X-601-78-21

NIMBUS-4 BUV DARK CURRENT STUDY

DATA FILTERING

E. G. STASSINOPOULOS
O. J. ROSE
R. A. GOLDBERG

JUNE 1978

National Aeronautics and Space Administration
Goddard Space Flight Center
Greenbelt, Maryland 20771

NIMBUS-4 BUV DARK CURRENT STUDY

DATA FILTERING

E. G. Stassinopoulos

Ollie J. Rose

R. A. Goldberg

NASA/Goddard Space Flight Center

Sciences Directorate

National Space Science Data Center

June 1978

Goddard Space Flight Center
Greenbelt, Maryland 20771

Introduction

A study is currently being performed for the Ozone Processing Team (OPT) with the objective to ascertain the nature and extent of dark current contamination of the NIMBUS-4 BUV measurements, and to formulate an empirical correction algorithm as a function of satellite position and time.

A partially treated and processed data base has been made available by the project office for this study.

In order to insure that the results of the analysis will be correct and reliable, it is necessary to further screen and clean the data before attempting statistical correlations, cross-correlations with solar-magnetic phenomena, or probabilistic modelling.

For that purpose, an initial simple filtering scheme was developed. After copious quantities of data were reviewed, this scheme eventually evolved into the comprehensive filtering process described in subsequent sections. Some of the filters were instrument-conditioned, others were empirically determined. This process is by no means concluded and additional filters may have to be developed as the analysis proceeds.

Filtering Sequence

The sequence in which the different filters are being activated or applied and their location in the special software package is shown in Table 1.

The arrangement reflects the steps in the evolution of the final scheme, in their actual chronological order. It also indicates the priorities that prevailed along the way in the filter development process.

Unavailable data points in the original data set supplied for the study (Code -1) and instrument-conditioned filters (Codes -10 & -11) received primary consideration. Only after this screening was accomplished was the emphasis placed on the actual "cleaning" process: Codes -6, -7, -5, -4, -2, -3, -8, -9. These codes were first developed for the pulse-count data and then applied also to the analog data, but only after a thorough validity verification.

Table 2 lists the type of data available for analysis and indicates their usefulness in terms of "gain" status.

Finally in Table 3, the codes are being explained in detail and the reasoning for their design is discussed.

Table 1

FILTERING SEQUENCE

	Sequ.	Data Being Filtered	Filter Code	Routine Where Filtering is Done
Combined data	1	PC	-1	MAIN: during "MASTER TAPE" generation
	2	PC + A	-10	FLIP
	3	PC + A	-11	FLIP
	4	A	-1	FLIP
	5	PC + A	-6	REPEAT
PC data only	6	PC	-7	FILTER
	7-9*	PC	-5 -4 -2	FILTER
	10	PC	-3	STRIP
	11	PC	-8	REFINE
	12	PC	-9	REFINE
A data only	13	A	-7	FILTER
	14-16*	A	-5 -4 -2	FILTER
	17	A	-3	STRIP
	18	A	-8	REFINE
	19	A	-9	REFINE

Legend: PC = pulse count
A = analog

*The order in which codes -5, -4, and -2 are assigned varies depending on the nature of the data (see explanations of filter codes).

Table 2

BUV-Data

Pulse Counts:

Gain

Photometer
- Hi = valid
- Lo = not valid

Monochrometer
- Hi = valid
- Lo = not valid

Analog U-values:

U_{photo}
- Hi = valid
- Lo = valid

U_{mono}
- Hi = valid
- Lo = valid

DATA FILTERING CODES

Table 3

Code	Explanation	Reasoning
-1.0	Data unavailable on original U-tape (<0.0 on U-tape).	Negative values on user-tape supplied by OPT (-99, -77, etc.).
-2.0	Less than three data points within a scan, therefore all data discarded.	After careful and extensive examination of many data records, it was concluded that if there were two or fewer positive values out of a possible 12 within a scan, the data were probably unreliable for that scan.
-3.0	Discarded because the "order-of-magnitude" difference of adjacent points is equal to or exceeds the value of two, except that representative* order of magnitude points are not discarded.	On physical grounds, it is unreasonable to believe that two data points adjacent in wavelength and separated in time by ∿2.5 seconds** will differ by two or more orders of magnitude***.
-4.0	Entire record discarded because quality of data judged to be poor (i.e., variation of more than four orders of magnitude within a scan).	Upon examining copious quantities of data, it was decided that if a variation of four or more orders of magnitude occurred in a single scan, it was likely that the entire scan was suspect.

*See appendix.

**The time estimate is taken from: "Nimbus IV User's Guide", p. 159, NASA-TM-X-69887.

***For the purposes of the filtering program, we used the following definition of order of magnitude: Let X be the number in question, let N be a non-negative integer such that $0.5 \times 10^N \leq X < 5 \times 10^N$, the N is the order of magnitude of X.

DATA FILTERING CODES (cont.)

Code	Explanation	Reasoning
-5.0	Exact 50:50 split in orders of magnitude grouping, and o.m.'s are not sequential, e.g., 50% of 3 and 50% of 5. This applies only to records with even number of elements present. When record has odd number of elements (because 1,3,5 etc. may be missing), minority is discarded.	If the tally of points in the order of magnitude grouping is exactly split 50:50 into two counting bins, they must differ by no more than one order of magnitude for the same reasons under code -3.0. Thus, if they differ by two or more orders of magnitude, the entire record is discarded because it cannot be decided which half to retain.
-6.0	Identical data in sequential records, i.e., values of three or more corresponding channels exactly equal in two or more sequential scans; repeated data assumed irregular, repeated records discarded.	It is unlikely that in sequential records data in corresponding channels will be identical, i.e., same values for same wavelength but in succeeding scans. The occurrence of three or more such identical points in sequential records was selected as a threshold for considering the repeated record to be suspect.
-7.0	Any pulse count data with a value of less than 100 are discarded. Similarly, any converted analog data (u-values) of magnitude less than 100 are also discarded.	Magnitude of minumum (quiet condition) background pulse count has been empirically determined to be >100; thus, any values less than 100 were discarded.

DATA FILTERING CODES (cont.)

Code	Explanation	Reasoning
-8.0	Discarded when ratio of maximum data value to the scan average is greater than 3.0 and point in question is assumed to be a local fluctuation.	By considering the character of the data variations across scans, it was determined that abrupt fluctuations can be detected by examining the minimum-to-average and maximum-to-average ratios. The cut-off values were found to be: $\frac{\max}{\text{avg}} \leq 3.0$ (Code: -8.0) $\frac{\min}{\text{avg}} \geq 0.3$ (Code: -9.0) Any data exceeding the cut-off values towards the extrema were discarded.
-9.0	Discarded when ratio of minumum data value to the scan average is less than 0.3 and point in question is assumed to be a local fluctuation.	See reasoning for Code -8.0.
-10.0	Discarded when magnitude of high-gain pulse count greater than 2.5×10^5.	High-gain pulse count data > 2.5×10^5 are considered unreliable due to dead-time correction.*
-11.0	Low gain pulse count: data invalid.	Low gain pulse count data are unreliable for the same reason as in code -10.0.

*J. Gatlin, et. al., private communication

Appendix

Representative Order of Magnitude

For the definition of "Representative Order of Magnitude" (ROM), the monochrometer and photometer data are being treated separately; that is, a different ROM is established for each measurement type.

To compute the order of magnitude of a variable $X > 0$, the following criterion was used: let N be an integer such that $0.5 \times 10^N \leq X < 5.0 \times 10^N$. Then for classification purposes, N is the order of magnitude of X. It has been empirically determined that the order-of-magnitude range of all available data is limited to the values $1 \leq N \leq 8$.

Next, let M be the total number of valid positive data points in a 12-channel (or wavelength ν_j; j=1,12) scan. It is possible that $M < 12$ because not all 12 data points are always available. Then for every scan where $M > 0$, eight order-of-magnitude accumulation bins are set up, one for each possible value of N.

For these eight accumulation bins, a counting parameter n_i (i=1,8) is established, where the subscript i relates to the bin which corresponds to the order-of-magnitude value N=i; for example, i=1 relates to the bin corresponding to N=1, etc.

The counting parameter n_i indicates how many times the specific order-or-magnitude occurred in a given scan; for example, every time N=3 the n_3 value is incremented by one (1).

By definition:

$$M \equiv \sum_i n_i$$

Using this identity, a fractional distribution of data points f_i can be obtained in terms of "order-of-magnitude" values:

$$f_i = \frac{n_i}{M}$$

Finally, let R be the maximum fractional distribution value:

$$R = \max_{i=1,8} [f_i]$$

If R is unique, that is, only one maximum value exists, then this order of magnitude is said to be "representative". If however, more than one maxima occur, as for example a 40:40:20 split or an even 50:50 distribution, then the scan is discarded because a prevailing value can not be established and a "representative" order-of-magnitude is not meaningful.