A Binary Universal Form for the Representation of Meteorological Data - an introduction to FM 94 BUFR

by

Daniel Söderman and J.K. Gibson

European Centre for Medium-Range Weather Forecasts

ABSTRACT

A binary representation, internationally recognised, is necessary for the efficient exchange and storage of meteorological data. A representation form has been adopted for trial use based on simple but universal table driven concepts. To enhance efficiency, a compression method is included. The resulting representation form will be of particular relevance to observational data; it is envisaged that FM 94 BUFR with form FM 92 GRIB for the representation of grid point data sets will eventually become the principle representation forms for the exchange and storage of meteorological data.

1. What is FM 94 BUFR?

Meteorology addresses weather phenomena that respect no national boundaries. The international co-operation with respect to the observing of weather, the exchange of data, the production and exchange of forecast products, and the organisation of the World Weather Watch has long set an example for all to follow. WMO FM Code Forms exploited the opportunities for the international exchange of data presented by telegraphic communications. The Global Telecommunications System (GTS) evolved to handle these character messages.

With the advent of computerised telecommunications and internationally accepted protocols capable of handling binary data, two forms of representation, FM92 GRIB (GRId in Binary), and FM 94 BUFR (Binary Universal Form for the Representation of meteorological data) have emerged. They are designed to meet the needs for the efficient exchange and storage of meteorological data in binary form.

BUFR is designed to represent any meteorological data in a logical, efficient way. Binary representation, independent of any particular computer is employed. Each "BUFR message" is a string of binary numbers, independent of any computer physical record, block, or file structure.

BUFR includes both data and data description. Thus

- BUFR messages can be exchanged and interpreted;
- BUFR can represent all currently available data;
- BUFR can represent ad hoc sets of elements which would require new FM Character Code Forms for exchange by such means.

Because FM 94 BUFR represents data, simply defined, in a form suitable for computer processing, it is efficient in a data processing environment. Whereas huge decoding programs are needed to convert FM Character Code Forms to data processing entities, BUFR requires only a relatively simple expansion process.

2. The Representation of Meteorological Data

J. Pilon (1985) has demonstrated that the value of a meteorological element can be regarded as a function of

- time
- space
- meteorological parameter.

The specification for BUFR follows, and somewhat extends this concept. Users particularly of observational data may wish to be aware, additionally, of

- identification data (e.g. station number, ship call sign, etc.);
- instrumentation data, including observational practice;
- space/time qualifications, such as significant level for temperature, first cloud layer, etc.

3. Data Description in BUFR

The concept of a code table has proved a useful aid to data description and representation in meteorology. Traditional code tables simply associate a value or a range of values with a code figure. Using the code figure, the appropriate code table enables the associated value to be obtained.

BUFR uses a table (table B) to describe data elements; it contains information about how to encode or decode the values, and specifies:

- the parameter name
- the units used
- details of the means of representation, including the data width in bits.

Code figures used to describe data in BUFR are called "descriptors". Such a descriptor describing a single element is called an "element descriptor". A list of element descriptors together can define a set of binary data. This concept is illustrated by the following example:

001001	001002	002011	002012
004001	004002	004003	004004
004005	005002	006002	007001

This list of descriptors refers to the following extracts from BUFR tables:

Table Reference	Element Name	Units	Scale	Reference Value	Data Width (bits)
001001	WMO block number	numeric	–	–	7
001002	WMO station number	numeric	–	–	10
002011	Radiosonde type	code table	–	–	8
002012	Radiosonde computational method	code table	–	–	4
004001	Year	Year	0	0	12
004002	Month	Month	0	0	4
004003	Day	Day	0	0	6
004004	Hour	Hour	0	0	5
004005	Minute	Minute	0	0	6
005002	Latitude	degree	2	- 9000	15
006002	Longitude	degree	2	-18000	16
007001	Height of station	m	0	- 400	15

Thus, the list of descriptors describes the identification, instrumentation, and location with respect to time and 3 dimensional space of a radiosonde observation. Note that the data are defined to occupy 7, 10, 8, 4, 12, 4, 6, 5, 6, 15, 16 and 15 bits for each sequential item; thus the list of descriptors define 12 data elements, represented by 108 bits of binary data.

To support this means of data description BUFR table B includes entries corresponding to many meteorological parameters, together with entries for identification, instrumentation, and time-space location. Using lists of element descriptors, table B, and a simple set of rules to define the significance of the identification, instrumentation and time-space location, any meteorological data may be described and represented.

Extension of BUFR tables does not affect previously defined data. Thus the tables may be added to as the needs arise.

4. Additional Data Description Concepts

Lists of element descriptors could describe any meteorological data or product. The lists would become long and would occupy more space than the data they define. To overcome this potential problem, two additional concepts are available:

- the replication operator
- BUFR table D.

The replication operator is specified by means of a special descriptor. It defines a range of subsequent descriptors, together with a replication factor. This enables the appropriate descriptors to be considered to be repeated a number of times. For example:

102005 007004 010003

is equivalent to

```
007004  010003  007004  010003
007004  010003  007004  010003
007004  010003
```

and could be used to define data representing the geopotential height of 5 pressure levels. The descriptor 102005 is the replication descriptor; it conveys the information "replicate 2 descriptors 5 times". A special form of the replication operator allows the replication factor to be stored with the data rather than within the descriptor. This enables data to be described in a general way, with the number of replications being different from case to case; this special form is called "delayed replication".

BUFR table D is referenced in the same way as table B. Instead of defining an item of binary data, each entry in table D contains a list of descriptors. Where lists of descriptors are frequently required in the repeated definition of commonly associated data, they may be added to table D. A descriptor reference to table D is called a "sequence descriptor", because it refers to the list or sequence of descriptors stored at that table entry. Table D thus provides building blocks to shorter data description. For example:

302003

is equivalent to

011011	011012	012004	012006
013003	020001	020003	020004
020005			

and defines wind direction (10m), wind speed (10m), temperature (2m), dew point (2m), relative humidity, horizontal visibility, present weather, past weather (1), and past weather (2).

There is a potential need to re-define, temporarily, table B attributes, and to represent less well defined data such as quality control fields, character data, etc. To meet these needs additional operation techniques are defines in BUFR table C.

Where a number of data sub-sets have exactly the same data description, the data description only occurs once. The number of data sub-sets is given within the data description section. Data may then be either in compressed form, as described under 7. below, or as would be described by repeating the data description once for each data sub-set.

5. The Structure of BUFR Descriptors

A BUFR descriptor has a data width of 2 octets and contains 3 values:

F	X	Y
< 2 bits >	< 6 bits >	< 8 bits >

The numeric value of the 2 bit quantity, F, indicates the type of descriptor

F = 0	element descriptor
F = 1	replication operator
F = 2	operator descriptor
F = 3	sequence descriptor.

Tables B and D are divided into classes and categories respectively. This is mainly to assist in the maintenance, readability and presentation of the entries they contain.

In table B the elements which define identification, instrumentation, time, space and significance are contained in classes 0 to 9. Elements from these classes relate to the subsequent meteorological data. Thus, once defined, they remain in effect until re-defined.

In table D the first categories contain sequences of descriptors referring to similar or related data entities. The later categories build up combinations of unrelated entities structured according to the means of measurement or reporting.

For element descriptors and sequence descriptors the X value indicates the class or category, and the Y value the entry within that class or category with reference to tables B and D respectively.

6. The Reserved Areas of Tables B and D

Tables B and D can each accommodate 64 classes or categories with 256 entries in each. To allow room for future additions, the initial tables will use only classes or categories 0 to 31 inclusive, and entries 0 to 63 inclusive.

Since many data processing centres will need to represent data conforming to local individual requirements, specific areas of tables B and D are reserved for local use. Initially, these areas are defined as entries 192 to 255 inclusive of all classes. Centres defining classes or categories for local use should restrict their use to the range 54 to 63 inclusive (see Fig. 8.1).

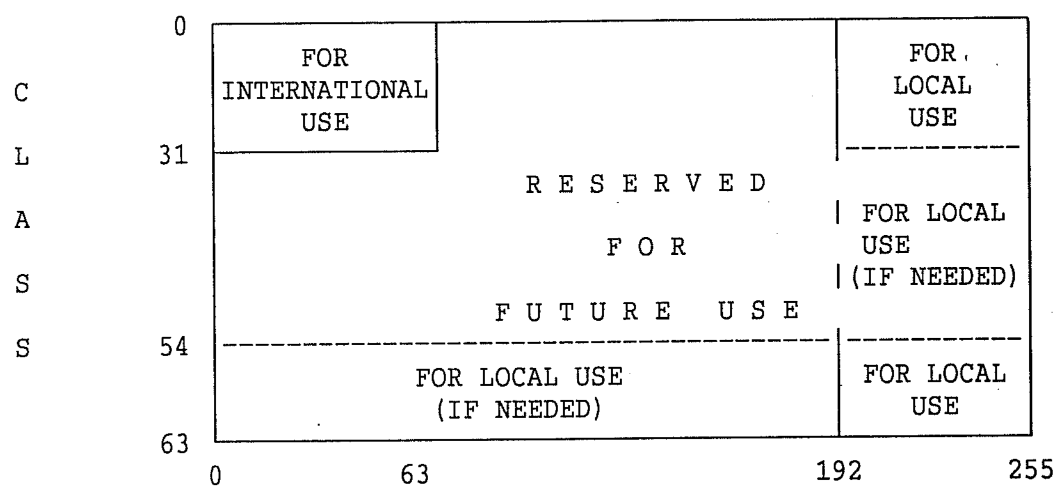

Fig. 8.1 Table B & D Reservation

7. BUFR Data Compression

Where a number of data sub-sets with identical descriptions are stored together in a BUFR message, data compression may be used, element by element. The BUFR data compression method is particularly effective where variations of each element between the sub-sets is small. The compression method employed is similar to that used in FM 92 GRIB. For each element the range is examined, and the minimum extracted. Corresponding elements for each data sub-set are represented in terms of differences from these minima. For example, consider the following group of identically defined data sub-sets:

	station number	station height	pressure	temperature	dew point
sub-set 1	101	296	10132	122	110
sub-set 2	125	291	10122	121	110
sub-set 3	127	310	10050	105	099
sub-set 4	136	295	10119	110	102
sub-set 5	138	350	10055	095	089
sub-set 6	141	325	10075	101	091

Extraction of the minimum value of each element gives:

101	291	10050	095	089

If we now represent each value as the difference from these minima, we obtain a set of "increments":

0	5	82	27	21
24	0	72	26	21
26	19	0	10	10
35	4	69	15	13
37	59	5	0	0
40	34	25	6	2

By representing this reduced range of values, less bits are required to represent each value. This is achieved in the actual data by representing

min (element 1), data width increments (element 1), inc_1, inc_2, etc.
min (element 2), data width increments (element 2), inc_1, inc_2, etc.
. . . . .
. . . . .
. . . . .
min (element n), data width increments (element n), inc_1, inc_2, etc.

inc_1, inc_2, etc. are the increments, each represented in the data width given.

The algorithm for accomplishing data compression follows:

(i) Start with a set of n like values of an element (with the appropriate table reference value R already subtracted out).

(ii) Scan the set and find the minimum value therein: the local reference value, R°.

(iii) Subtract the local reference value from each of n values giving a set of increments, I.

(iv) Determine the number of bits necessary to contain the largest of the increments plus 1.

(v) Repeat the process for the other like values in the BUFR message.

The original values V can be recovered by

$$V = R + R^{o} + I$$

R = Table reference value
R^{o} = Local reference value
I = Increment

Missing values in the original data, provided all values are not missing values, are to be ignored when finding the local reference value (Step ii), and calculating the increments (step iii).

Local reference values are represented according to the units, scale, reference value and data width defined for the corresponding data element.

Missing values are indicated by fields of all ones.

A set of identical values is represented by entering the single value for the set as the local reference value, the data width for the increments as zero, and omitting the set of increments. A set of missing values is represented by entering the missing value (field of ones) as the local reference value, the width for the increments as zero, and omitting the set of increments.

8. The Supplementary Information within BUFR

A BUFR message is comprised of up to 6 sections. Section 0 and section 5 contain the characters BUFR and 7777 to indicate the beginning and the end of the BUFR message. In addition to the data and data description, information such as data type, date/time, origin, and details of the version of BUFR tables used are included. This enables the message to be handled without decoding the data.

9. What is FM 94 BUFR?

Many years ago a Danish meteorologist, J.W. Larsen, proposed the idea of a universal code form to represent meteorological data. Since this idea was first put forward:

- meteorological telecommunications have become increasingly automated
- communications protocols have been developed and standardised
- remote sensing technology has become a principle source of meteorological observational data
- computer analysis and assimilation of meteorological data has advanced
- computer costs have been reduced to the point where a powerful micro computer is as cheap as a teleprinter
- the storage of large amounts of data in computer processable form has become important for meteorological research

The advent of automated data processing techniques enabled meteorological products to be exchanged and used for many purposes, including automated flight planning for aviation. Many forms of data representation resulted - FM 47-V GRID, FGGE III, aviation "stack" codes, etc. The need to develop an efficient, standard binary representation form resulted in FM 92 GRIB (Grid in binary), wich is now extensively used.

FM 94 BUFR is intended to meet similar needs for data which cannot necessarily be represented using a regular, two dimensional, horizontal grid. Being self-defining with regard to data contents, it is universal. Thus it can represent any combination of data elements. This means that:

- BUFR can represent currently exchanged data
- BUFR can represent any new observation form
- BUFR can represent the raw data resulting from remote sensing
- BUFR can be decoded by one single computer programme, which does not need modification for changes to the data content resulting from changes to observational practise.

In consequence, FM 94 BUFR is ideal for the storage and exchange of data within an automated environment.

10. Future trends

The decoding of BUFR does not require extensive computer resources. Investigations into decoding methods (Gibson and Dragosavac, 1987) show that a fully automated "display form" can result such that no further recourse to code tables is required. Thus, a BUFR message can be displayed in a form which is completely comprehensible. Such a form could be produced with the aid of a microcomputer, and either displayed on a VDU or printed.

Data base methods will enable the selection of appropriate BUFR messages. The decoding process will allow either complete decoding, or the decoding of selected elements. Thus BUFR will prove to be an ideal representation form for the storage of data, both on-line and archival.

With the implementation of standard telecommunications protocols (e.g. X.25) it will become increasingly possible to use binary forms for the transmission of data between automated centres. In consequence it is envisaged that FM 94 BUFR and FM 92 GRIB will eventually become the principle representation forms for the exchange and storage of meteorological data. It follows that the development of new remote sensing equipment should endeavour to make use of BUFR for the storage and transmission of the data produced.

Support for less automated centres will be required for many years. This can be effected by the selection of required data sub-sets, and the presentation of these data in an efficient, printable character form.

BUFR is extremely efficient when representing sets of closely correlated data, such as high resolution satellite soundings over an area. The data compression possible in these circumstances enables many times the efficiency of the traditional SATEM code to be achieved. Since such data is becoming increasingly important and available, a means of effecting its efficient exchange is essential.

The World Weather Watch long term plan envisages the development of new data management techniques to maximise the use of all sources of meteorological data, and the exchange of resulting products. Many years ago the invention of the character code forms enabled meaningful data to be exchanged in near real time by telegraphic means. The ability of meteorologists of any nation speaking any language to interpret these messages contributed greatly to international co-operation. With the advent of BUFR and GRIB, the meteorological community is again pioneering co-operation in the field of automated processing. The new binary forms will enable meaningful data to be exchanged in real time by automated means. The ability of computers of any make using any form of internal representation to interpret these messages will be of supreme importance for future progress.

Glossary

BUFR message	a single complete BUFR entity
category	the lists of sequence descriptors tabulated in BUFR table D are categorised according to their application; categories are provided for non-meteorological sequences, for various types of meteorological sequences, and for sequences which define reports, or major subsets of reports
class	the elements tabulated in BUFR table B are classified according to type.
co-ordinate class	classes 0-9 inclusive in BUFR table B define elements which assist in the definition of elements from subsequent classes; each of these classes is referred to as a co-ordinate class
data description operator	operators which define operations resulting from the application of the replication operator or the operations listed in BUFR table C
data entity	a single data item
data sub-set	a set of data corresponding to the data description in a BUFR Message; for observational data a data sub-set usually corresponds to one observation
descriptor	a 16 bit unsigned positive integer entered within the data description section to describe or define data; a descriptor may take the form of an element descriptor, a replication operator, an operator descriptor, or a sequence descriptor
element descriptor	a descriptor containing a code figure reference to BUFR table B; the referenced entry defines an element, together with the units, scale factor, reference value and data width to be used to represent that element as data
operator descriptor	a descriptor containing a code figure reference to BUFR Table C, together with data to be used as an operand
reference value	all data are represented within a BUFR message by unsigned positive integers; to enable negative values to be represented, suitable negative base values are specified as reference values. The true value is obtained by addition of the reference value and the data as represented

replication descriptor	a special descriptor is reserved to define the replication operation; it is used to enable a given number of subsequent descriptors to be replicated a given number of times
section	a logical sub-division of a BUFR message, to aid description and definition
sequence descriptor	a descriptor used as a code figure to reference a single entry in BUFR table D; the referenced entry contains a list of descriptors to be substituted for the sequence descriptor

References

J. Pilon (1985)	New concepts in Data Management, WMO GDPS Report No. 2 "New Trends in Data Representation"
(Awaiting publication)	Binary Universal Form for data Representation FM 94 BUFR
J.K. Gibson and M. Dragosavac (1987)	Decoding Data Represented in FM 94 BUFR, ECMWF Technical Memorandum No. 134