CHAPTER 4

DATA COMPRESSION

4.1 Introduction. Even though BUFR makes efficient use of space by virtue of binary numbers that take only as many bits as are necessary to hold the largest expected value, a further compression may be possible. Data compression is indicated by setting the second bit of octet 7 in Section 3 to a value of one.

4.2 Method Used for Data Compression. The method employed by BUFR for data compression is similar to that used in the WMO Code FM 92 GRIB (GRidded Binary fields). Like elements from the full set of observations are collected together, their minimum values subtracted out, and the difference from the minimum are then encoded with a bit length selected to hold the largest difference from the minimum value. This is repeated for all the elements.

Using the following group of identically defined data subsets:

	Station Number	Station Height	Pressure	Temperature	Dew Point
subset 1	**101**	296	10132	122	110
subset 2	103	**291**	10122	121	110
subset 3	107	310	**10050**	105	099
subset 4	112	295	missing	110	102
subset 5	114	350	10055	**095**	**089**
subset 6	116	325	10075	101	091

Extraction of the minimum value of each element gives:

Station Number	Station Height	Pressure	Temperature	Dew Point
101	291	10050	095	089

Each value can now be represented as the difference from these minimum values:

	Station Number	Station Height	Pressure	Temperature	Dew Point
subset 1	0	5	82	27	21
subset 2	2	0	72	26	21
subset 3	5	19	0	10	10
subset 4	11	4	missing	15	13
subset 5	13	59	5	0	0
subset 6	15	34	25	6	2

After each difference from the minimum value has been determined for each element, the number of bits necessary to store the largest of the difference values for each element is established. For the station number column above, the largest difference is 15 which is

equivalent to 1111_2, or 4 bits. However this presents a small problem. All four bits being set on, as is the case for the number 15, is properly interpreted as "missing," not as a numeric value of 15. What is done in such cases is to simply add one bit to the number needed to store the largest difference value; thus 15 gets stored in 5 bits, as 01111. It is not necessary to add one bit to the bit lengths for all the elements; it is only necessary when one of the numbers to be encoded "fills" the available space; that is, if the number 3 is to be stored in 2 bits, 7 in 3 bits, 15 in 4 bits, 31 in 5 bits, etc. A convenient way to do this and assure that there is always room for "missings" (if needed) is to add 1 to the largest difference value and figure the number of bits based on this larger-by-one value.

In the example above, the station height would be placed in 6 bits; the pressure in 7 (with the "missing" indicated as 1111111), etc., as in the following table:

	Station Number	Station Height	Pressure	Temperature	Dew Point
largest difference value +1	16	60	83	28	22
number of bits	5	6	7	5	5

Whereas in the non-compressed storage of data in Section 4 there is a continuous bit stream for all parameters for an entire observation, in the compressed form all elements of the same parameter from each observation form a continuous stream (Figure 4-1). In order to determine what the minimum value is that has to be added back to each of the following elements, and how many bits are being used for the storage of these elements, there are two additional items appearing in the compressed form of storage in Section 4 that do not appear in the non-compressed form.

These items are:

(1) the minimum value of this parameter and,
(2) the number of bits that are being used for the storage of each element.

These items of information precede the element values. The Section 4 representation for compressed data for each parameter used in the example above is:

Station number minimum value (101) occupying 10 bits as specified by the Table B data width for entry 0 01 002 followed by:

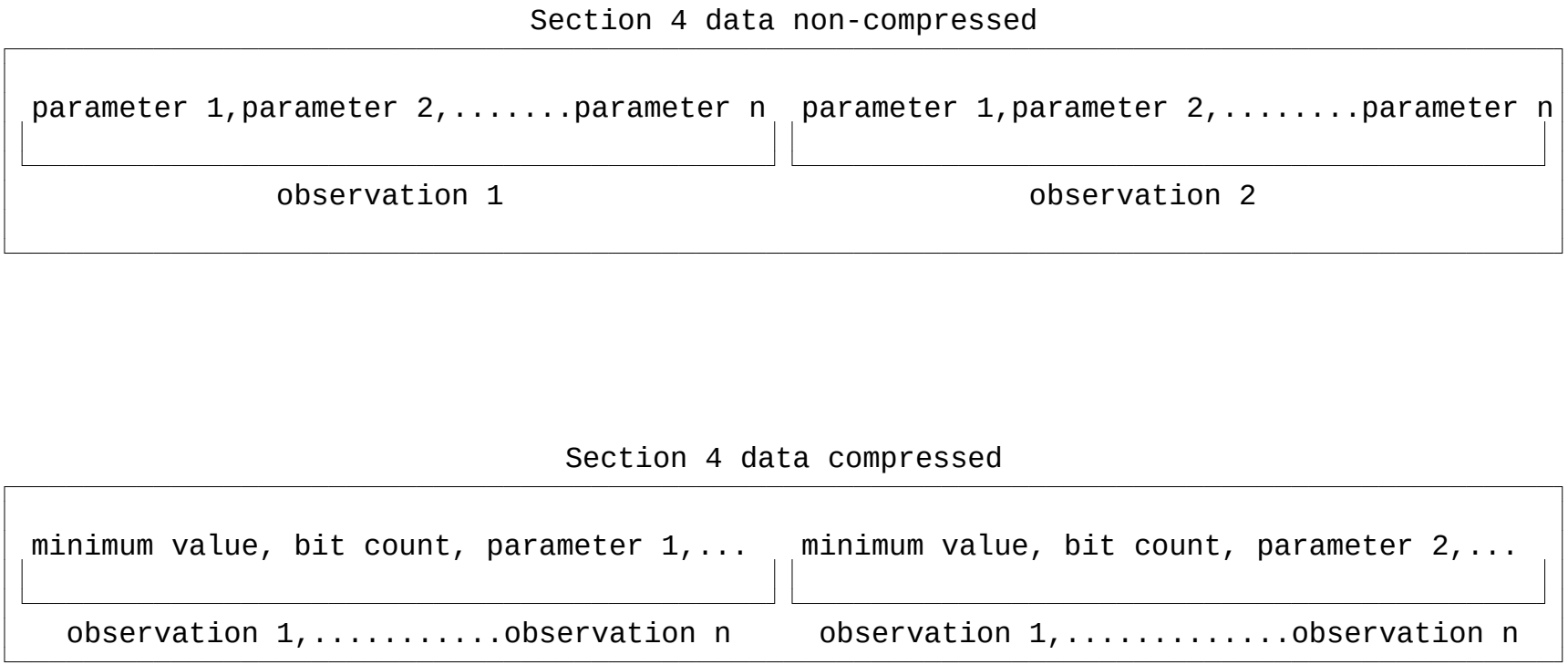

Figure 4-1. Comparison of non-compressed and compressed data in Section 4

6 bits containing the count in bits (5) that each of the station numbers will occupy, followed by: the 6 station number differences from the minimum values (0, 2, 5, 11, 13 and 15), where each value occupies 5 bits.

After the last station number difference (15), the next 15 bits (Table B data width for entry 0 07 001) will be taken by the minimum value for station height (291) followed by the count of bits to represent the differences (6) and then each of the elements occupying 6 bits apiece (5, 0, 19, 4, 59, 34).

Continuing the process for all 5 parameters would produce within Section 4 the following bit counts:

	station number		station height		pressure		temperature		dew point	
Table B descriptor	0 01 002		0 07 001		0 10 004		0 12 004		0 12 006	
data width to contain minimum value	10		15		14		12		12	
6 bits containing bit count of parameter	6		6		6		6		6	
Total bits preceding each parameter	16		21		20		18		18	
data width to represent difference from minimum	5		6		7		5		5	
compressed data representation for 6 subsets	30		36		42		30		30	
total bit count for 6 subsets including compression bit counts	46	+	57	+	62	+	48	+	48	= 261

To represent all 6 subsets in compressed form in Section 4, 261 bits are necessary.

Using the same set of values for the 6 subsets in non-compressed form, the bit counts in Section 4 would be as follows:

	station number		station height		pressure		temperature		dew point	
Table B descriptor data width	10		15		14		12		12	
total bit count for 6 subsets	60	+	90	+	84	+	72	+	72	= 378

A total of 378 bits are necessary to represent all 6 subsets in non-compressed form.

There are other conditions that can occur when encoding compressed data. If all elements of a set of parameters are missing, the minimum value occupying the specified Table B data width in Section 4 will be set to all 1's, the 6 bits specifying how many bits are used for each value will be set to 0, and the difference values will be omitted. If, for example, all the dew points were missing from the 6 subsets, the number of bits to represent dew point would be reduced to include only the Table B data width for dew point (12 bits) and the 6 bits specifying the bits used for each value.

	station number	station height	pressure	temperature	dew point
Table B descriptor	0 01 002	0 07 001	0 10 004	0 12 004	0 12 006
data width to contain minimum value	10	15	14	12	12
6 bits containing bit count parameter will occupy	6	6	6	6	6
Total bits preceding each parameter	16	21	20	18	18
compressed data (difference from minimum)	5	6	7	5	0
compressed data representation for 6 subsets	30	36	42	30	0
total bit count for 6 subsets including compression identifiers	46 +	57 +	62 +	48 +	18 = 231

In the non-compressed form, storage of the missing dew point values would still occupy 12 bits each, with all bits set to 1.

	station number	station height	pressure	temperature	dew point
Table B descriptor data width	10	15	14	12	12
total bit count for 6 subsets	60 +	90 +	84 +	72 +	72 = 378

The other condition that may occur is, if all the difference values are identical, then, the 6 bits specifying the count of bits for each difference value will set to 0, and difference values will be omitted. This condition would produce the same bit count as if all elements were missing. In summary:

Set of parameters missing:

minimum value occupying number of bits as indicated in Table B set to all 1's

6 bits specifying how many bits are used for each value set to 0

difference values omitted

Set of identical parameters:

minimum value occupying number of bits as indicated in Table B set to minimum value (actual value for all parameters)

6 bits specifying how many bits are used for each value set to 0

difference values omitted

Data compression is most effective when the range of values for the parameters is small. In the example of the 6 subsets, each parameter has a difference from the minimum value, where the number of bits to represent the difference is half, or less than half, the number of bits required in non-compressed form for storage in Section 4, as indicated by the Table B entry data width. If the 6 subsets were put into a message where compression was not applied, the length of the message would be 100 octets (Figure 4-2). By applying compression, the length of the message would be reduced to 86 octets (Figure 4-3).

Using the range of values for the same 6 subsets, not realistic, but to show the effect of compression for a large data set, a total of 4267 subsets could be put into a BUFR message not exceeding 15000 octets (Figure 4-4). In non-compressed form there would be only 1898 subsets within the 15000 octet limit (Figure 4-5).

	Section Octet No.	Octet in Message	Encoded Value	Description
Section 0 (indicator section)	1-4	1-4	BUFR	encoded international CCITT Alphabet No. 5
	5-7	5-7	100	total length of message (octets)
Section 1 (identification section)	8	8	2	BUFR edition number
	1-3	9-11	18	length of section (octets)
	4	12	0	BUFR master table
	5-6	13-14	58	originator (U.S. Navy - FNOC)
	7	15	0	update sequence number
	8	16	0	indicator for no Section 2
	9	17	0	Table A - surface land data
	10	18	0	BUFR message sub-type
	11	19	2	version number of master tables
	12	20	0	version number of local tables
	13	21	92	year of century
	14	22	4	month
	15	23	18	day
	16	24	0	hour
	17	25	0	minute
	18	26	0	reserved for local use by ADP centers (also needed to complete even number octets for section
Section 3 (Data description section)	1-3	27-29	18	length of section (octets)
	4	30	0	reserved
	5-6	31-32	6	number of data subsets
	7	33	bit 1=1	flag indicating observed data
			bit 2=0	**flag indicating <u>no compression</u>**
	8-17	34-43	0 01 002	WMO station no.
			0 07 001	height of station
			0 10 004	pressure
			0 12 004	temperature
			0 12 006	dew point
	18	44	0	needed to complete section with an even number of octets
Section 4 (Data section)	1-3	45-47	52	length of section (octets)
	4	48	0	reserved
	5-52	49-96	data	continuous bit stream of data for 6 subsets, 63 bits per subset plus 6 bits to end on even octet
Section 5 (End section)	1-4	97-100	7777	encoded CCITT international Alphabet No. 5

Figure 4-2. BUFR message of 6 subsets in non-compressed form

	Section Octet No.	Octet in Message	Encoded Value	Description
Section 0 (indicator section)	1-4	1-4	BUFR	encoded international CCITT Alphabet No. 5
	5-7	5-7	86	total length of message (octets)
	8	8	2	BUFR edition number
Section 1 (identification section)	1-3	9-11	18	length of section (octets)
	4	12	0	BUFR master table
	5-6	13-14	58	originator (U.S. Navy - FNOC)
	7	15	0	update sequence number
	8	16	0	indicator for no Section 2
	9	17	0	Table A - surface land data
	10	18	0	BUFR message sub-type
	11	19	2	version number of master tables
	12	20	0	version number of local tables
	13	21	92	year of century
	14	22	4	month
	15	23	18	day
	16	24	0	hour
	17	25	0	minute
	18	26	0	reserved for local use by ADP centers (also needed to complete even number octets for section
Section 3 (Data description section)	1-3	27-29	18	length of section (octets)
	4	30	0	reserved
	5-6	31-32	6	number of data subsets
	7	33	bit 1=1	flag indicating observed data
			bit 2=1	**flag indicating <u>compression</u>**
	8-17	34-43	0 01 002	WMO station no.
			0 07 001	height of station
			0 10 004	pressure
			0 12 004	temperature
			0 12 006	dew point
	18	44	0	needed to complete section with an even number of octets
Section 4 (Data section)	1-3	45-47	38	length of section (octets)
	4	48	0	reserved
	5-52	49-82	data	261 continuous bits of compressed data plus 11 bits to end on even octet
Section 5 (End section)	1-4	83-86	7777	encoded CCITT international Alphabet No. 5

Figure 4-3. BUFR message of 6 subsets in compressed form

	Section Octet No.	Octet in Message	Encoded Value	Description
Section 0 (indicator section)	1-4	1-4	BUFR	encoded international CCITT Alphabet No. 5
	5-7	5-7	15000	total length of message (octets)
	8	8	2	BUFR edition number
Section 1 (identification section)	1-3	9-11	18	length of section (octets)
	4	12	0	BUFR master table
	5-6	13-14	58	originator (U.S. Navy - FNOC)
	7	15	0	update sequence number
	8	16	0	indicator for no Section 2
	9	17	0	Table A - surface land data
	10	18	0	BUFR message sub-type
	11	19	2	version number of master tables
	12	20	0	version number of local tables
	13	21	92	year of century
	14	22	4	month
	15	23	18	day
	16	24	0	hour
	17	25	0	minute
	18	26	0	reserved for local use by ADP centers (also needed to complete even number octets for section
Section 3 (Data description section)	1-3	27-29	18	length of section (octets)
	4	30	0	reserved
	5-6	31-32	4267	number of data subsets
	7	33	bit 1=1	flag indicating observed data
			bit 2=1	**flag indicating <u>compression</u>**
	8-17	34-43	0 01 002	WMO station no.
			0 07 001	height of station
			0 10 004	pressure
			0 12 004	temperature
			0 12 006	dew point
	18	44	0	needed to complete section with an even number of octets
Section 4 (Data section)	1-3	45-47	14952	length of section (octets)
	4	48	0	reserved
	5-52	49-14996	data	119569 continuous bits of compressed data plus 15 bits to end on even octet
Section 5 (End section)	1-4	14997-15000	7777	encoded CCITT international Alphabet No. 5

Figure 4-4. BUFR message of 4267 subsets in compressed form

	Section Octet No.	Octet in Message	Encoded Value	Description
Section 0 (indicator section)	1-4	1-4	BUFR	encoded international CCITT Alphabet No. 5
	5-7	5-7	15000	total length of message (octets)
	8	8	2	BUFR edition number
Section 1 (identification section)	1-3	9-11	18	length of section (octets)
	4	12	0	BUFR master table
	5-6	13-14	58	originator (U.S. Navy - FNOC)
	7	15	0	update sequence number
	8	16	0	indicator for no Section 2
	9	17	0	Table A - surface land data
	10	18	0	BUFR message sub-type
	11	19	2	version number of master tables
	12	20	0	version number of local tables
	13	21	92	year of century
	14	22	4	month
	15	23	18	day
	16	24	0	hour
	17	25	0	minute
	18	26	0	reserved for local use by ADP centers (also needed to complete even number octets for section
Section 3 (Data description section)	1-3	27-29	18	length of section (octets)
	4	30	0	reserved
	5-6	31-32	1898	number of data subsets
	7	33	bit 1=1	flag indicating observed data
			bit 2=0	**flag indicating <u>no compression</u>**
	8-17	34-43	0 01 002	WMO station no.
			0 07 001	height of station
			0 10 004	pressure
			0 12 004	temperature
			0 12 006	dew point
	18	44	0	needed to complete section with an even number of octets
Section 4 (Data section)	1-3	45-47	14952	length of section (octets)
	4	48	0	reserved
	5-52	49-14996	data	continuous bit stream of data for 1898 subsets, 63 bits per subset plus 10 bits to end on even octet
Section 5 (End section)	1-4	14997-15000	7777	encoded CCITT international Alphabet No. 5

Figure 4-5. BUFR message of 1898 subsets in non-compressed form