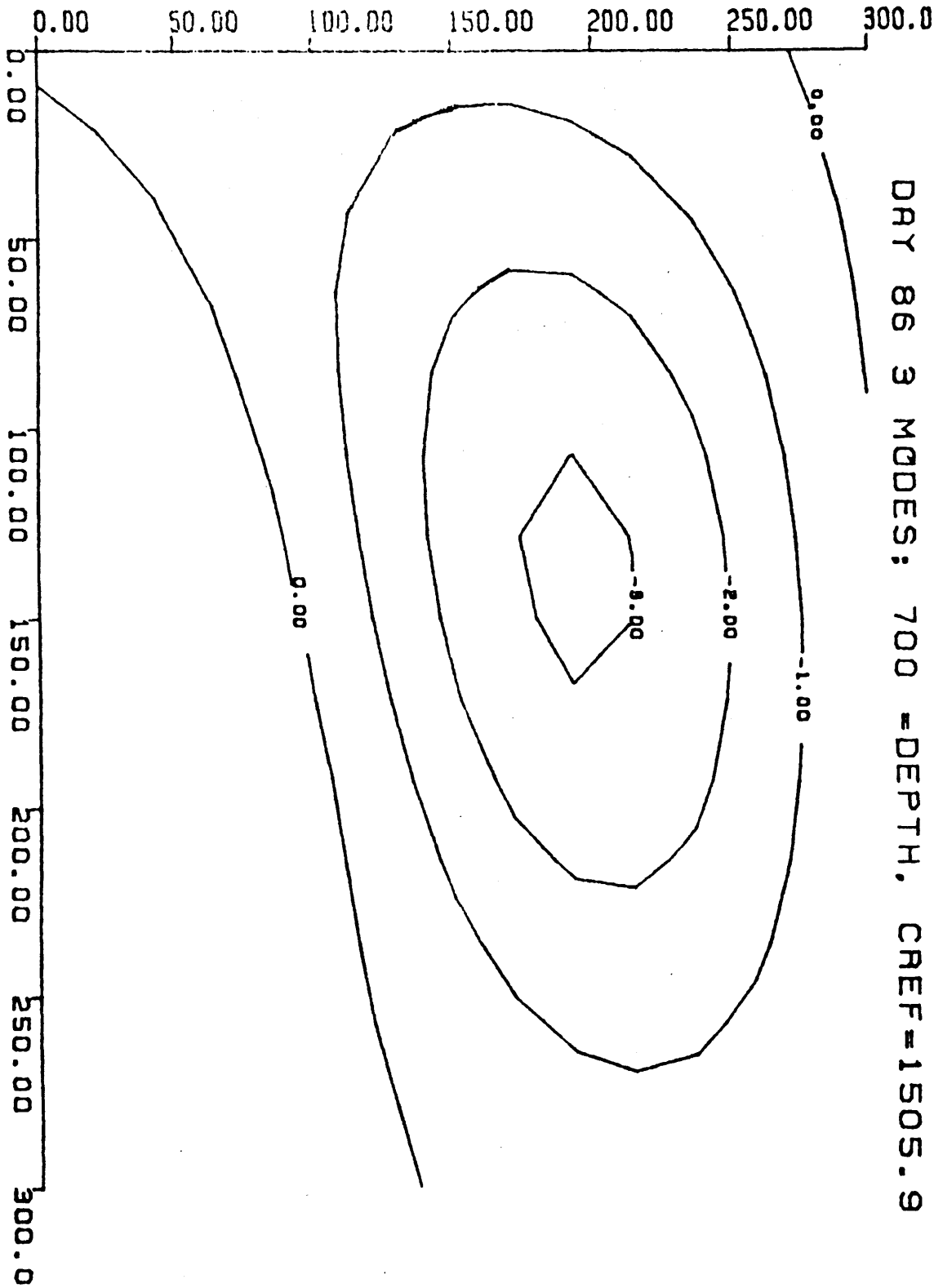


FIGURE 10.10 B: ESTIMATE OF FIELD SHOWN AS FIGURE 10.9 USING DATA CONSTRUCTED BY TRACING RAYS IN THE 3-DIMENSIONAL "OCEAN" REFERRED TO BY FIGURE 10.9. ERROR = 2 MSEC., CORRECTED DATA.

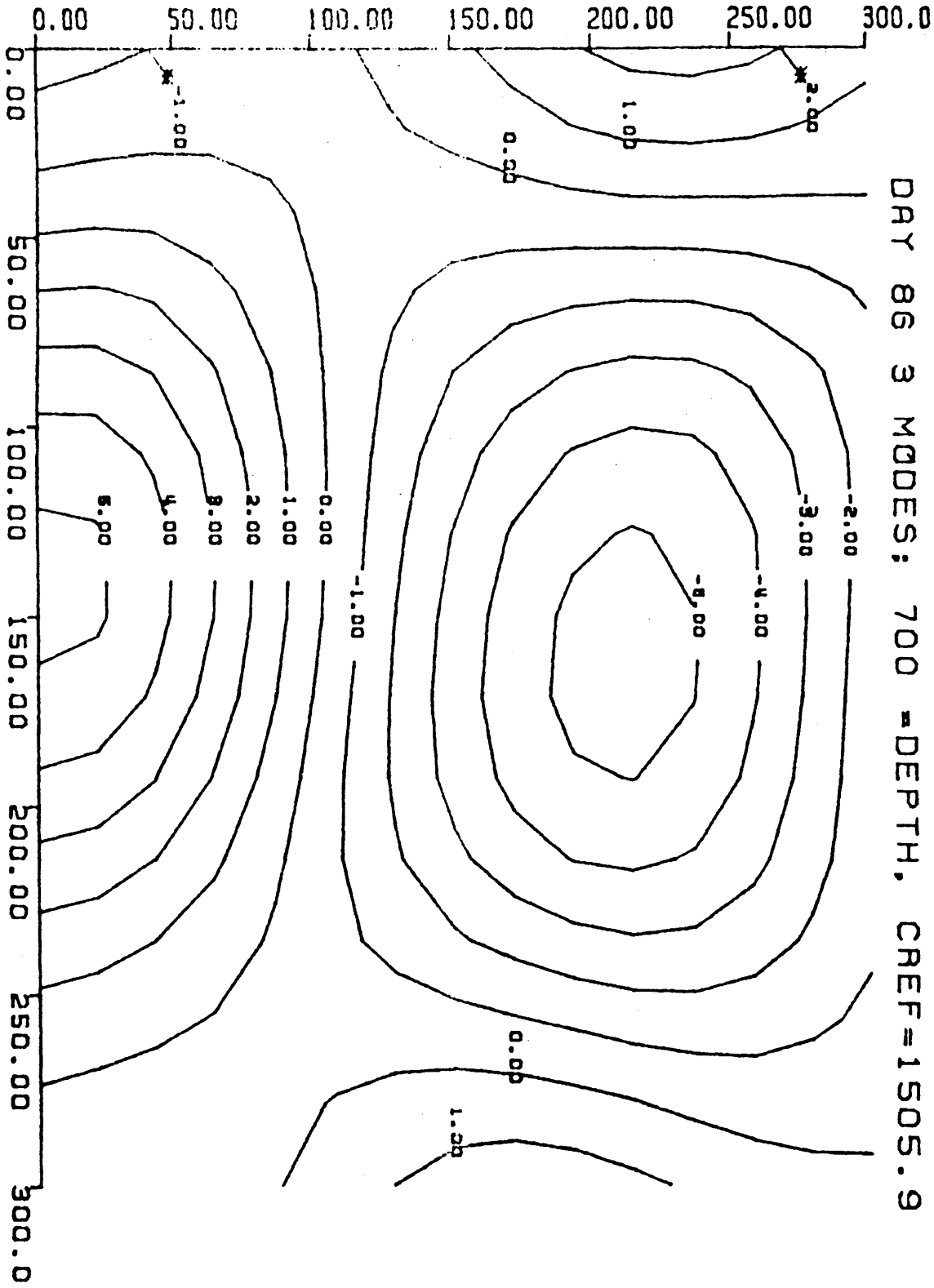


FIGURE 10.10 C: ESTIMATE OF FIELD SHOWN AS FIGURE 10.9 USING DATA CONSTRUCTED BY TRACING RAYS IN THE 3-DIMENSIONAL "OCEAN" REFERRED TO BY FIGURE 10.9. ERROR = 2 MSEC., NO MOORING MOTION CORRECTIONS

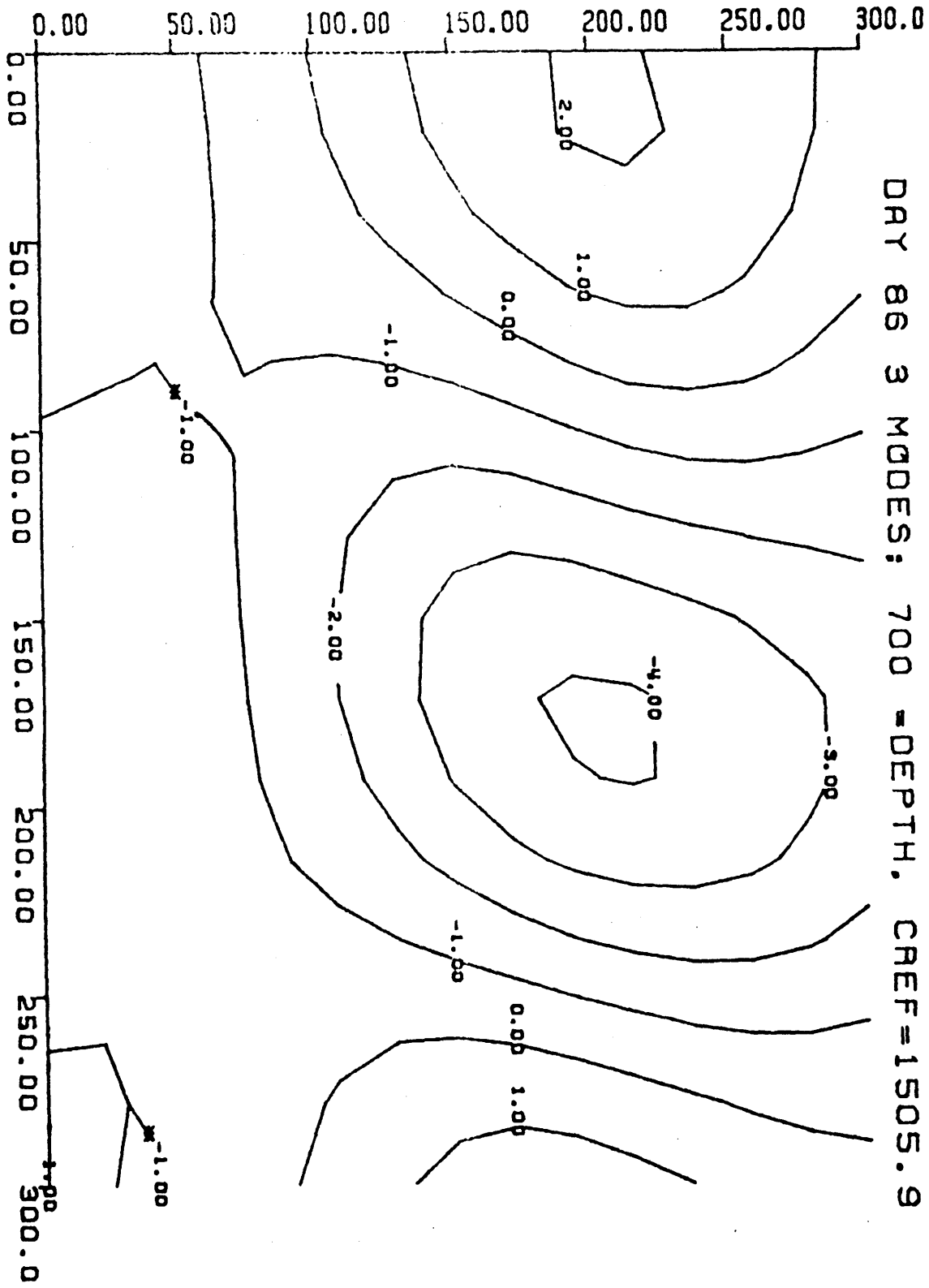


FIGURE 10.11: TEST "OCEAN" CONSTRUCTED FROM 150-KM WAVELENGTH ROSSBY WAVES. CONTOURS ARE OF SOUND SPEED ANOMALY RELATIVE TO AVERAGE.

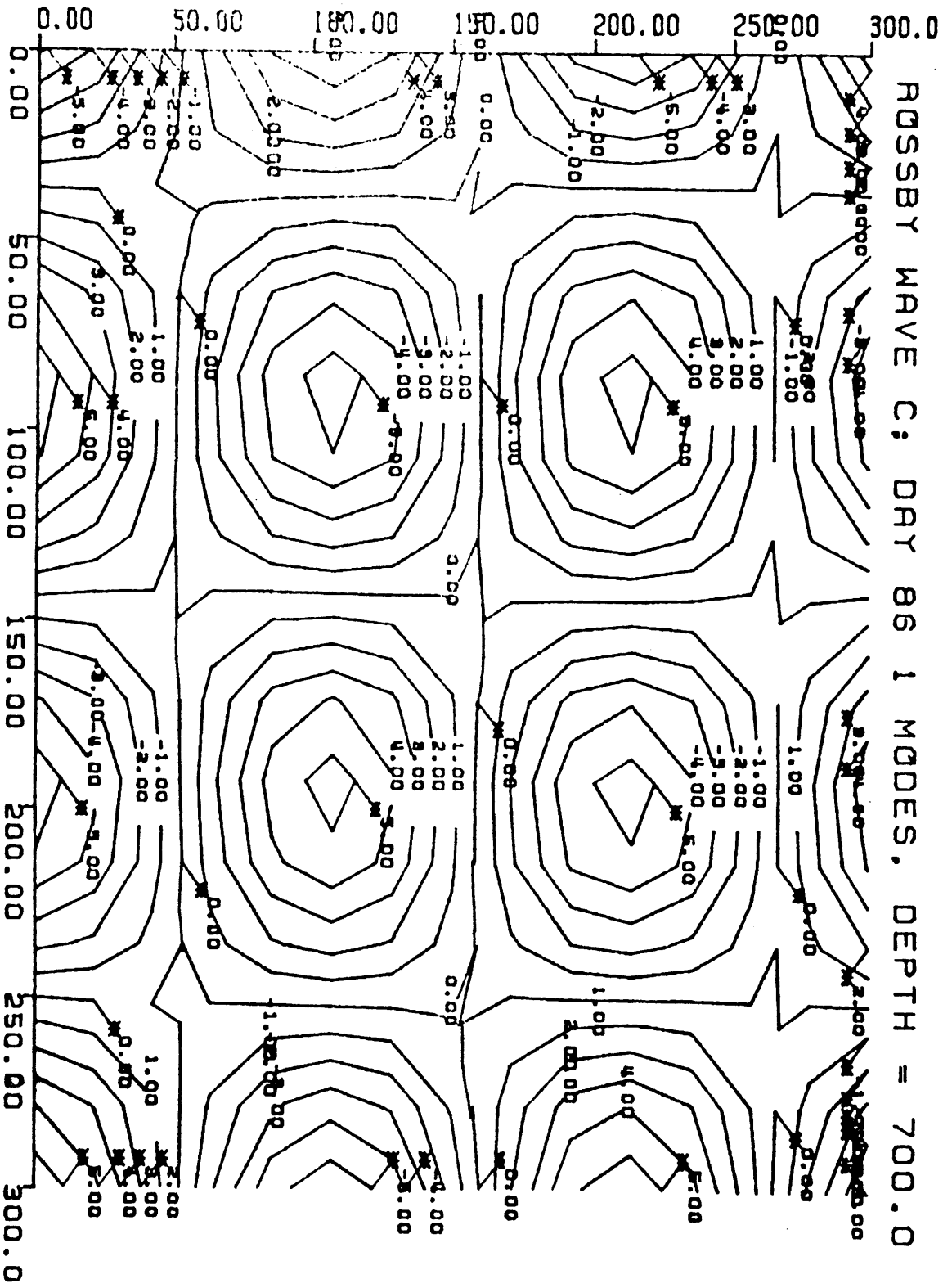
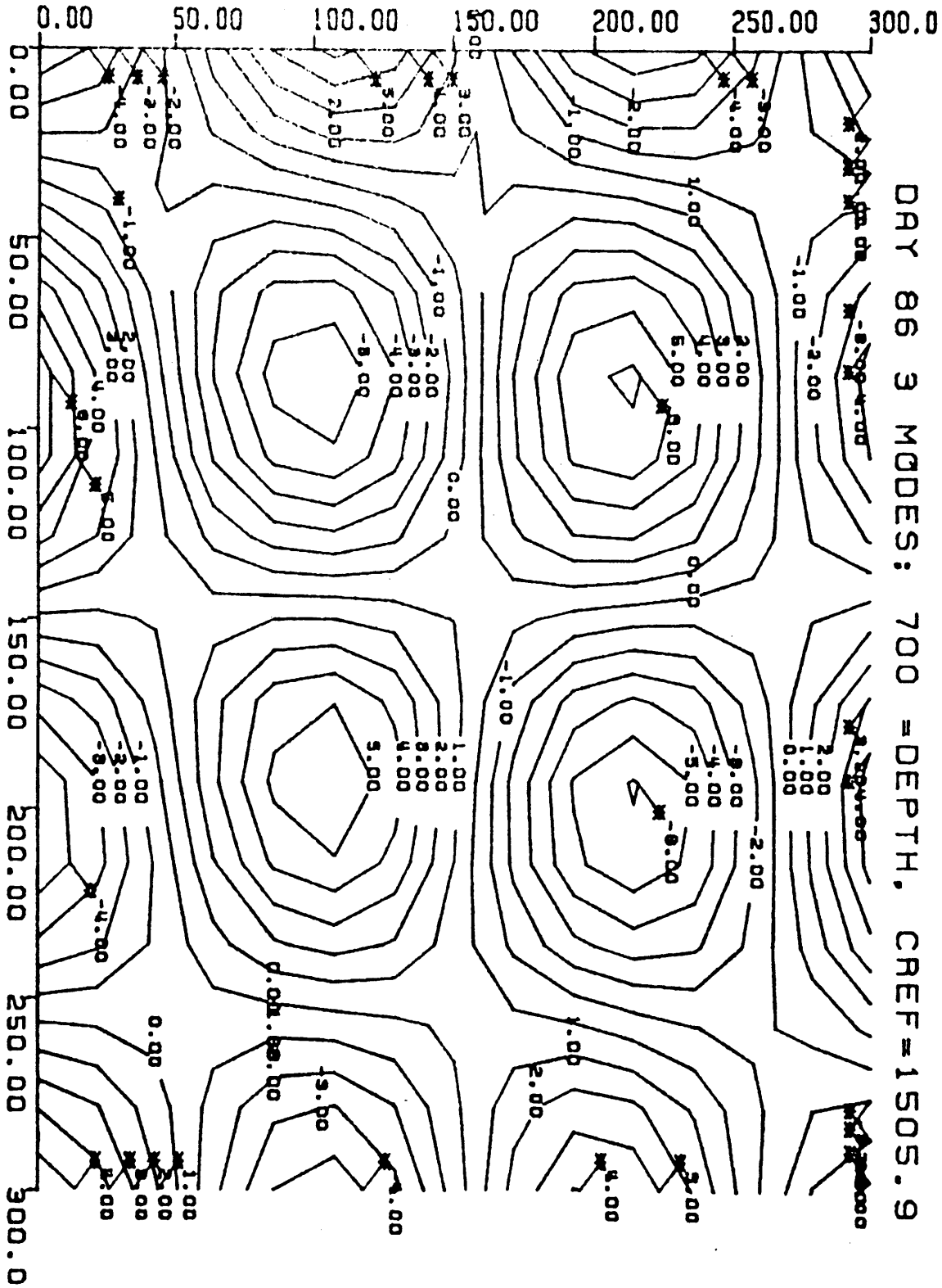


FIGURE 10.12 : ESTIMATE OF FIELD SHOWN AS FIGURE 10.11 USING DATA CONSTRUCTED BY TRACING RAYS IN THE 3-DIMENSIONAL "OCEAN" REFERRED TO BY FIGURE 10.11. ERROR = 0.5 MSEC., CORRECTED DATA.



to unknown mooring position. Mesoscale tomography is limited only by the precision of the travel time determination, and not by complicated mooring hardware. The sources and receivers have no exposed moving parts, and the precision is limited by the available level of digital electronic technology, which is increasing at a rapid rate. The present inverse framework is designed to include rigorous self-evaluation, in the forms of both error maps and results from simulated data, so that it is possible to juggle the engineering trade-offs in a very rational manner, much as objective mapping provided a means for evaluating array layouts for current meters.

10.3 FUTURE APPLICATIONS OF TOMOGRAPHY

The methods discussed in this thesis suggest a basis for designing all oceanography experiments, and they are being used at present to explore possibilities for future applications of the tomographic techniques. Because tomography is a form of remote sensing, the most obvious uses are in cases where it is inconvenient to directly sample the region of study. In the 1981 application, the acoustics represented a way to gather a synoptic data set over an extensive region, without instrumenting the volume at the required spacing. This same argument applies, with greater force, to the problem of observing an entire ocean basin (Munk and Wunsch, 1982). In some high-current areas, such as the Gulf Stream, it is difficult to moor instruments directly in the current, so that the capability to study the current using instruments moored out of harm's way is important.

Munk and Wunsch (1982) proposed a scheme for monitoring a basin-sized region using equipment similar to the 1981 experiment, but transmitting reciprocally to heighten the resolution of current velocity. They point out that, because acoustic tomography uses ray travel time data which average the ocean over long distances, tomography should be most effective in estimating averaged

quantities, and may in fact be the most practical way to obtain such averages for a large basin. They propose a simple "array" of 5 instruments to measure large-scale heat content and other climatological quantities. This array of transceivers can in fact estimate large-scale averaged vorticity by measuring circulation around regions enclosed by sets of three instruments. The engineering requirements for the large scale experiment are not unreasonable, given the knowledge acquired during the 1981 experiment. Peter Worcester (1977) has already demonstrated reciprocal transmission in one instance, and the Tomography Group is currently engaged in developing the capability to transmit reciprocally over long ranges using moored instruments.

The basin scale experiment is planned for several years in the future, and simulations have not yet been done, but Gulf Stream monitoring is also an engineering possibility, and has been examined in some detail. The strong currents of the Gulf Stream make it more difficult to instrument than the relatively quiet mid-gyre areas, and it is attractive to consider placing acoustic moorings near the bottom under the Stream and/or outside the high-velocity regions.

One possible arrangement is shown in Figure 10.13. Each instrument is a transceiver, so that all paths are reciprocal, and the surface bounces ensure that the rays gather data at all depths. Figure 10.14 shows an averaged sound speed profile from archived stations, Figure 10.15 shows an actual Gulf Stream section expressed as sound speed anomalies relative to this averaged profile, while Figure 10.16 is the estimate of the section using travel times from the rays shown in Figure 10.13.

The steep angles of the rays from bottom-mounted instruments minimize path changes, so that re-linearization is not necessary, even in the presence of strong, O(40 m/sec) perturbations. These estimates are based on a model of the Gulf Stream built up of vertical modes (Figure 10.17), and a horizontal covariance (Figure 10.18), just as in the mesoscale case. The mode amplitude estimates can be used to estimate density, velocity, or transport as well, while the reciprocal paths should provide good resolution of cross-stream velocities. Although no vertical rays are shown in Figure 10.13, they can be timed extremely accurately, and, since the sound speed structure is determined by the side-looking rays, the inverted echo soundings can be converted accurately to surface height, providing another version of altimeter for monitoring variability in the total flow field.

FIGURE 10.13 PARTIAL PLOT OF RAYS FOR A SET OF 7 BOTTOM-MOUNTED TRANSCIEVERS MOUNTED ON A SECTION UNDER THE GULF STREAM.

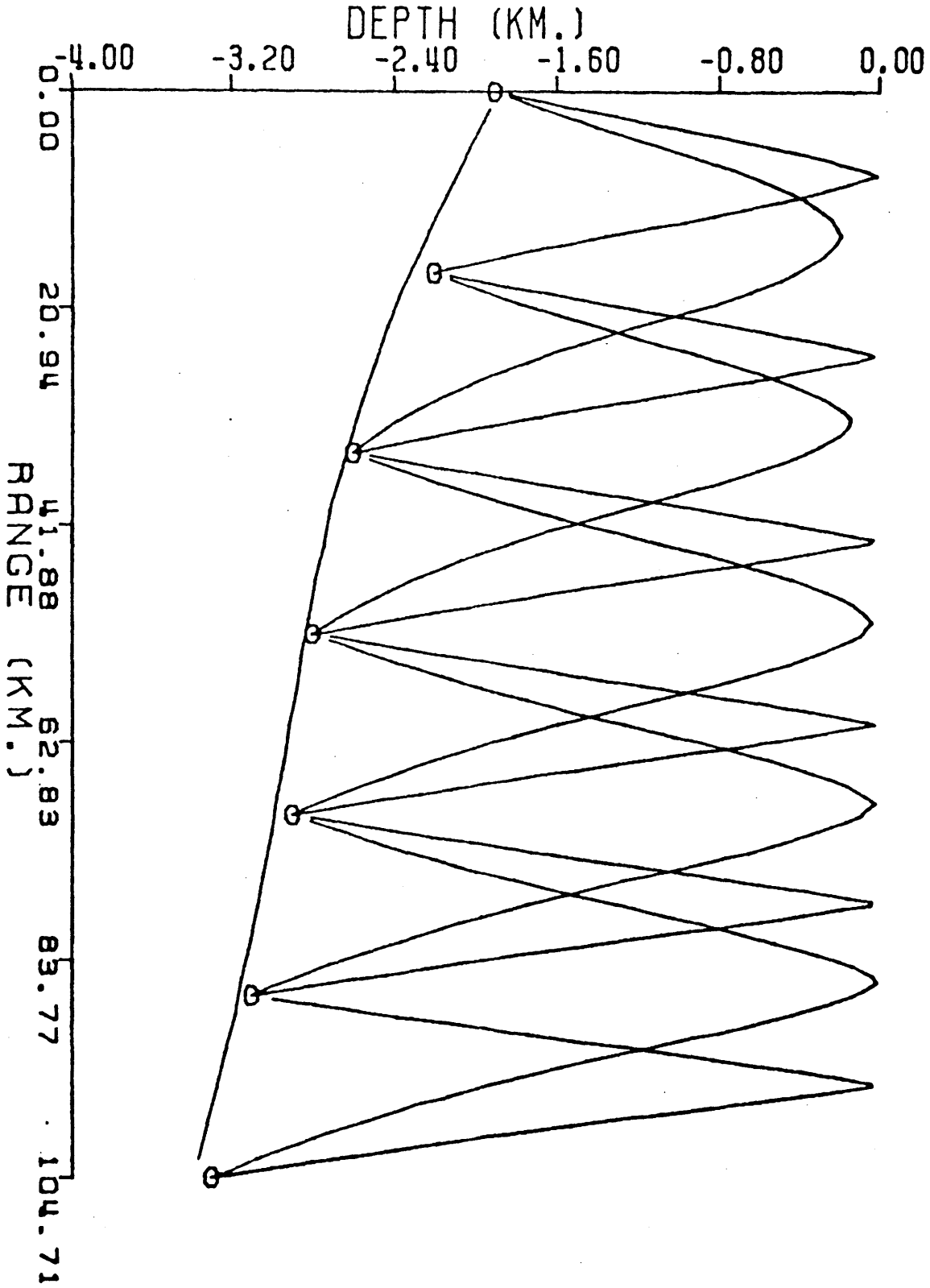


FIGURE 10.14 AVERAGED SOUND SPEED PROFILE FOR THE SECTION SHOWN IN FIGURE 10.13. BASED ON ARCHIVED HYDROGRAPHIC SECTIONS.

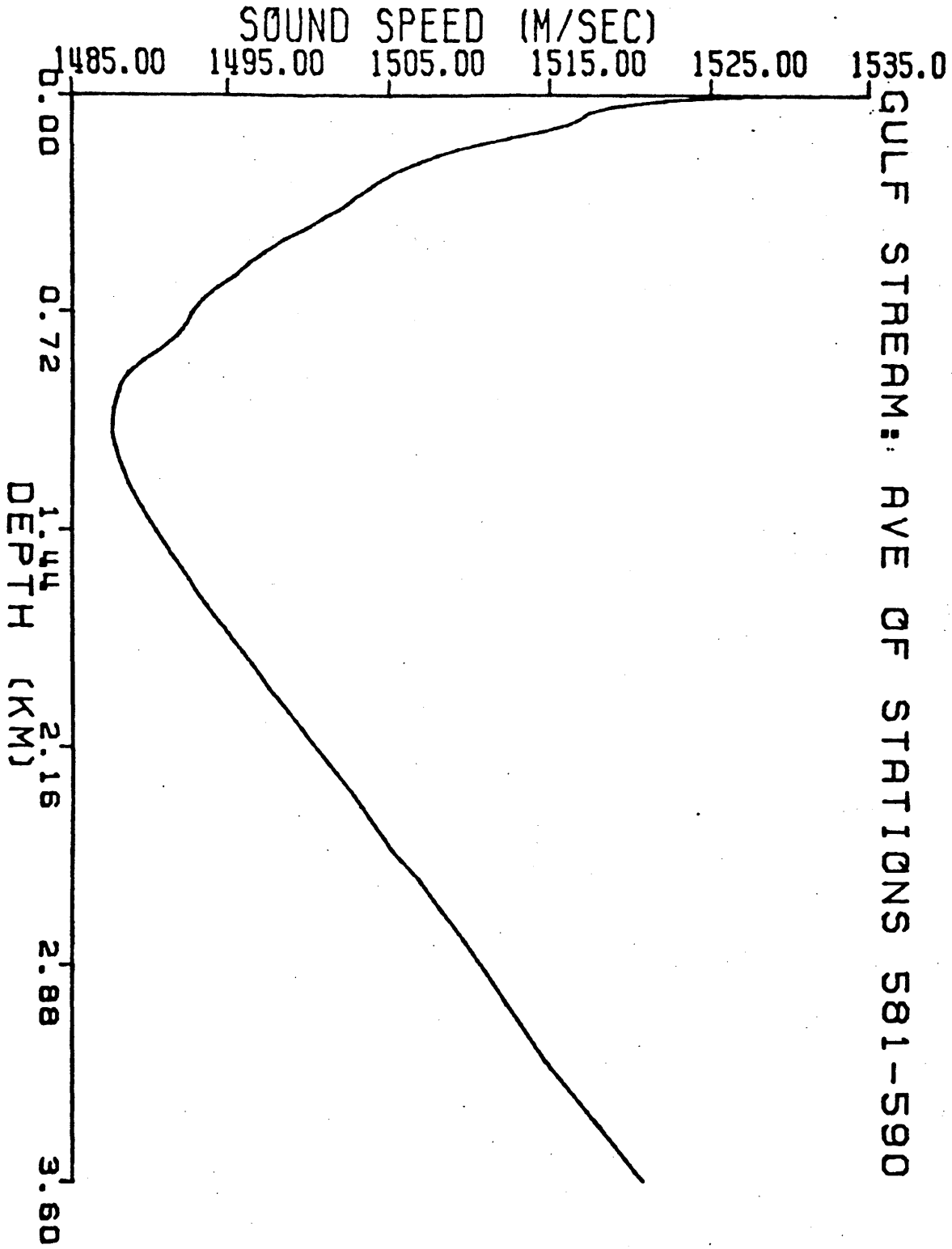


FIGURE 10.15 PLOT OF SECTION ACROSS GULF STREAM SHOWN IN FIGURE 10.13. CONTOURS ARE SOUND SPEED ANOMALY RELATIVE TO THE AVERAGE PROFILE SHOWN IN FIGURE 10.14. CONTOUR INTERVAL IS 8 M/SEC.

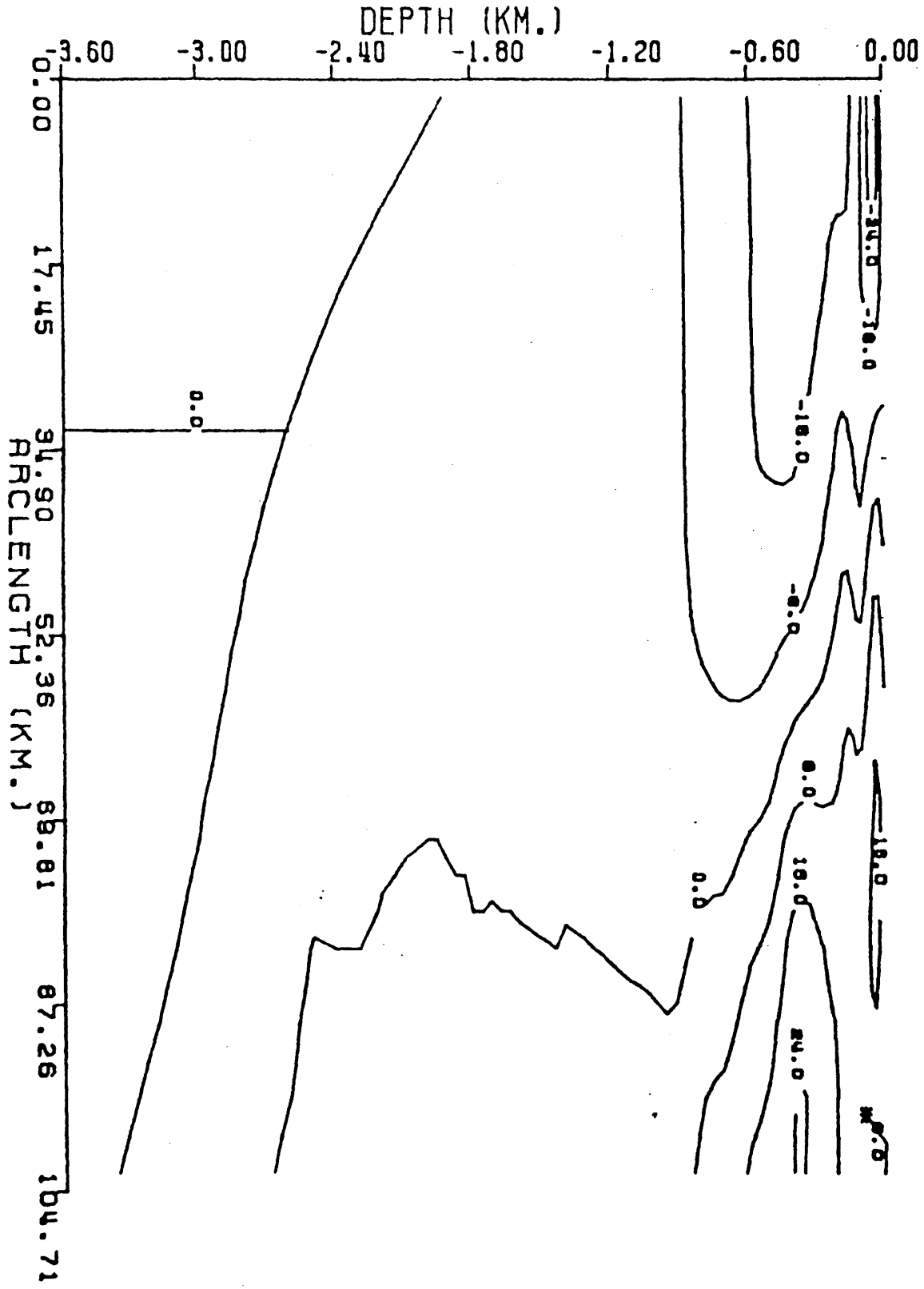


FIGURE 10.16 ACOUSTIC INVERSE ESTIMATE FOR SECTION SHOWN IN FIGURE 10.15. CONTOURS ARE THE SAME AS IN FIGURE 10.15.

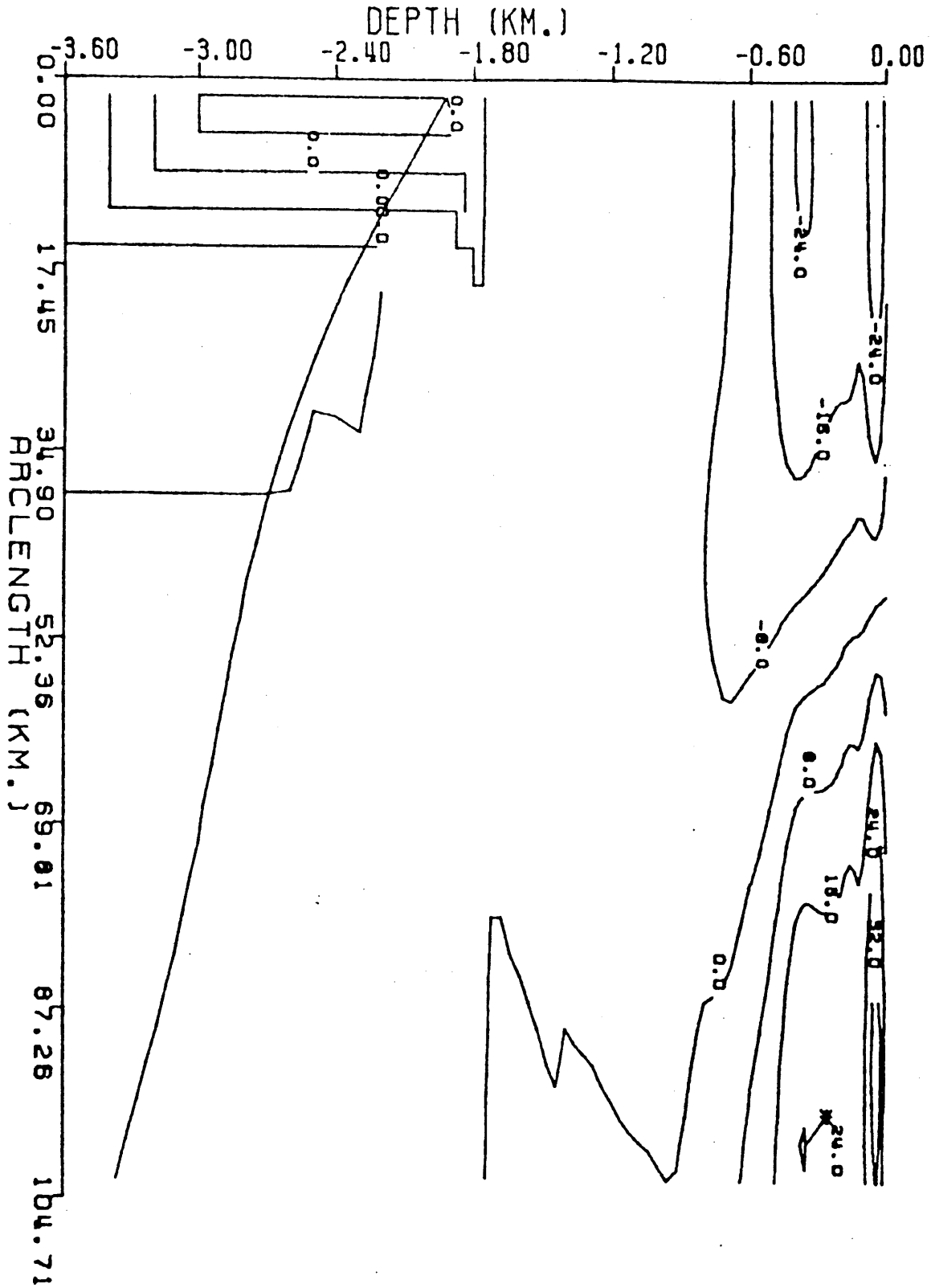
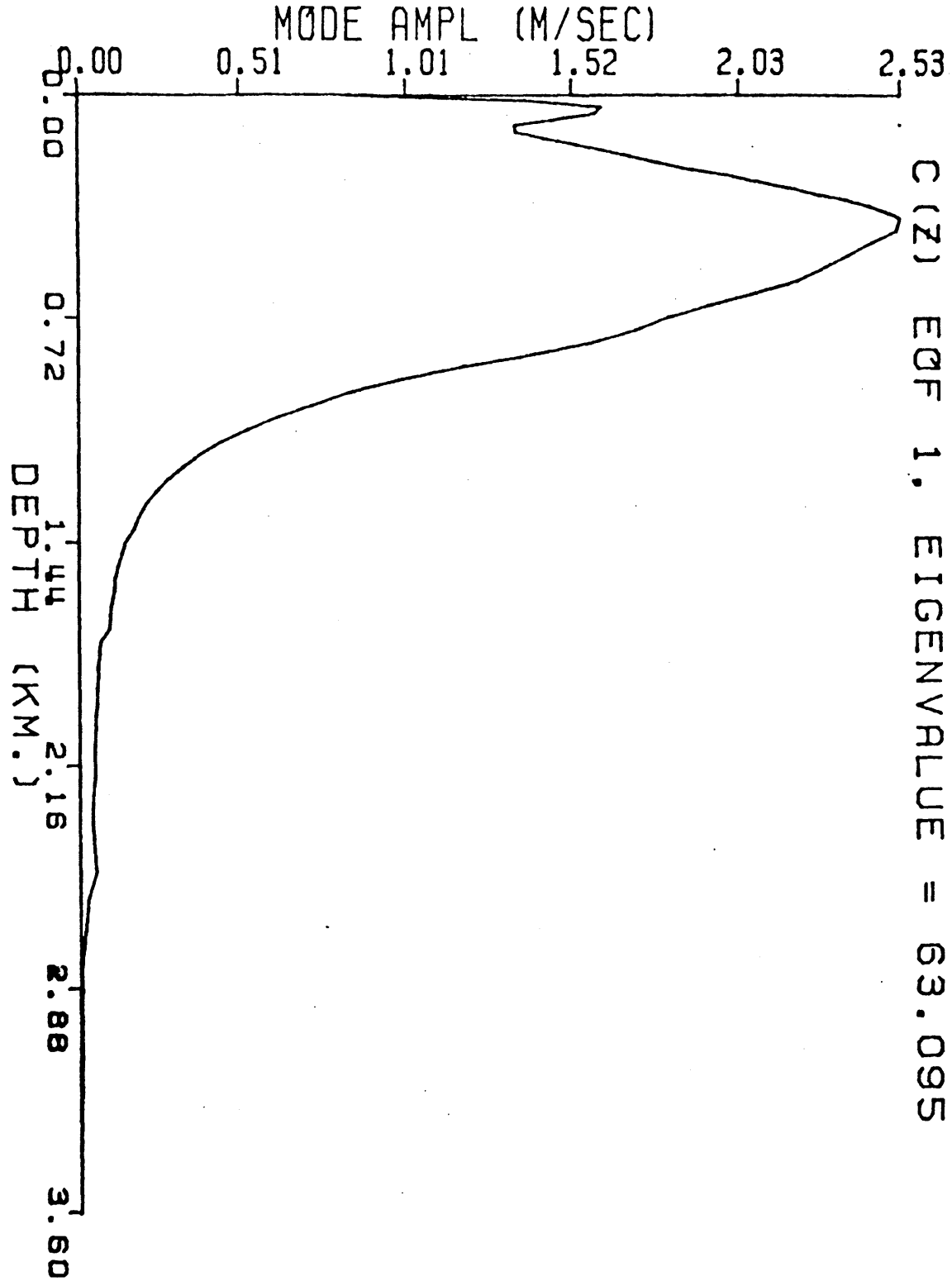
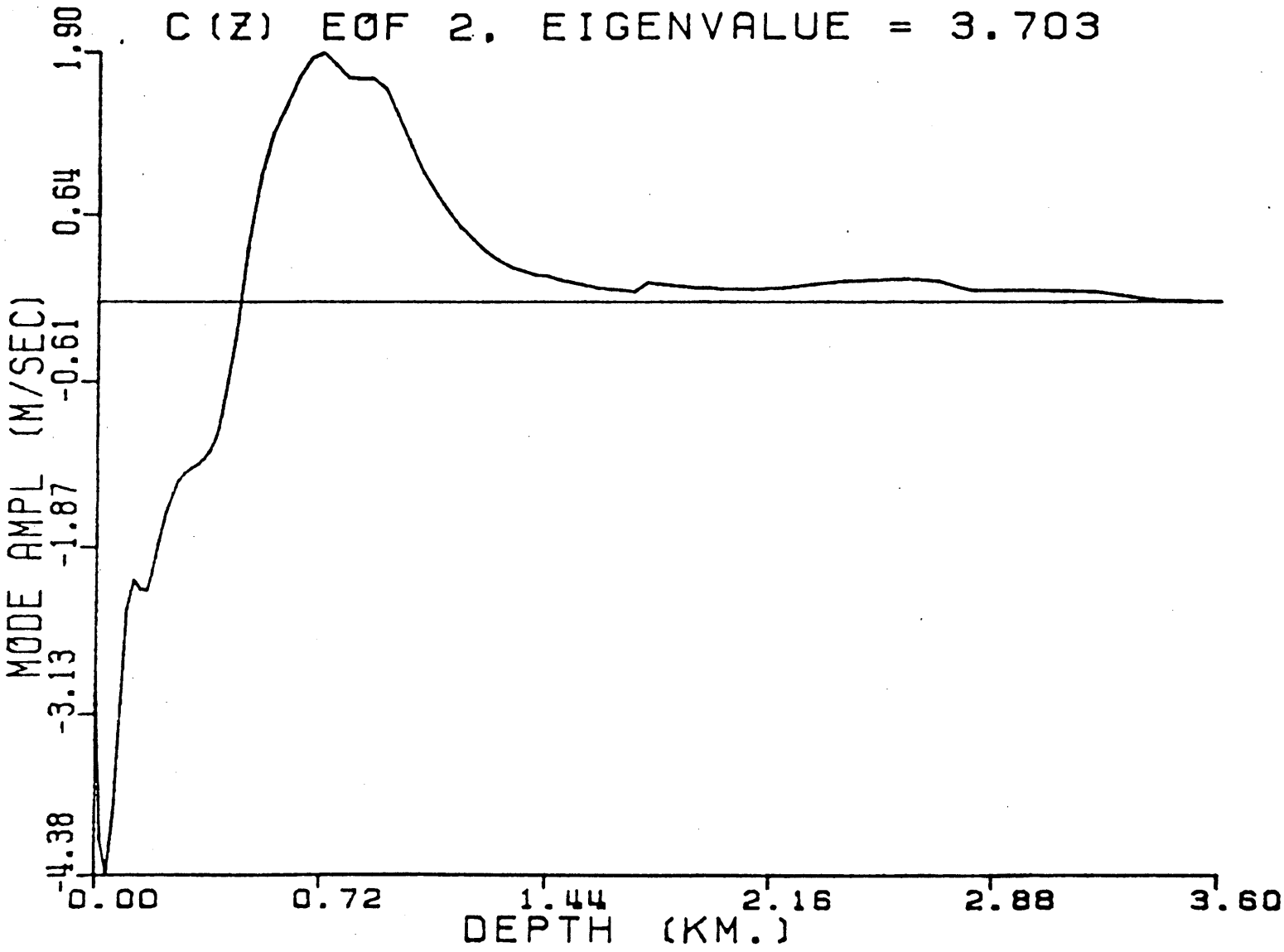
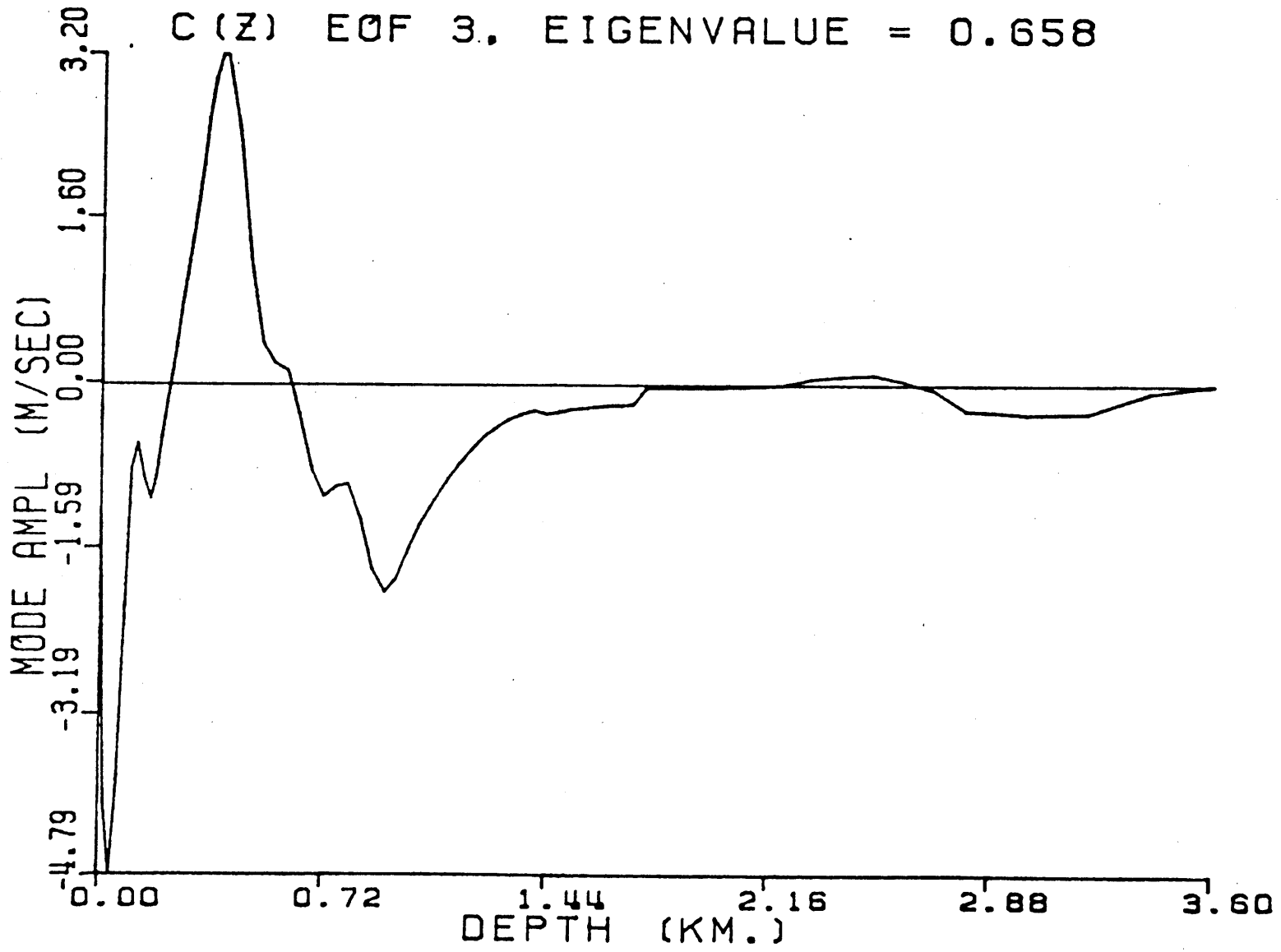


FIGURE 10.17 A,B,C: FIRST 3 MODES FOR THE GULF STREAM SECTION. CALCULATED FROM THE ARCHIVED HYDROGRAPHIC DATA.




C(Z) EOF 2. EIGENVALUE = 3.703
MODE AMPL (M/SEC)
1.90
0.64
-0.61
-1.87
-3.13
-4.38
0.00
0.72
1.44
2.16
2.88
3.60
DEPTH (KM.)


C(Z) EOF 3. EIGENVALUE = 0.658
MODE AMPL (M/SEC)
3.20
1.60
0.00
-1.59
-3.19
-4.79
0.00
0.72
1.44
2.16
2.88
3.60
DEPTH (KM.)

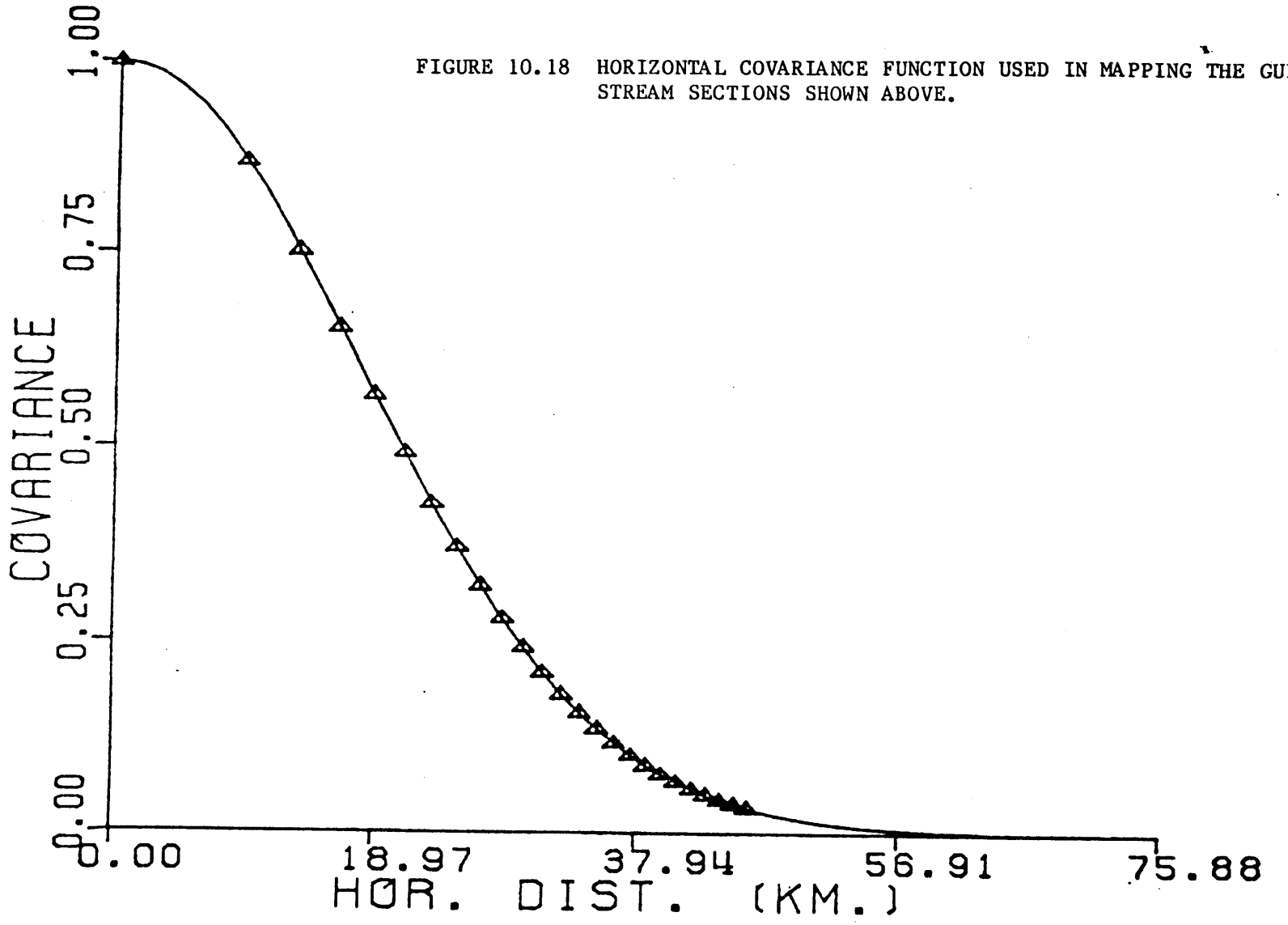


FIGURE 10.18 HORIZONTAL COVARIANCE FUNCTION USED IN MAPPING THE GULF STREAM SECTIONS SHOWN ABOVE.

These simulations were constructed using archived data, but test cases have also been run using a channel model to simulate the Gulf Stream (Rizzoli, Cornuelle, and Haidvogel, 1982). At present, the model has only been used to construct synthetic oceans for raytracing and evalution of the estimators. For the future, however, combining oceanographic measurements with analytical or numerical models is potentially powerful. One example has already been discussed--using a planetary wave basis for the inversions, so that the acoustic data are used only to update the amplitudes and phases of the waves. The more general case, combining a dynamic model (which evolves in time) with data taken periodically, has been considered, in a simple form, by Ghil, et. al. (1982) for the meteorological case.

Ghil used the Kalman filter, which is a technique from control theory in which an estimate of the unknown field, made by a linearized model, for a given time is optimally combined with the data taken at that time, and the resulting field is then used as the basis of the next estimate. The Kalman filter is designed to minimize the squared error between the estimate and the true field, just as in the stochastic inverse, and the time-dependent

stochastic inverse with the proper constraints should reproduce the Kalman filter. The Kalman filter is simple to implement, and is well-understood, but the length of the state vector for a primitive equation or quasi-geostrophic ocean model is perhaps too large to reasonably apply the Kalman filter blindly.

The field of stochastic and deterministic control theory is growing rapidly, and there are many error-minimization algorithms available, depending on the assumptions that are reasonable to make. Future observations of the oceans or atmosphere should be made with these techniques in mind, deciding on the goal of the measurements and choosing a mix of instruments to maximize the resolution of the field or balance under study, subject to economic constraints. If a body of theory is well understood and accepted, it can be used as a substitute for much data if it is incorporated in the estimation procedure.

ACKNOWLEDGEMENT

I would like to thank my thesis advisor, Carl Wunsch, for the freedom and resources he gave me, for his uniformly accurate and important suggestions, and for his many painstaking readings of many quasi-legible versions of this thesis. The Ocean Tomography Group, and especially Bob Spindel and Peter Worcester, labored mightily before turning over all their data to me, and I owe them both thanks and congratulations for their success with an entirely new type of experiment. Barbara Grant taught me all I know about programming, although she naturally doesn't want it known. The discussion of inverse methods owes much to conversations with A. Tarantola during the summer of 1982. Finally, I want to thank my parents for supporting me completely in whatever I decided to do (provided it didn't take too long.)

My support for the first 3 years came from an NSF graduate fellowship, and I was then supported as a research assistant by NSF Grant OCE-8017791.

APPENDIX

DETAILS OF THE PROBABILISTIC ESTIMATION

From completing the square in equation (10), chapter 4 obtained expressions for the covariance matrix of the result of the estimation:

$$\hat{\underline{\underline{C}}}^{-1} = \underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1} \tag{1}$$

$$\hat{\underline{\underline{C}}}^{-1}\hat{\underline{\lambda}} = \underline{\underline{C}}_a^{-1}\tilde{\underline{\lambda}} + \underline{\underline{C}}_T^{-1}\bar{\underline{\lambda}} \tag{2}$$

$$\hat{\underline{\lambda}} = (\underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1})^{-1} (\underline{\underline{C}}_a^{-1}\tilde{\underline{\lambda}} + \underline{\underline{C}}_T^{-1}\bar{\underline{\lambda}}) \tag{3}$$

we also have

$$(\underline{\underline{A}}^{-1} + \underline{\underline{B}}^{-1})\,\underline{\underline{A}}^{-1} = \underline{\underline{B}}(\underline{\underline{B}}+\underline{\underline{A}})^{-1} \tag{4}$$

Applying this to (3), we obtain

$$\hat{\underline{\lambda}} = \underline{\underline{C}}_T(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\tilde{\underline{\lambda}} + \underline{\underline{C}}_a(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\bar{\underline{\lambda}} \tag{5}$$

Using the partitioned inverse, $(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}$ becomes

$$\begin{matrix} (\beta - \underline{\underline{C}}_{pd}\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd})^{-1} & -\beta^{-1}\underline{\underline{C}}_{pd}(\underline{\underline{C}}_o - \underline{\underline{C}}^T_{pd}\beta^{-1}\underline{\underline{C}}_{pd})^{-1} \\ -\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd}(\beta - \underline{\underline{C}}_{pd}\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd})^{-1} & (\underline{\underline{C}}_o - \underline{\underline{C}}^T_{pd}\beta^{-1}\underline{\underline{C}}_{pd})^{-1} \end{matrix} \tag{6}$$

$$\equiv \begin{matrix} (\underline{\underline{C}}_1)^{-1} & -\beta^{-1}\underline{\underline{C}}_{pd}(\underline{\underline{C}}_n)^{-1} \\ -\underline{\underline{C}}_o^{-1}\underline{\underline{C}}^T_{pd}(\underline{\underline{C}}_1)^{-1} & (\underline{\underline{C}}_n)^{-1} \end{matrix} \tag{7}$$

Equation (7) defines β, $\underline{\underline{C}}_o$, C_1, and $\underline{\underline{C}}_n$,

$$\beta \equiv \alpha + C_p \tag{8}$$

$$\underline{\underline{C}}_o \equiv \underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon \tag{9}$$

$$C_1 \equiv \beta - \underline{C}_{pd}\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd} \tag{10}$$

$$\underline{\underline{C}}_n \equiv \underline{\underline{C}}_o - \underline{C}^T_{pd}\beta^{-1}\underline{C}_{pd} \tag{11}$$

Recall that (5) has two parts:

$$\hat{\underline{\lambda}} = \underline{\underline{C}}_T(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\tilde{\underline{\lambda}} + \underline{\underline{C}}_a(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\bar{\underline{\lambda}} \tag{5}$$

The first part multiplies $\tilde{\underline{\lambda}}$;

$$\underline{\underline{C}}_T(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1} =$$

$$\begin{matrix} C_pC_1^{-1} - \underline{C}_{pd}\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd}C_1^{-1} & -C_p\beta^{-1}\underline{C}_{pd}\underline{\underline{C}}_n^{-1} + \underline{C}_{pd}\underline{\underline{C}}_n^{-1} \\ \underline{C}^T_{pd}C_1^{-1} - \underline{\underline{C}}_d\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd}C_1^{-1} & -\underline{C}^T_{pd}\beta^{-1}\underline{C}_{pd}\underline{\underline{C}}_n^{-1} + \underline{\underline{C}}_d\underline{\underline{C}}_n^{-1} \end{matrix} \tag{12}$$

$$= \begin{matrix} (C_1 - \alpha)C_1^{-1} & (\alpha\beta^{-1})\underline{C}_{pd}\underline{\underline{C}}_n^{-1} \\ \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd}C_1^{-1} & (\underline{\underline{C}}_n - \underline{\underline{C}}_\varepsilon)\underline{\underline{C}}_n^{-1} \end{matrix} \tag{13}$$

The second part, multiplying $\bar{\underline{\lambda}}$, is:

$$\underline{\underline{C}}_a(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1} = \begin{matrix} \alpha C_1^{-1} & -\alpha\beta^{-1}\underline{C}_{pd}\underline{\underline{C}}_n^{-1} \\ -\underline{\underline{C}}_\varepsilon\underline{\underline{C}}_o^{-1}\underline{C}^T_{pd}C_1^{-1} & \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_n^{-1} \end{matrix} \tag{14}$$

Thus,

$$\hat{p} = (C_1 - \alpha)C_1^{-1}\tilde{p} + (\alpha\beta^{-1})\underline{\underline{C}}_{pd}\underline{\underline{C}}_n^{-1}\tilde{\underline{d}} + \alpha C_1^{-1}\bar{p} - \alpha\beta^{-1}\underline{\underline{C}}_{pd}\underline{\underline{C}}_n^{-1}\bar{\underline{d}}$$

$$= [(C_1 - \alpha)\tilde{p} + \alpha\bar{p}]C_1^{-1} + \alpha\beta^{-1}\underline{\underline{C}}_{pd}\underline{\underline{C}}_n^{-1}(\tilde{\underline{d}} - \bar{\underline{d}}) \qquad (15)$$

$$\hat{\underline{d}} =$$

$$\underline{\underline{C}}_\varepsilon\underline{\underline{C}}_o^{-1}\underline{\underline{C}}_{pd}^T C_1^{-1}\tilde{p} + (\underline{\underline{C}}_n - \underline{\underline{C}}_\varepsilon)\underline{\underline{C}}_n^{-1}\tilde{\underline{d}} - \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_o^{-1}\underline{\underline{C}}_{pd}^T C_1^{-1}\bar{p} + \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_n^{-1}\bar{\underline{d}}$$

$$= \tilde{\underline{d}} + \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_n^{-1}(\bar{\underline{d}} - \tilde{\underline{d}}) + \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_o^{-1}\underline{\underline{C}}_{pd}^T C_1^{-1}(\tilde{p} - \bar{p}) \qquad (16)$$

In the case where no a priori information about a particular value of p is available, ($\alpha \to \infty$) then

$C_1 \to \beta \to \alpha \to \infty$ and $\underline{\underline{C}}_n \to \underline{\underline{C}}_o$, so that

$$\hat{p} = \bar{p} + \underline{\underline{C}}_{pd}\underline{\underline{C}}_n^{-1}(\tilde{\underline{d}} - \bar{\underline{d}}) \qquad (17)$$

$$\hat{\underline{d}} = \tilde{\underline{d}} + \underline{\underline{C}}_\varepsilon\underline{\underline{C}}_o^{-1}(\bar{\underline{d}} - \tilde{\underline{d}}) \qquad (18)$$

$$= \tilde{\underline{d}} + \underline{\underline{C}}_\varepsilon(\underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon)^{-1}(\bar{\underline{d}} - \tilde{\underline{d}}) \qquad (19)$$

$$= \tilde{\underline{d}} - \hat{\underline{\varepsilon}} \qquad (20)$$

Where $\hat{\underline{\varepsilon}}$ is the optimal estimate of the error in the data:

$$\hat{\underline{\varepsilon}} = \underline{\underline{C}}_\varepsilon(\underline{\underline{C}}_d + \underline{\underline{C}}_\varepsilon)^{-1}(\tilde{\underline{d}} - \bar{\underline{d}}) \qquad (21)$$

$\hat{\underline{\varepsilon}}$ is often referred to as the vector of "residuals" in discussions of inverse methods, and is usually calculated by substituting the estimated field into the

the forward problem, and subtracting the data calculated in this way from the measured data. When the model is continuous, this simple-minded calculation can become quite expensive, and the direct estimate is certainly more rigorous.

The a posteriori probability density function for both the data and the unknowns defines the expected variance of the true value, $\underline{\lambda}$, around the estimate, $\hat{\underline{\lambda}}$:

$$\sigma(\underline{\lambda}) \propto \exp\{-1/2[(\underline{\lambda}-\overline{\underline{\lambda}})^T \underline{\underline{C}}_T^{-1}(\underline{\lambda}-\overline{\underline{\lambda}}) + (\underline{\lambda}-\tilde{\underline{\lambda}})^T \underline{\underline{C}}_a^{-1}(\underline{\lambda}-\tilde{\underline{\lambda}})]\} \qquad (22)$$

This can be put in the form:

$$\sigma(\underline{\lambda}) \propto \exp[-1/2(\underline{\lambda}-\hat{\underline{\lambda}})^T \hat{\underline{\underline{C}}}^{-1}(\underline{\lambda}-\hat{\underline{\lambda}})] \qquad (23)$$

where $\hat{\underline{\lambda}}$ is the maximum likelihood, minimum variance estimate of $\underline{\lambda}$, and $\hat{\underline{\underline{C}}}$ is the estimated covariance around the true value. We are most interested in the expected variance of $\hat{p}(\underline{x},t)$ around $p(\underline{x},t)$,

$$E_p^2 = \langle[p(\underline{x},t) - \hat{p}(\underline{x},t)]^2\rangle, \qquad (24)$$

but it is informative to sketch out the complete $\hat{\underline{\underline{C}}}$. The expression for $\hat{\underline{\underline{C}}}^{-1}$ has already been derived;

$$\hat{\underline{\underline{C}}}^{-1} = \underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1}, \qquad (25)$$

but we need $\hat{\underline{\underline{C}}}$ directly:

$$\hat{\underline{\underline{C}}} = (\underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1})^{-1} \qquad (26)$$

It is possible to take advantage of the partitioning to calculate (26) out as written, but it is more efficient to re-use the identity (4).

$$\hat{\underline{\underline{C}}} \cdot \underline{\underline{C}}_a^{-1} = (\underline{\underline{C}}_a^{-1} + \underline{\underline{C}}_T^{-1})\underline{\underline{C}}_a^{-1} \tag{27}$$

$$= \underline{\underline{C}}_T(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1} \tag{28}$$

so that

$$\hat{\underline{\underline{C}}} = \underline{\underline{C}}_T(\underline{\underline{C}}_a + \underline{\underline{C}}_T)^{-1}\underline{\underline{C}}_a \tag{29}$$

$$= \begin{pmatrix} C_p & \underline{C}_{pd} \\ \underline{C}^T_{pd} & \underline{\underline{C}}_d \end{pmatrix} \cdot \begin{pmatrix} (\underline{\underline{C}}_1)^{-1} & -\beta^{-1}\underline{C}_{pd}(\underline{\underline{C}}_n)^{-1} \\ -\underline{C}_o^{-1}\underline{C}^T_{pd}(\underline{\underline{C}}_1)^{-1} & (\underline{\underline{C}}_n)^{-1} \end{pmatrix} \cdot \begin{pmatrix} \alpha & 0 \\ 0 & \underline{\underline{C}}_\varepsilon \end{pmatrix}$$

$$= \begin{pmatrix} C_p & \underline{C}_{pd} \\ \underline{C}^T_{pd} & \underline{\underline{C}}_d \end{pmatrix} \cdot \begin{pmatrix} \alpha(\underline{\underline{C}}_1)^{-1} & -\beta^{-1}\underline{C}_{pd}(\underline{\underline{C}}_n)^{-1}\underline{\underline{C}}_\varepsilon \\ -\alpha\underline{C}_o^{-1}\underline{C}^T_{pd}(\underline{\underline{C}}_1)^{-1} & (\underline{\underline{C}}_n)^{-1}\underline{\underline{C}}_\varepsilon \end{pmatrix} \tag{30}$$

The product requires much space to write out, but we are most interested in the top left element of $\underline{\underline{C}}$, which is the variance of the estimated value of p around the true value:

$$E_p{}^2 = \alpha C_p(\underline{\underline{C}}_1)^{-1} - \alpha\underline{C}_{pd}\underline{C}_o^{-1}\underline{C}^T_{pd}(\underline{\underline{C}}_1)^{-1} \tag{31}$$

$$= \alpha\cdot(C_p - \underline{C}_{pd}\underline{C}_o^{-1}\underline{C}^T_{pd})\cdot(C_p + \alpha - \underline{C}_{pd}\underline{C}_o^{-1}\underline{C}^T_{pd})^{-1} \tag{32}$$

For completeness, I will write out the bottom right-hand element of $\hat{\underline{\underline{C}}}$, which describes the variance of the estimated data values around the true values:

$$E_d^2 = \underline{\underline{C}}_d \, (\underline{\underline{C}}_n)^{-1} \underline{\underline{C}}_\varepsilon \; - \; \underline{C}^T_{pd} \, \beta^{-1} \underline{\underline{C}}_{pd} (\underline{\underline{C}}_n)^{-1} \underline{\underline{C}}_\varepsilon \tag{33}$$

$$= (\underline{\underline{C}}_d \; - \; \underline{C}^T_{pd} \, \beta^{-1} \underline{\underline{C}}_{pd}) \cdot (\underline{\underline{C}}_n)^{-1} \underline{\underline{C}}_\varepsilon \tag{34}$$

$$= (\underline{\underline{C}}_n \; - \; \underline{\underline{C}}_\varepsilon) \cdot (\underline{\underline{C}}_n)^{-1} \underline{\underline{C}}_\varepsilon \tag{35}$$

$$= \underline{\underline{C}}_\varepsilon \; - \; \underline{\underline{C}}_\varepsilon \cdot (\underline{\underline{C}}_n)^{-1} \underline{\underline{C}}_\varepsilon \tag{36}$$

Note the exact symmetry with the estimate of the model field uncertainty.

REFERENCES

The Ocean Tomography Group: Scripps Institution of Oceanography: M. Brown, R. Knox, W. Munk, J. Spiesberger, P. Worcester; Woods Hole Oceanographic Institution: R. Spindel, D. Webb; M.I.T.: B. Cornuelle, R. Heinmiller, C. Wunsch; University of Michigan: T. Birdsall, K. Metzger; NOAA-AOML: D. Behringer; Draper Laboratory: J. Dahlen.

Aki, K., and P. G. Richards: Quantitative Seismology, Theory and Methods. W. H. Freeman, San Francisco, 1980.

Backus, G.E. and J.F. Gilbert: Numerical applications of a formalism for geophysical inverse problems. Geophys. J. Roy. astr. Soc., 13, 247-276, 1967.

Backus, G. E., and F. Gilbert: The resolving power of gross Earth data. Geophys. J. R. astr. Soc., 16, 169-205, 1968.

Backus, G. E., and F. Gilbert: Uniqueness in the inversion of inaccurate gross Earth data. Phil. Trans. R. Soc. London Ser. A, 266, 123-192, 1970.

Birdsall, T.G.: On understanding the matched filter in the frequency domain. IEEE Transactions on Education, 19, 168-169, 1976.

Bretherton, F., R.E. Davis, and C.B. Fandry: A technique for objective analysis and design of experiments applied to MODE-73. Deep Sea Res., 23, 559-582, 1976.

Bretherton, F., and J.C. McWilliams: Estimations from irregular arrays. Rev. Geophys. Space Phys. 18, 789-812, 1980.

Brown, M.: Linearized travel time, intensity, and waveform inversions-A comparison. J.A.S.A., submitted, 1983.

Brown, M.: application of the WKBJ Green's function to acoustic propagation in horizontally stratified oceans. J. Acoust. Soc. Am., 71, 1427-1432, 1982.

Carter, E.F. and A.R. Robinson: Synoptic maps of the main thermocline from POLYMODE XBTs: A time series via space-time objective analysis, unpublished manuscript, March, 1983.

Cornuelle, B.D.: Acoustic Tomography. IEEE Trans. Geosci. Rem. Sens., GE-20, 326-332.

Clark, J.G. and M. Kronengold: Long-u eriod fluctuations of CW signals in deep and shallow water. J. Acoust. Soc. Amer., 56, 1071-1083, 1974.

Eisler, T.J., R. New, and D. Calderone: Resolution and variance in acoustic tomography. J. Acoust. Soc. Am. 72, 1965-1977, 1982.

Ewing, M. and J.L. Worzel: Long-range sound transmission. Geol. Soc. Am. Mem., 27, 1948.

Flatte, S.M., R. Dashen, W.H. Munk, K.M. Watson, and F. Zachariasen: Sound transmission through a fluctuating ocean. Cambridge University Press, Cambridge, 1978.

Flierl, G.R.: Models of vertical structure and the calibration of two-layer models. Dyn. Atm. Oc., 2, 341-381, 1978.

Ghil, M., et.al.: Applications of estimation theory to numerical weather prediction. In: Dynamic Meteorology data assimilation methods, Bengtsson, Ghil, and Kallen, eds., Springer-Verlag, New York, 1981.

Hamilton, G.R.: Time variations of sound speed over long paths in the ocean. In: International workshop on low-frequency propagation and noise, Woods Hole, Massachusetts, Oct. 14-19, 1974, pp. 7-30, 1977.

Hamilton, K.G., W.L. Siegmann, and M.J. Jacobson: Simplified calculation of ray-phase perturbations due to ocean-environmental variations. J. Acoust. Soc. Am., 67, 1193-1206, Apr. 1980.

Hua, B.-L., and W.B.Owens, Unpublished manuscript, 1983.

Lanczos, C.: Linear Differential Operators. Van Nostrand, New York, 1961.

Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 591-612, 1970.

McWilliams, J.C.: Maps from the Mid-Ocean Dynamics Experiment: Part II. Potential vorticity and its conservation. J. Phys. Ocean., 6, 828-846, 1976.

McWilliams, J.C. and W. Owens: Estimation of spatial covariances from the MODE esperiment. NCAR Tech. Rep. No. 115+STR, 25 pp., 1976.

Mercer, J.A., and J.R. Booker: Long-range propagation of sound through oceanic mesoscale structures, J. Geophys. Res. 88, 689-700, 1983.

MODE Group: Mid-ocean dynamics experiment. Deep-Sea Research, 25, 859-910, 1978.

Munk, W.H. and P.F. Worcester: Monitoring the ocean acoustically. In: Science, technology, and the modern navy-thirtieth anniversary 1946-1976, (ONR-37), Office of Naval Research, Arlington, VA, pp. 497-508; also appears as: Weather and climate under the sea-the Navy's habitat. In: Science and the future navy-a symposium, 30th Anniversary Volume, Office of Naval Research, National Academy of Sciences, Washington, D.C., pp. 42-52, 1976.

Munk, W. and C. Wunsch: Ocean acoustic tomography: a scheme for large-scale monitoring. Deep-Sea Res., 26A, 123-161, 1979.

Munk, W.: Horizontal deflection of acoustic paths by mesoscale eddies. J. Phys. Ocean., 10, 596-604, 1980.

Munk, W. and C. Wunsch: Observing the ocean in the 1990s. Phil. Trans. R. Soc. Lond. A 307, 439-464, 1982.

Munk, W. and C. Wunsch: Acoustic tomography: rays and modes. Rev. Geophys. Space Phys., 1983, in press.

Officer, C.B.: Introduction to the Theory of Sound Transmission with Application to the Ocean. New York:McGraw-Hill, 1958.

Papoulis, A.: Maximum entropy and spectral estimation: A review. IEEE Trans. Acoust. Speech and Sig. Proc., ASSP-29, 1176-1186, 1981.

Parker, R.L.: Understanding Inverse Theory. Ann. Rev. Earth Planet. Sci., 5, 32-64, 1977.

Pedlosky, J.: Geophysical Fluid Dynamics, New York: Springer-Verlag, 1980.

Porter, R.P., R.C. Spindel, and R.J. Jaffee: CW beacon system for hydrophone motion determination. J. Acoust. Soc. Amer., 53, 1691-1699, 1973.

Richman, J.G., C. Wunsch, and N.G. Hogg: Space and time scales of mesoscale motion in the western North Atlantic. Rev. Geophys. and Space Phys., 15, 385-420, 1977.

Malanotte-Rizzoli, P., B.D. Cornuelle, and D.B. Haidvogel: Gulf Stream Acoustic Tomography: Modeling Simulations. unpublished manuscript, Oct. 1982.

Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal, 27, 623-656, 1948.

Spiesberger, J.L., R.C. Spindel, and K. Metzger: J. Acoust. Soc. Am., 67, 2011, 1980.

Spindel, R.C., R.P. Porter, and D.C. Webb: IEEE J. Oceanic Engng., OE-2, 331, 1977.

Spindel, R.C.: IEEE Trans. Acoust. Speech Signal Processing, ASSP-27, 723, 1979.

Spindel, R.C.: Proc. IEEE Electron. Aerospace Convent., 80 CH, 1587-4AES, 165, 1980.

Spindel, R.C., and J.L. Spiesberger: Multipath variability due to the Gulf Stream. J. Acoust. Soc. Am., 69, 982-988, 1981.

Stommel, H. and F. Schott: The beta spiral and the determination of the absolute velocity field from hydrographic station data. Deep Sea Res., 24, 325-329, 1977.

Steinberg, J.C. and T.G. Birdsall: Underwater sound propagation in the Straits of Florida. J. Acoust. Soc. Amer., 39, 301-315, 1966.

Swindell, W. and H.H. Barrett: Computerized tomography: taking sectional X-rays. Physics Today, 32-41, 1977.

Tarantola, A. and B. Valette: Generalized nonlinear inverse problems solved using the least squares criterion. Rev. Geophys. Space Phys., 19, 219-232, 1982a.

Tarantola, A. and B. Valette: Inverse problems = Quest for information. J. Geophys., 50, 159-170, 1982b.

Van Trees, H.: Detection, Estimation and Modulation Theory, Part 1. 697 pp., J. Wiley, 1968.

Warren, B.A., and C. Wunsch (ed.): Evolution of Physical Oceanography. Scientific Surveys in Honor of Henry Stommel. 623 pp. The MIT Press, Cambridge. 1981.

Webb, D.C.: Proc. Oceans, '77 Conf. Rec. MTS-IEEE, 2, 44B, 1977.

Worcester, P.F.: Reciprocal acoustic transmission in a mid-ocean environment. J. Acoust. Soc. Amer., 62, 895-905, 1977.

Worcester, P.F.: An example of ocean acoustic multipath identification at long range using both travel time and vertical arrival angle. J. Acoust. Soc. Amer., 70, 1743-1747, 1981.

Worcester, P.F., and Cornuelle, B.D.: Ocean acoustic tomography: currents. Proceedings of the IEEE Second Working Conference on Current Measurement, IEEE, 1982.

Wunsch, C.: Determining the general circulation of the oceans: a preliminary discussion. Science 96, 871-875, 1977.

Wunsch, C.: The North Atlantic general circulation west of 50°W determined by inverse methods. Rev. Geophys. and Space Phys., 16, 583-620, 1978.

Wunsch, C.: Low-frequency variability of the sea. In Evolution of Physical Oceanography, B.A. Warren and C. Wunsch (ed.). The MIT Press, Cambridge, Mass., 623 pp., 1981.

Zlotnicki, V., B. Parsons, and C. Wunsch: The inverse problem of constructing a gravimetric geoid. J. Geophys. Res., 87, 1835-1848, 1982.

Zlotnicki, V.: The oceanographic and geiodal components of sea surface topography. Ph.D. Thesis, M.I.T., Cambridge, Mass, 1983.