II

Present state and prospects of data processing

by

Y. ADAM

(Based on work by Y. ADAM, J.P. FOGUENNE, C.J. FRANKIGNOUL, A. FOLLENTIER, Y. RUNFOLA, M. VANDENBOSSCHE)

1.- Data treatment : an overview

In the frame of a research program built around a general mathematical model, which is fed by huge amounts of data coming from a great number of laboratories and automatic data acquisition systems, the disponibility of a large and easily accessible data base is obviously a useful tool for the scientists.

The only way to make a compromise between the terrific amount of available data and the accessibility of this information to the scientist, is to build a computerized data base structured with the help of the mathematical model, to develop a set of management and analysis computer programs which will free the scientist from the painful tasks of retrieving data from other laboratories, handplotting his own measurements, performing boring statistical computations and give him a practical tool to easily edit and correct his data, compare them to results of theoretical simulations.

The great variety of data makes this task uneasy; data from automatic acquisition systems (like buoys, current meters) are usually time series of regularly spaced samples; data issued from *in situ* measurements through laboratories are usually batches of samples collected at different positions and times during a cruise or a survey; some data are counts (*e.g.* number of bacteria), others are decimal numbers (*e.g.* nutrient concentration), some are even synoptic codes (*e.g.* meteorological observations).

One must thus, first design a rather flexible format of data storage, to cope with the variety of data one is faced with; then develop general routines to format, retrieve, access and select every kind of data, each routine being built in such a way that it can easily be tailored by the user to treat the kind of data he is interested in.

2.- Present state of the data base and its associated software

We have described the goal we are aiming at. It must be emphasized that we are still on the first stages of design and development of the data base. The scientists involved in data treatment and information processing have up to now been too busy with other fundamental research (like mathematical models design and optimization, numerical tools creation) or with technological tasks (data acquisition systems implementation) to dedicate their time to the data base.

However, some files already exist and are easily accessible on computer readable storage media (mainly magnetic tape) and an elementary software is able to store, access, edit and visualize these data. The software has no unity and consists in several programs chains, each being specialized in the treatment of a particular type of data. Up to now, mathematical or display tools developed for one type of data cannot be immediately adapted to another type; moreover, computer programs have been designed for particular types of computers and nothing has been made to insure the portability, neither of the software, nor of the data (let us remind that two types of computers, at least, are available in the frame of the program for the data processing : a HP2100A and an IBM 370/158).

The automated data chains handle

1) buoy data,

2) current meter data,

3) meteorological data.

All are time series (data every 15 minutes for current meters, every hour for buoy, every 3 hours for meteo).

The first chain has been implemented in Ostende, the two others in Liège.

2.1.- Meteorological data

The meteorological data handling chain is described in Appendix I. It is the most sophisticated of all because of the complex structure of the time series (28 samples and 14 variables at each time step).

2.2.- Buoy data

The flow diagram of the buoy data handling chain is shown on figure 1.

The first step of the chain is the data acquisition itself, described elsewhere, together with the signal transmission and reception.

In the second step raw data are fed into the computer on paper tape support; the first program (DASS1) stores them on a magnetic tape in a fixed format and updates the catalog of the tape. No correction and no editing are performed on this stage.

In the third step the MT containing the raw data may be

- copied by program MCØPY, for the sake of safety,
- listed by program MLIST,
- edited by program DASS3 (scratches wrong files) and MEDIT (eventually corrects erroneous date or values).

Every file is then tested by program PURIF, which corrects data where erroneous ASCII characters appear, and rewritten at the same place on MT.

In the fourth step a series of programs CH01, CH02, ... decode the information given by every channel of the buoy (CH01 translates channel 1 data, CH02 translates channel 2 data, ...). Results are stored on disc.

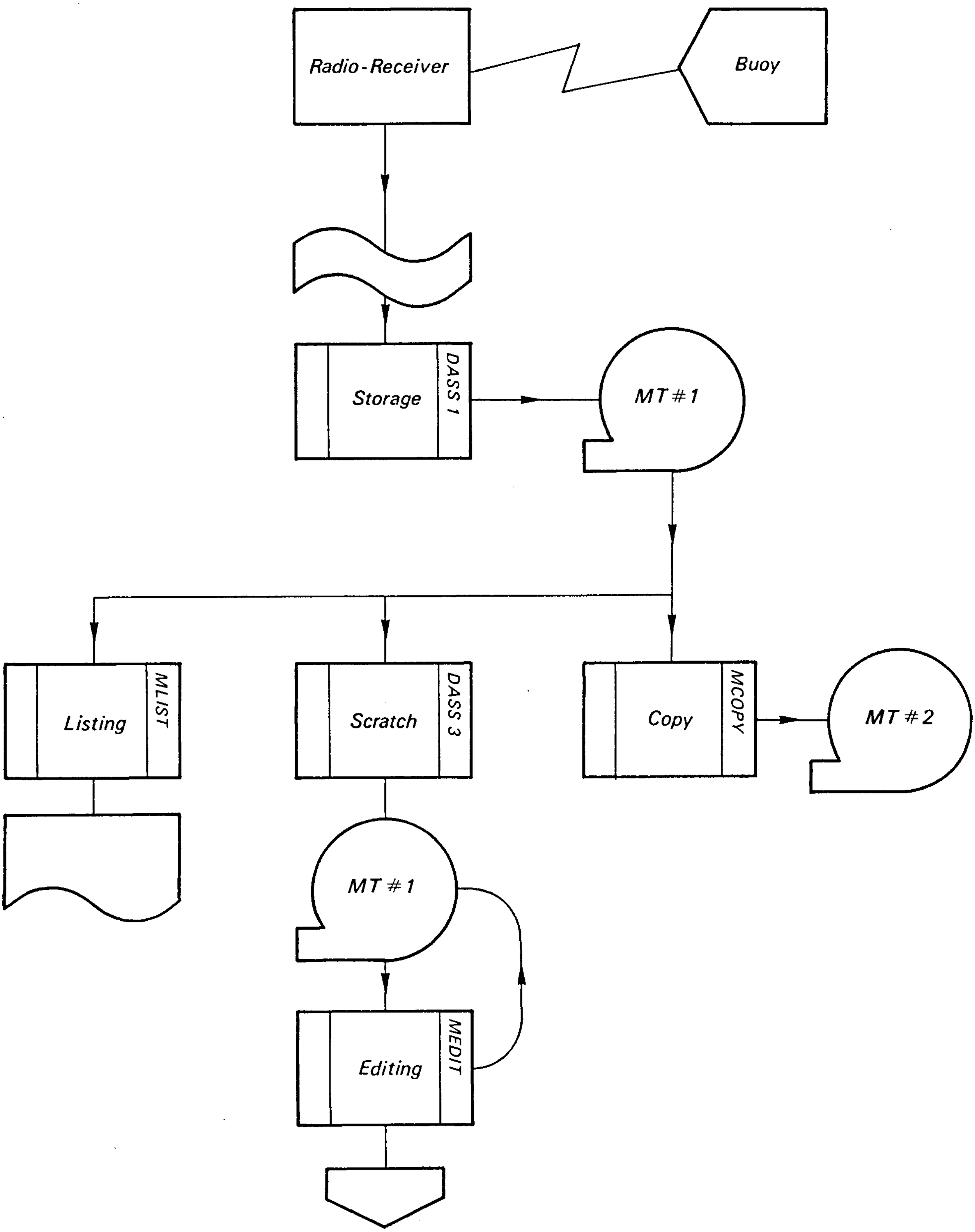

Radio-Receiver
Buoy
Storage
DASS 1
MT # 1
Listing
MLIST
Scratch
DASS 3
Copy
MCOPY
MT # 2
MT # 1
Editing
MEDIT

fig. 1.

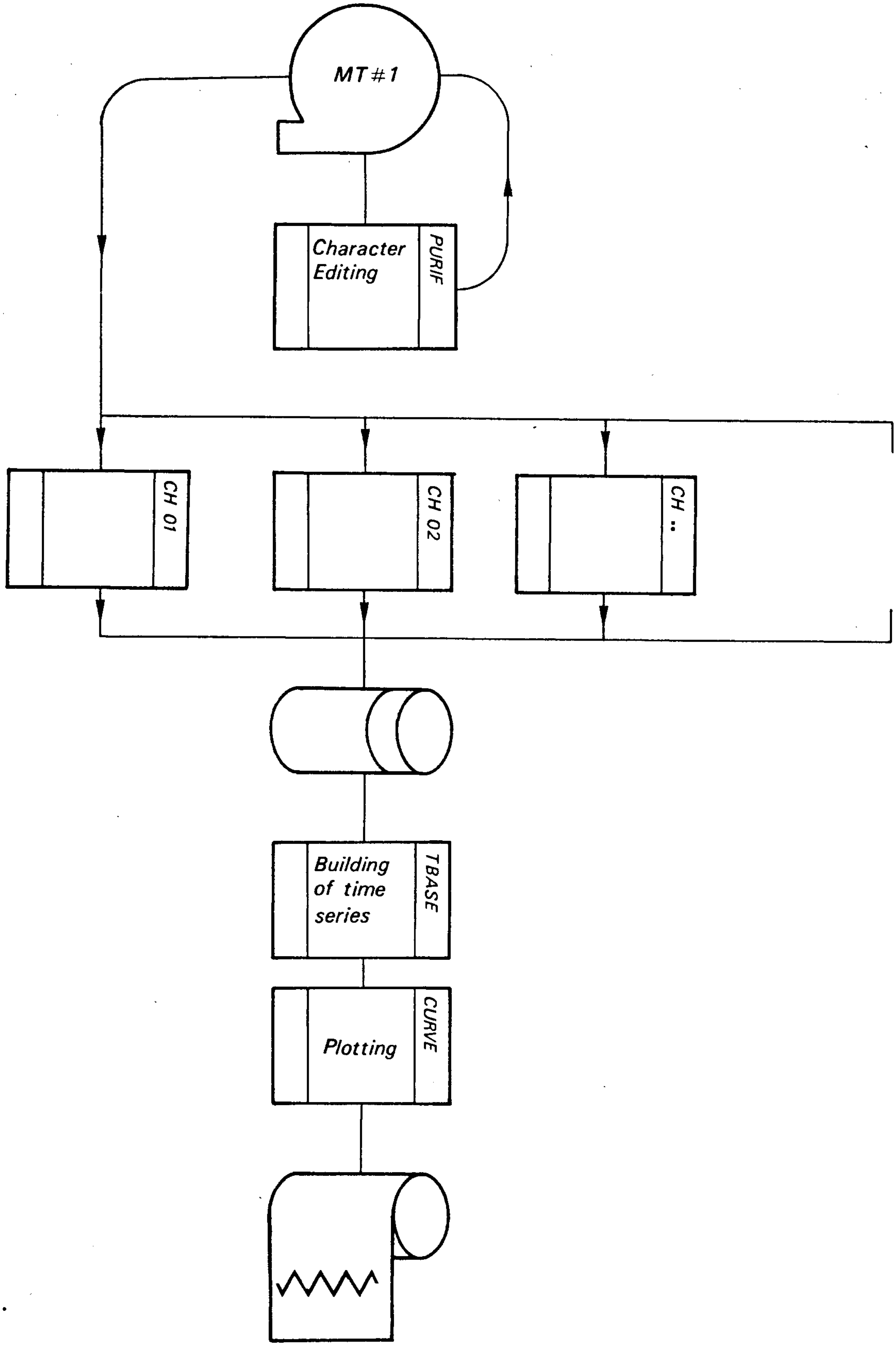

MT#1
Character Editing
PURIF
CH 01
CH 02
CH ..
Building of time series
TBASE
Plotting
CURVE

fig. 1.

Program TBASE builds then from that information time series, which can be plotted by program CURVE.

This software is rather simple and operates fast. Its main drawbacks are :

1) it is exclusively buoy oriented,

2) the storage format is rigid,

3) no actual treatment is performed.

It can be improved in two ways :

1) the first step can be replaced by on line data acquisition. This will be worked out in the near future : signals from the buoy will be immediately stored on disc files (acting as buffers) and then transferred to MT;

2) the last step must be completed by a program which stores the data on a mass storage device in a standard format we will describe later. Buoy information could then be treated by a standard software still to be developed but common to all types of data.

2.3.- Current meter data

This handling chain is the less sophisticated. Current meter data are received in different ways, according to the type of instrument.

VACM data arrive in a ready-to-plot-and-analyse form from WHOI when they have been decoded, edited and pretreated (interpolation of missing values) on a MT written in EBCDIC.

NBA raw data arrive on a paper tape in ASCII characters. No edition or pretreatment has been performed.

Plessey data arrive in Plessey code form (on paper tape or MT); they have to be translated in ASCII or EBCDIC before being available to processing programs.

For each kind of instrument, there exists a program which builds the time series, edits them and stores them on disk-files which serve as input for plotting and analysing programs (namely a Fast Fourier Transform).

3.- Guidelines for the building of the data base and processing software

The utility of a data base and related software in the frame of an interdisciplinary research program has been stressed. We dare not call this structure a data bank, because we do not intend to achieve a very sophisticated software, completely transparent to the users, like some systems used in financial, commercial and banking environments. The software will not be extended so far for two main reasons :

1) environmental data (like oceanographic or meteorological data) present a greater variety of types and structures than business data and need more complex processing. It seems to us to be overweening to design a very sophisticated software which would be computer time greeding with the limited means (material and human) we dispose of;

2) the data base and processing software is intended to be used, not by bank clerks, but by scientists with a minimum experience of computer systems who are assumed to know which kind of information they exactly need and how they want to handle the data. A minimum staff of software specialists would however be necessary to help the scientists in assembling their program chains and to advise them on the choice of computing tools.

This solution has already been experimented in several oceanographic research centers, like WHOI, University of Southampton, and BNDO at the *Centre océanologique de Bretagne*.

The system we aim at will however be flexible enough to enable the storage, retrieval, selection and processing of present and future data sets, whatever be their origin and structure. It will also be designed to facilitate the work of the scientist by giving him standard tools for the analysis of his material and for the validation and comparison of mathematical models and to be easily extended with programs for specific purposes.

Finally, the data storage and the software are thought to be portable (*i.e.* able to be implemented with the least possible modifications) on most families of computer systems in order that the effort made for designing them will benefit the largest number of research groups. This implies that the structure of the programs themselves and of the data records be compatible with the great majority of the computers used in scientific research centers. We explain later how this can be achieved, for both the software and the storage format.

4.- Requirements for the portability of the storage format and of the software

Computer specialists know very well that almost everything that has been written by any computer system can be read by any other, provided that the 'translation' software exists. The portability of the software is less evident : not all languages are implemented on all computer systems and translating a package from one language to another often needs a lot of time and money. Generally, oceanographic research centers do not have the men needed to work out such a conversion. That is why the software package we are developing is designed to be used, with only a few slight modifications relative to input/output operations, on the great majority of general purpose or scientific computers.

Due to the internal structure of the routines, to the flexible shape of the storage format and to the fact that it is hardly impossible to write data processing programs fully independant from the hardware, the following hardware and software features are needed :

1) a direct memory access to a magnetic tape and disk is available,

2) single precision real words are stored in 2 integer words,

3) a compiler exists for ANSI Fortran IV,

4) records of variable length can be read and written on magnetic tape (eventually through a limited set of special assembly language routines).

ANSI Fortran IV is the most widespread computer language for scientific and technical applications : it is thus the best language to write a scientific package.

As to condition 2), it means that the core storage of a single precision floating point number is twice the core storage of an integer word. That is what happens automatically on 16-bits or 24-bits word oriented computer, and what can be forced by software on 32-bits or 60-bits word-oriented machines.

The magnetic tape format is portable in the sense that the structure of the records on magnetic tape is the same for all computer systems. It is intended as a data processing and a mass storage format, not as a data exchange format[1]. As such, it uses binary recording mode for compactness and input/output performance, *i.e.* the data on mag-tape are an exact copy of their core storage representation.

A data exchange format should be written in interchange code as EBCDIC or ASCII. Such a code is unsuitable for fast processing and is less compact than binary code.

However, despite the binary recording mode, tapes written by one computer can easily be read by another under several conditions.

1) magnetic tape specifications are compatible on both systems (*i.e.* same number of tracks, same density, same encoding technique, same shape of end-of-record and end-of-file marks);

2) both machines are either 16-bits (word) or 32-bits (byte) oriented; both machines are 24-bits word oriented; both machines are 60-bits word oriented (computers are then called "word compatible");

3) routines exist for the translation of internal formats for real numbers and characters.

Provision is made in the storage format for defining the internal code used to write the numbers and the characters in order that the appropriate routines (condition 3) can be selected. The keys defining these codes can be decoded by any computer word-compatible with the source computer because they are positive integer numbers that have the same format in all systems with the same integer word length.

1. Data exchange formats are now being developed by international scientific organizations.

5.- General description of the software

The software being presently on the first stages of development, it cannot be described in detail. It can be defined as an extended subset of the WHOI standard buoy format package, simpler than the latter one under some aspects. The original is being modified to be more general (to treat a greater variety of data structures) and more portable; it will be extended to meet mathematical modelling requirements.

It will consist in :

1) a set of general purpose routines to read, write date records, label and retrieve data files (both on magnetic tape and on disk) and to transfer the data between mass storage and core memory (in arrays accessible to the user), in both directions;

2) a set of data handling programs (using the general purpose routines) to select, edit, list and plot the data in various ways, depending on their structure;

3) a set of mathematical programs to analyse the data and to compare and fit models to experimental features.

6.- The data storage standard format

The format is self-describing : each data file contains in itself every information needed to be read properly whatever be the number of data variables, of samples, their nature, their structure (continuous function of a variable or not), their record type, and the number of records in the file. Magnetic tapes written in standard format also contain at the beginning of the data, an identification file and, after the last data file, an end-of-data file.

Each data file consists of :

1) two label records,

2) data records (in any number).

The first label record has a fixed length and defines :

- the keys for binary data conversion,
- the file creation date,

- the name of the file,
- the number of variables,
- the number of samples,
- the data origin date,
- the type of the second label.

The second label record defines :

- the position where the data were taken (latitude/longitude or other information),
- sampling parameters,
- the initial value of the continuous sampling variable,
- for every variable :
    - its name,
    - its units,
    - the measurement code,
    - the instrument code,
    - the serial,
    - the type in which the data are recorded (integer, real, synoptic),
    - the depth or pressure (if constant),
    - three attributes.

A data record consists of :

- a data file name (to check whether the read data are those described in the first label record),
- the number of samples in the preceeding records,
- the number of samples in the record itself,
- the initial value of the continuous variable,
- the data.

The detailed structure of each record type is described in Appendix II. Only the structure for 16-bits or 32-bits computers is given.

Appendix I — Specification of the wind field for real-time processing and forecasting

I.1.- Introduction

In a preliminary report [Frankignoul (1971)], the present state of knowledge on wind stress determination was reviewed. Approximate formulae for the drag coefficient in different wind regimes were given, thus allowing to calculate from wind data the surface interaction terms for the ecological model of the Southern North Sea [Nihoul (1973)]. Preliminary information on the practical problem of collecting the required meteorological data was also given.

Arrangements have been made with the *Régie des Voies Aériennes* for real-time transmission via telex of all relevant meteorological data. Regular transmission to Liège University began in May 1973. The system that has been worked out for the specification of the wind field for real-time processing and forecasting is now operational. It is described in the present report. For illustration, computer plots of the observed and geostrophic wind field during the first three days of the JONSDAP experiment in September 1973 are reported. Mention is made of further improvements that will be permitted by an extensive study of the recorded data.

I.2.- Description of the available meteorological data

A number of regular weather stations have been selected for real-time specification of the wind field. These are mainly coastal stations and anchored light vessels within the zone comprised between 50 ° and 54 °N , 0 ° and 5 °E . Location and nature of the stations which are operational now is given in figure 2. A large number of coastal stations have been retained so far, although the relevance of their measurements to marine conditions is often questionable, due to considerable change of surface roughness. Work is undertaken to determine and eliminate those coastal stations that give meteorological data which are too influenced by the disturbing presence of land. Synoptic data is emitted

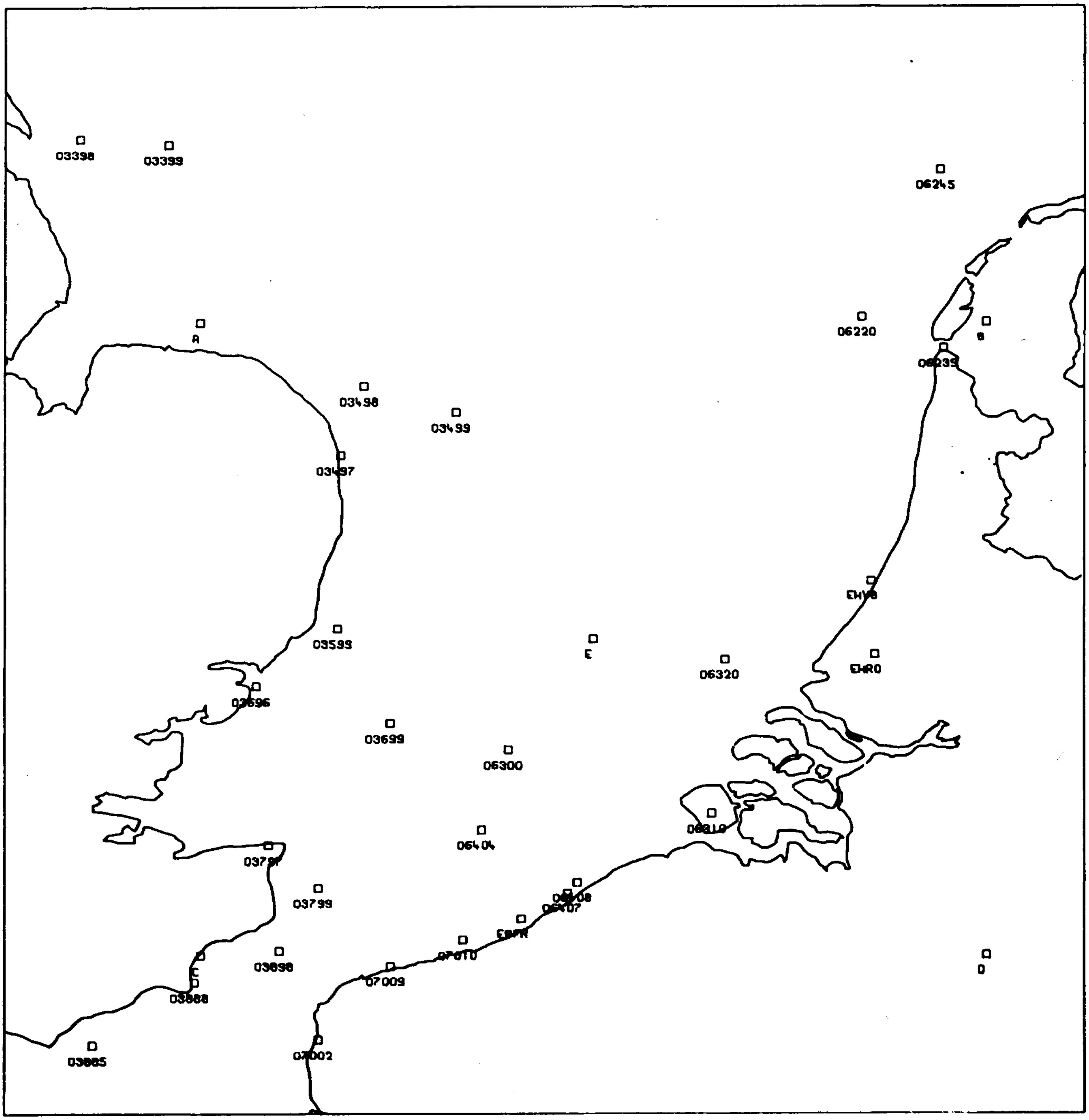


fig. 2.

by most stations every three hours, although some of them show much irregularity. Parameters retained for use in the mathematical model are listed in table 1.

Table 1

List of stored synoptic data

This table shows the various information stored on magnetic tape after the second data handling program. Some stations give their information in "english code", others in "Synop code", others in "Metar code". The set of information varies slightly with the used code. In order to store the information in a standard format, inexistant information is replaced by a dummy code (9999). Some stations, although they use a particular code, do not send all the usual information. The asterisks in the table point the information not always in the messages. A cross denotes an ever-existing information.

Type of data	Unit	English code	Synop code	Metar code
Station name	-	X	X	X
Surface horizontal visibility	meters	X	X	X
Wave height	half-meters	*	9999	9999
period	seconds	*	9999	9999
direction of propagation	tens of degrees	X	9999	9999
Temperature air	degrees °C	X	X	*
sea	degrees °C	X	9999	9999
dew-point	degrees °C	9999	X	*
Barometric pressure	tenths of millibars	*	X	*
Wind (at 10 m height) speed	knots	X	X	X
direction	tens of degrees	X	X	X
Weather present	coded	X	X	9999
past	coded	X	X	9999
Barometric trend	coded	9999	X	9999
Caracteristic of barometric trend	coded	9999	X	9999

Geostrophic wind inferred from pressure maps is also transmitted every three hours, at five locations : 51 °N , 05 °E ; 51° N , 05 °E ; 52 °N , 03 °E ; 53 °N , 01 °E and 53 °N , 05 °E . So far this wind is hand-analysed by experienced meteorologists. In a later stage, the calculation will be computerized at the R.V.A.

Twelve hourly geostrophic wind forecasts at the five locations are similarly transmitted every six hours. More extended forecasts might be used in case of urgency but they are not very reliable at present.

Information coming from ships at random positions and times cannot be handled so far but this is felt to be unimportant, due to the dense distribution of regular stations and the poor quality of ship recorded meteorological data. On the other hand, data from moored buoys is very reliable. A system will be set up to incorporate such data in case of long term mooring. Work will begin on this problem as soon as the first Belgian buoy becomes operational.

I.3.- Data handling

The data is transmitted every three hours via telex from the R.V.A. It is recorded automatically on punched tape in telex code. This punched tape is then stored on a magnetic tape in a chronologic sequence. At regular time intervals (every week or every day, if there is some urgent need) the first program of the data handling chain transfers the results :

a) on another magnetic tape which will be interpreted later;

b) on a listing which provides information for subsequent treatment.

This latter magnetic tape is the input of the second program which interprets the coded messages issued in different codes from the various field stations into numeric values usable by the following programs.

Output of this second step is :

1) data sets on magnetic tape which are stored in chronologic order; there are three kinds of these :
   - data sets containing the results of measurements in field stations;
   - data sets containing 'measured' geostrophic wind at five points;
   - data sets containing predicted geostrophic wind at the same five points.

   Every data set is defined by the time (year, month, day, hour) at which the measurements were made. A data set includes the most complete possible set of observations (even if some are missing or never produced by a particular station) *cf.* table 1, of 28 meteorological stations (27 before January 1st 1974).

2) an optional listing of messages in readable form.

These two first programs form the very first steps of the data processing. These steps must be executed whatever is needed and whatever the following treatment programs are.

The intermediate output is a set of 'crude' data stored in a standard format usable as input by any subsequent program.

It must be emphasized that the rough data from the meteorological stations is only decoded and translated into numerical form. No kind of smoothing, interpolation or debugging has been performed at this point of the treatment. This is intended to store the most original possible measurements; doing so, it will be possible later to determine the eventual bias or systematic errors in data issued by some stations and introduce the corrections in the data handling programs themselves.

I.4.- Observed and geostrophic wind

Weather forecasting provides information on the surface pressure field, which is then converted into a field of geostrophic wind. An important question is then how the surface wind at sea may be derived from the geostrophic wind.

Under ideal conditions (no frontal situation, quasi-steady pressure gradients, weakly curved isobars), the ratio of surface wind speed to geostrophic wind speed depends mainly on the geostrophic wind speed, the stability of the density stratification and the latitude. Hasse and Wagner (1971) have investigated this problem from observations at the German Bight near 54 °N 7°E . The similarity with our test-region and the proximity in latitude of their observation zone suggest the direct applicability of their results. Using standard regression techniques, they found that the relation between the surface wind speed U_{10} and the geostrophic surface wind U_g was (in m/s) :

- under stable conditions : $U_{10} = 0.56\ U_g + 3.0$

(1) - near neutral : $U_{10} = 0.56\ U_g + 2.4$

- stable : $U_{10} = 0.56\ U_g + 1.5$

with standard deviation 2.1 m/s approximately independent of wind speed in the range $5\ m/s \leq U_g \leq 32\ m/s$. Air-sea temperature difference was used as stability parameter; the conditions are unstable if

$$\Delta T = T_{air} - T_{water} \ll 0 ,$$

near neutral if

$$\Delta T \simeq 0 ,$$

and stable if

$$\Delta T \gg 0 .$$

The stability can be established readily from the synoptic data, for a number of sea stations. Conditions are considered unstable or stable if the mean temperature difference ΔT is less than $- 1.2\ ^{o}C$ or greater than $0.7\ ^{o}C$ respectively and otherwise neutral.

Wagner and Hasse disregarded all cases where U_g was below 5 mph as the pressure gradient is then too small to be determined reliably. As those cases are of littly dynamical significance, we will use (1) for any value of the geostrophic wind. Note that at small wind speed, U_{10} is larger than U_g mainly because the surface wind is a local variable (influence by grestiness and small scale circulation) whereas U_g is a mesoscale variable.

Under general circumstances, there appear occasionally cases with fronts, noticeable curvatures or marked non stationarities in the pressure field. Curvature effects can be taken into account by computing the gradient wind instead of the geostrophic wind. Approximate formulae for the gradient wind are known. However the work of Neiburger *et al.* (1948) suggests that, over land, the geostrophic wind is more suitable for the determination of the actual wind than the gradient wind. Aagaard (1969) estimates that the error in neglecting curvature is no more than a few percent. Thus it will be neglected although it might play a role at high wind speed [Deacon (1973)]. Formulae (1) will be used, keeping in mind that under general circumstances, the standard error is larger than given.

Formulae (1) describe how to relate the magnitude of observed and geostrophic wind. There is also a small angle between the two vectors which has also to be taken into account. In most studies, the coded directions (in terms of degrees) do not permit a meaningful investigation of this small effect. From the results of Aagaard (1969), Smith (1970) and Deacon (1973), a reasonable estimate is a $15^\circ - 20^\circ$ counterclockwise rotation of the surface wind U_{10} from the geostrophic wind U_g. We shall adopt

$$\alpha = +20^\circ . \tag{2}$$

Further test and improvement of relations (1) and (2) will be undertaken using the data whose acquisition system is described.

I.5.- Example of application

As an example of use of meteorological data, we have computed the wind field over the Southern Bight during the first days of JONSDAP survey. The values of wind speed and direction have been interpolated from observed data. The points where the wind field has been calculated are the points of a grid used for tidal computations in the Southern Bight (this grid covers well the studied region). To each point of this computational grid are associated the five nearest meteorological stations. Three of these stations (the nearest ones yielding required information at a given time) are used to compute the value of a function at a point (i,j) of the grid with the formula :

$$f_{ij} = \frac{\sum_{K=1}^{3} \frac{f_K}{d_{ij,K}}}{\sum_{K=1}^{3} d_{ij,K}}$$

where $d_{ij,K}$ is the distance between the (i,j) point and the station K . If the data from any station is obviously wrong, it is then automatically dropped off. Three-hourly information is then integrated over one day in order to get an average daily value. The results of the computation are shown on fig. 3. For the sake of readability, the wind vectors (underlined) are drawn only every two points. The vectors point to the point where the wind is computed and their direction is that of the coming wind.

fig. 3.

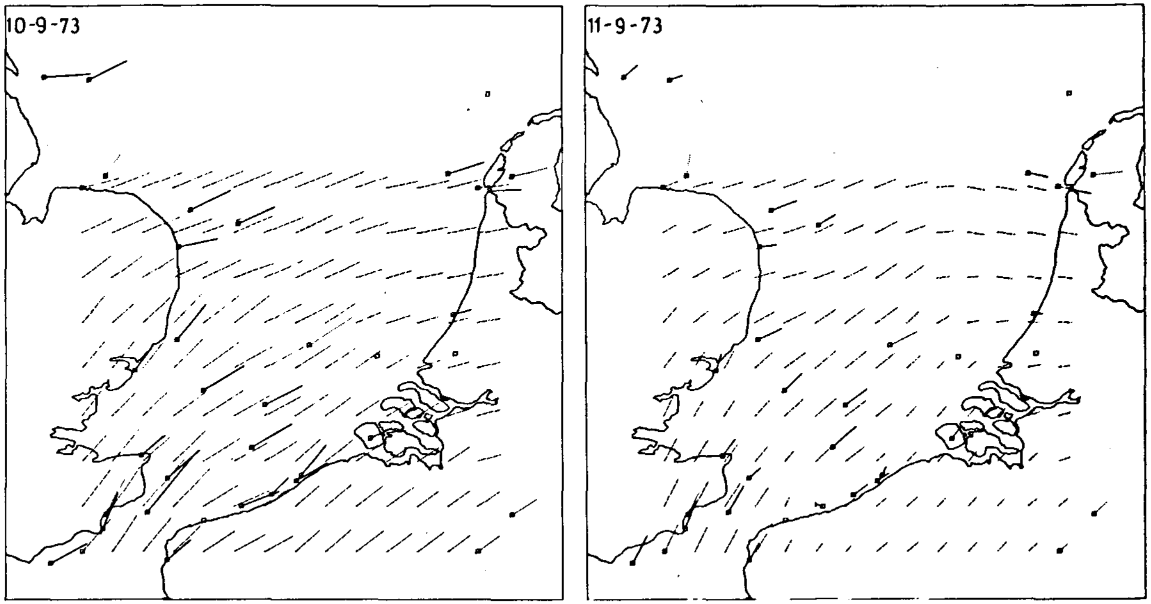
WIND FIELD OVER THE SOUTHERN BIGHT OF THE NORTH SEA DURING JONSDAP 73

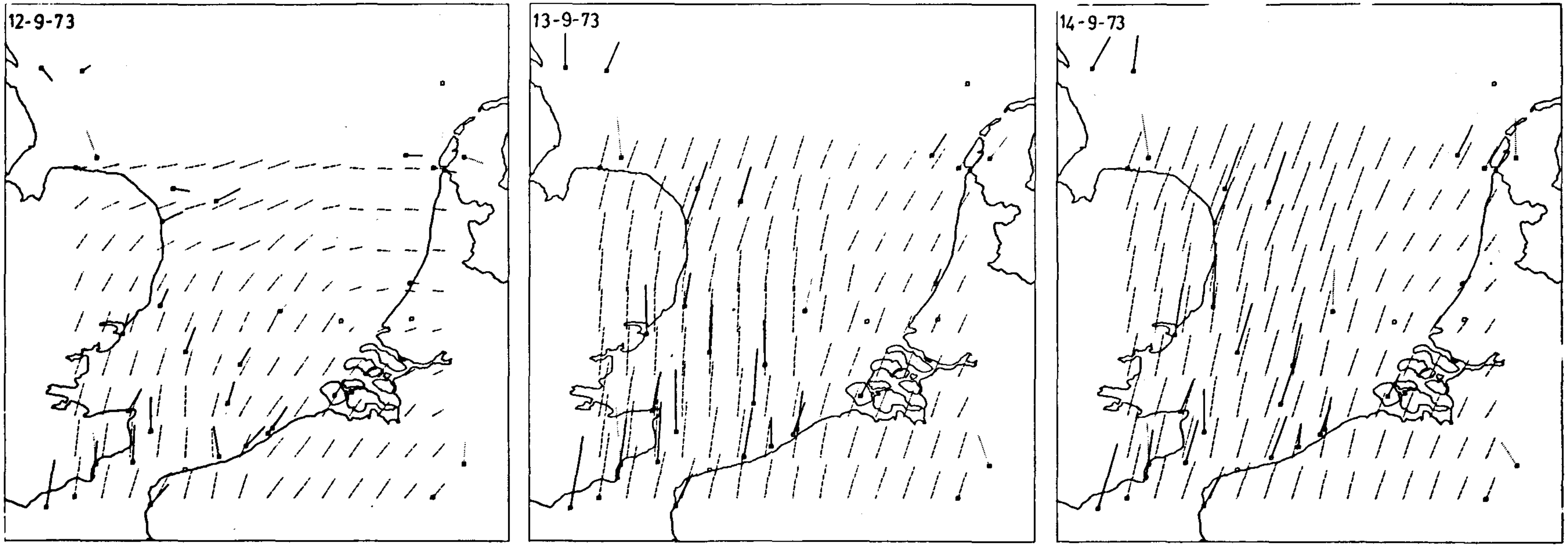
—— OBSERVED WIND

.... OBSERVED GEOSTROPHIC WIND

------ INTERPOLATED WIND

SCALE —— 10 M/S





Appendix II

Label record # 1 (Magnetic Tape format)

Word	Meaning
1	Record length
2	0 — this information is needed by some sophisticated
3	Record length — Operating Systems
4	0
5	Key-word : integer word (value = 1) : defines a Label # 1 record
6	Interchange code (code of the internal character representation) Integer : (1) EBCDIC (2) ASCII
7	Internal code : integer word
8-17	Data Name (20 bytes) - 4 bytes for a mnemonic code (nature of data) - 4 bytes for a station - 7 bytes for the radio identification of the data acquisition support (buoy, bart, ...) - 5 bytes for a mnemonic code (history of data)
18-22	Date of File creation (5 integer words : year-day-hour-minute-second)
23-24	Data source (4 bytes)
25	# variables (integer)
26	# samples (integer)
27	Format of the second Label record (integer) (standard value = 1)
28-32	Date of data origin (5 integer words : year-day-hour-minute-second)
33-104	Comments (144 bytes)

Label # 2

Word	Meaning
1	Record Length
2	0
3	Record Length — see Label # 1 description
4	0
5	Keyword : integer word (value = 2) : defines a Label #2 record
6	of sub samples (integer) (value = 1 if no sub sampling)
7-8	Sampling interval (real word) : meaningfull only for data depending on a continuous variable
9-10	Subsampling interval (real word) : meaningfull only if both former informations are meaningfull
11-12	Julian day or initial value of continuous variable other than time or 0 (no continuous variable)
13-14	Seconds or - 1 or 0
15-18	Latitude : 8 bytes : degrees-min-seconds — for standard format 1
19-22	Longitude : idem — Otherwise, 24 bytes available to
23-26	Magnetic variation : idem — describe another standard format

Label # 2 (continuation)

Word	Meaning
27-32	Variable name : 12 bytes
33-38	Dimensional units : 12 bytes
39-40	Measure code : 4 bytes
41-42	Instrument code : 4 bytes
43-44	Serial # : 4 bytes
45	Data Type : 2 bytes 'I' : integer; 'R' : real; 'S' : synoptic
46	Depth (in meters) or pressure (in mb) or 0 (integer word)
47-48	1st attribute (real)
49-50	2nd attribute (real)
51-52	3rd attribute (real)

Words 33-52: This set of information is repeated for every variable in the data file (up to 64)

Data record

Word	Meaning
1	Record length (see Label # 1 description)
2	0 (see Label # 1 description)
3	Record length (see Label # 1 description)
4	0 (see Label # 1 description)
5	Keyword : integer word (value = 3) : defines a data record
6-15	Data Name (see Label # 1 description)
16	# samples in preceeding data records (integer)
17	# samples in this record
18	Not used
19-20	Julian day or initial value of the continuous variable in this record
21-1172	Data

References

AAGAARD, K;, (1969). Relationship between geostrophic and surface winds at weather ship M, *Journal of Geophysical Research, 74,* 3440-3442.

DEACON, E.L., (1973). Geostrophic drag coefficients, *Boundary-Layer Meteorology, 5,* 321-340.

FRANKIGNOUL, C.J., (1971). Preliminary report on air-sea interaction, *Technical Report N7,* C.I.P.S.

HASSE, L. and WAGNER, V., (1971). On the relationship between geostrophic and surface wind at sea, *Monthly Weather Review, 99*, 255-260.

NEIBURGER, M., SHERMAN, L., KELLOGG, W.W., GUSTAFSON, A.F., (1948). On the computation of wind from pressure data, *Journal of Meteorology, 5*, 87-92.

NIHOUL, J.C.J., (1973). *Interactions at the sea boundaries*, in *Proceedings of the 5th Liège Colloquium on Ocean Hydrodynamics*, (Mémoires de la Société Royale des Sciences de Liège).

NIHOUL, J.C.J., (1974). *Réflexions sur la finalité d'une banque de données*, in *Proceedings du Colloque 'Informatique et Environnement'* (Arlon, 1974).

MALTAIS, J.A., (1969). *A nine-channel digital magnetic tape format for storing oceanographic data*. Unpublished manuscript n° 69-55 WHOI.