MIT Kavli Institute

Chandra X-Ray Center

MEMORANDUM
June 12, 2013

To: Jonathan McDowell, SDS Group Leader
From: John C. Houck, SDS
Subject: A better row-loss upper limit for `destreak`
Revision: 1.1
URL: `http://space.mit.edu/cxc/docs/docs.html#s4streak_loss`
File: destreak_limit_criterion.tex

1 Introduction

Images of the `ACIS-8` chip show a variable pattern of linear streaks that fall along rows of pixels with constant `CHIPY`. The `destreak` tool identifies and removes these streaks by exploiting the fact that each streak occurs within one frame and deposits multiple events along one `CHIPY` row of a node. For low count-rate observations, source events rarely cluster in this way. The `destreak` filtering algorithm[1] and the properties of the streak events[2] are described elsewhere. The main purpose of this memorandum is to suggest a change in the threshold that limits the degree of `destreak` event filtering when the streak rate is unusually high or when a bright source is present on `ACIS-8`.

2 `destreak` is Too Conservative

In most cases, `destreak` removes the majority of `ACIS-8` streak events with negligible loss of source X-ray events. However, a few observations have experienced a streak rate high enough to cause `destreak`'s self-limiting mechanism to reduce the effectiveness of the streak filter more than was necessary (*e g.* obsid 15543). Further examination of the streak rate distribution in archival data suggests that the current default row-loss upper limit criterion is too conservative.

By definition, a streak is the occurrence of *more than* $N_{\rm streak}$ events in a single `CHIPY` row, of a single node, within a single frame. For an observation, `destreak` determines the value of $N_{\rm streak}$ separately for each node by iteratively examining the population of streaks defined by a particular choice of $N_{\rm streak}$. Initially, $N_{\rm streak} = 1$, meaning that the occurrence of two or more events in a single row/node/frame is interpreted as a streak. Therefore, on the first iteration `destreak` removes as many streaks as possible. A streak is removed by discarding all events in the relevant row/node/frame. The

[1] `http://space.mit.edu/cxc/docs/docs.html#s4streak_alg`
[2] `http://space.mit.edu/cxc/docs/docs.html#s4streak_prop`

max_rowloss_fraction parameter specifies the allowable upper limit on the total number of rows discarded from each node as a fraction of the total number of rows read out. destreak enforces this limit by increasing N_{streak} on each node, reducing the number of streaks detected on each node, until the fraction of rows discarded falls below max_rowloss_fraction.

To define the limit criterion more precisely, consider an observation in which N frames are read out, with M rows per frame so that a total of NM rows are read out. Define $\delta_{iyn} = 1$ if row y is discarded from node i in frame n, and $\delta_{iyn} = 0$, otherwise. The total number of streak rows discarded from node i in the entire observation is then,

$$\delta_i = \sum_{y=1}^{M} \sum_{n=1}^{N} \delta_{iyn}, \tag{1}$$

so that a fraction, $\delta_i/(NM)$, of all rows read out are discarded from node i. Note that this fraction may be interpreted as the mean row-loss fraction (or row-loss probability) for node i, averaged over the duration of the observation and over all CHIPY rows. To simplify notation, drop the explicit node index and define the mean row-loss fraction, $\mathrm{mean}(f_y) \equiv \delta_i/(NM)$. The parameter max_rowloss_fraction specifies the maximum allowable value of $\mathrm{mean}(f_y)$; the current default is max_rowloss_fraction $= 5 \times 10^{-5}$.

In an effort to minimize the number of source events that might be discarded along with the streak events, the current default max_rowloss_fraction imposes a very conservative upper limit on the total number of rows discarded. Observations that trigger this self-limiting mechanism will be incompletely filtered because the "faintest" streak rows will not be removed. When such incomplete filtering has been noticed during manual validation and verification of new observations, the problem has been corrected by re-running destreak with a larger value of max_rowloss_fraction, provided that such a change did not adversely affect the scientific utility of the data.

3 A Simple Near Term Solution

In the near term, increasing the default value of the limit parameter to max_rowloss_fraction $= 10^{-3}$ will ensure more effective and consistent streak filtering. The same parameter default should work for both timed-exposure (TE) and continuous clocking (CC) mode, and for both grating (HETG or LETG) and non-grating data. See §5 for some discussion justifying this choice.

4 A Better Long Term Solution

In the longer term, a better solution is to use a more sensitive limit criterion. The main problem with the current implementation is that the limit criterion on the total number of rows discarded is effectively node-averaged, making it less sensitive for distinguishing the presence of a bright source from an unusually high streak rate. A more sensitive criterion would impose a limit based on the maximum exposure time lost by any single row or, equivalently, the maximum number of frames in which any single row is discarded.

Using the notation established above, the number of frames in which row y is discarded from node i is

$$\delta_{iy} = \sum_{n=1}^{N} \delta_{iyn}, \tag{2}$$

so that the maximum number of frames discarded from any row on node i may be written as $\max_y(\delta_{iy})$. Expressing this as a fraction of the total number frames, and dropping the explicit node index, we can define the maximum single-row loss fraction as, $\max(f_y) \equiv \max_y(\delta_{iy})/N$. Note that after multiplying the numerator and denominator by the frame time, τ, this fraction may also be interpreted as the maximum fraction of the total exposure time lost by any single row.

As long as the fraction of the total exposure time lost in any single row is small compared to the calibration uncertainty of the relevant effective area, any loss of source events must be negligible — with the exception of sources that are detected only through a small number of short, bright flares. However, events from such faint, flaring dominated sources are already at risk of removal by the hot-pixel/afterglow filter and in any case, a search for such sources using data strongly affected by `ACIS-8` streaks is questionable at best.

For the vast majority of practical cases, a conservative limit on the maximum exposure loss in any single row should be sufficient to guarantee that `destreak` has done no significant harm.

Implementing this new limit criterion would require only minor changes to `destreak`. Unfortunately, such changes would then require new regression tests, and some existing regression tests may need updating. A new `destreak` interface would require updates to `ahelp` documentation, data processing caveats and relevant threads and eventually, small changes to pipeline processing scripts may be required.

5 Analysis of Existing Data

To investigate the impact of an increase in `max_rowloss_fraction`, I used `destreak` to filter all public, non-calibration, `TE` mode `ACIS` observations with `ACIS-8` turned on, a set of 4683 obsids (3722 imaging, 961 grating) as of this writing in early June 2013. I also examined all public, non-calibration, `CC` mode `ACIS` observations with `ACIS-8` turned on, a set of 235 obsids (125 imaging, 110 grating). The streak rate distribution seen in the `CC` mode data is consistent with that seen in the `TE` mode data, but because the `TE` mode data is so much more common, all results presented here refer to the `TE` mode data unless otherwise stated.

To examine the worst-case impact, I processed all the level 1 event files with the self-limiting mechanism turned off, using `max_rowloss_fraction =1`. For each obsid, the `timefile` output from `destreak` records the total exposure time discarded from each row on each node. For each row, the exposure time loss reflects the number of frames in which a streak was flagged in that row. For example, consider an obsid with `EXPTIME`$= \tau$ and `EXPOSURE`$= T = N\tau$. Suppose that in m separate frames, `destreak` flags a streak in row $y =$ `CHIPY` of node i. Discarding these m rows reduces the total exposure in that row by an amount $\Delta T = m\tau$, corresponding to the loss of a fraction, $f_y = \Delta T/T = m/N$, of the total exposure in row y.

The upper panels of Figure 1 and Figure 2 show $\mathrm{mean}(f_y)$ for each node as a function of exposure time for imaging and grating data, respectively. Note that the two sets of observations are affected by the same streak rate distribution. The lower panels of Figure 1 and Figure 2 show the cumulative distribution of $\mathrm{mean}(f_y)$. Inspection of the cumulative distributions shows that even with worst-case filtering the majority of observations have $\mathrm{mean}(f_y) < 10^{-3}$.

The upper panels of Figure 3 and Figure 4 show $\max(f_y)$ for each node as a function of exposure time for imaging and grating data, respectively. By inspection, it is clear that the maximum exposure time discarded from any single row is almost always less than 1% of the total exposure — a loss that is comparable to or smaller than the uncertainty in the effective area. The lower panels of Figure 3 and Figure 4 show the cumulative distribution of $\max(f_y)$. Inspection of the cumulative distribution confirms that $> 99\%$ of obsids have $\max(f_y) < 0.01$, even when the limiting mechanism is turned off.

Comparing the distribution of $\mathrm{mean}(f_y)$ with the distribution of $\max(f_y)$ clearly demonstrates that the current limit parameter default of `max_rowloss_fraction` $= 5 \times 10^{-5}$ is too conservative. The current default causes the self-limiting mechanism to reduce the level of streak filtering for about 50% of all obsids, even though $\max(f_y)$ is almost always less than 1% with the self-limiting mechanism turned off.

Why then was such a conservative default chosen? Because, in principle, the node-averaged limit criterion can fail when the number of discarded rows are concentrated in a small range of `CHIPY` as might happen if the identified "streaks" are actually associated with a bright X-ray source. For example, if the same three adjacent rows are discarded from every third frame, the exposure in each of those rows is reduced by one third, but the node-averaged loss rate is only $\sim 10^{-3}$. Choosing a conservative `max_rowloss_fraction` $\ll 10^{-3}$ reduces the impact of such a failure, but also reduces the effectiveness of the streak filter for the common case where no bright source is present. The insensitivity of the node-averaged $\mathrm{mean}(f_y)$ limit criterion is the main motivation to switch to the more sensitive $\max(f_y)$ criterion.

Figure 5 compares the two diagnostics directly, showing that, compared with the distribution of $\mathrm{mean}(f_y)$, the upper cutoff of $\max(f_y)$ is flatter and farther away from the outliers. Figure 6 shows that the `CC` mode data yields a similar distribution of $\max(f_y)$ vs. $\mathrm{mean}(f_y)$.

Because a majority of existing observations have $\mathrm{mean}(f_y) < 10^{-3}$ and also satisfy $\max(f_y) < 0.01$, enforcing $\mathrm{mean}(f_y) < 10^{-3}$ on future observations lends confidence that $\max(f_y) < 0.01$ will also be satisfied, except for faint, flaring sources as mentioned above. Directly enforcing the criterion on $\max(f_y)$ would be better, but the simplest short term improvement is to increase the default value of the limit parameter to `max_rowloss_fraction` $= 10^{-3}$ to enforce the stricter criterion on $\mathrm{mean}(f_y)$.

The few obsids for which the worst-case $\max(f_y) > 0.01$ are listed in Table 1. Note that all of these outliers have exposure times $< 6\,\mathrm{ksec}$ (short exposures make it easier to exceed the 1% threshold needed to get on this list) and that most have an extremely bright source on `ACIS-8`— *e g.* Cas A, the Crab Pulsar, Cyg X-1, or LMC X-1 (extremely bright sources can mimic streaks when $>N_{\mathrm{streak}}$ source events occur in a single node/row/frame).

Table 1: Obsids with Worst-Case $\max(f_y) > 0.01$

				$\max(\Delta T)^c$			
		τ^a	T^b	`node 0`	`node 1`	`node 2`	`node 3`
`obsid`	Source	[sec]	[sec]	[sec]	[sec]	[sec]	[sec]
Imaging obsids (TE)							
231	Cas A	3.2	1036.8	16.0	316.8	272	6.4
1057	PKS0637-752	3.2	483.2	6.4	6.4	3.2	3.2
1078	LMC X-1	1.2	739.2	214.8	1.2	1.2	2.4
1079	LMC X-1	3.2	755.2	3.2	3.2	6.4	406.4
1080	LMC X-1	1.1	737.0	213.4	1.1	1.1	1.1
1081	LMC X-1	3.2	755.2	3.2	3.2	9.6	451.2
4016	NGC4419	3.2	810.6	6.4	6.4	9.6	6.4
13349	SDSS J0303-0023	3.1	1540.7	12.4	12.4	15.5	12.4
Grating obsids (TE)							
107	Cyg X-1	0.3	207.3	7.2	0.6	1.5	1.2
169	Crab Pulsar	3.2	3193.6	1008.0	16.0	3.2	19.2
334	Q0836+7104	3.2	185.6	3.2	3.2	3.2	3.2
433	MCG -6-30-15	3.2	1916.8	12.8	16.0	19.2	16.0
2741	Cyg X-1	1.7	1810.5	782.0	596.7	593.3	15.3
2742	Cyg X-1	1.7	1849.6	775.2	608.6	620.5	360.4
2743	Cyg X-1	1.7	2471.8	804.1	10.2	572.9	13.6
12313	Cyg X-1	1.3	3712.8	1381.9	13.0	5.2	3.9
13219	Cyg X-1	1.8	4465.8	1688.4	46.8	1411.2	19.8
Imaging obsids (CC)							
3463	RX J170930.2-263927	–	5491.3	49.4	59.6	49.4	37.8

[a] τ is the frame time, `EXPTIME`.
[b] T is the `ACIS-8` exposure time, `EXPOSUR8`.
[c] $\max(\Delta T) = T \max(f_y) = m_{\max}\tau$ is the maximum exposure time lost from a single row on the given node.

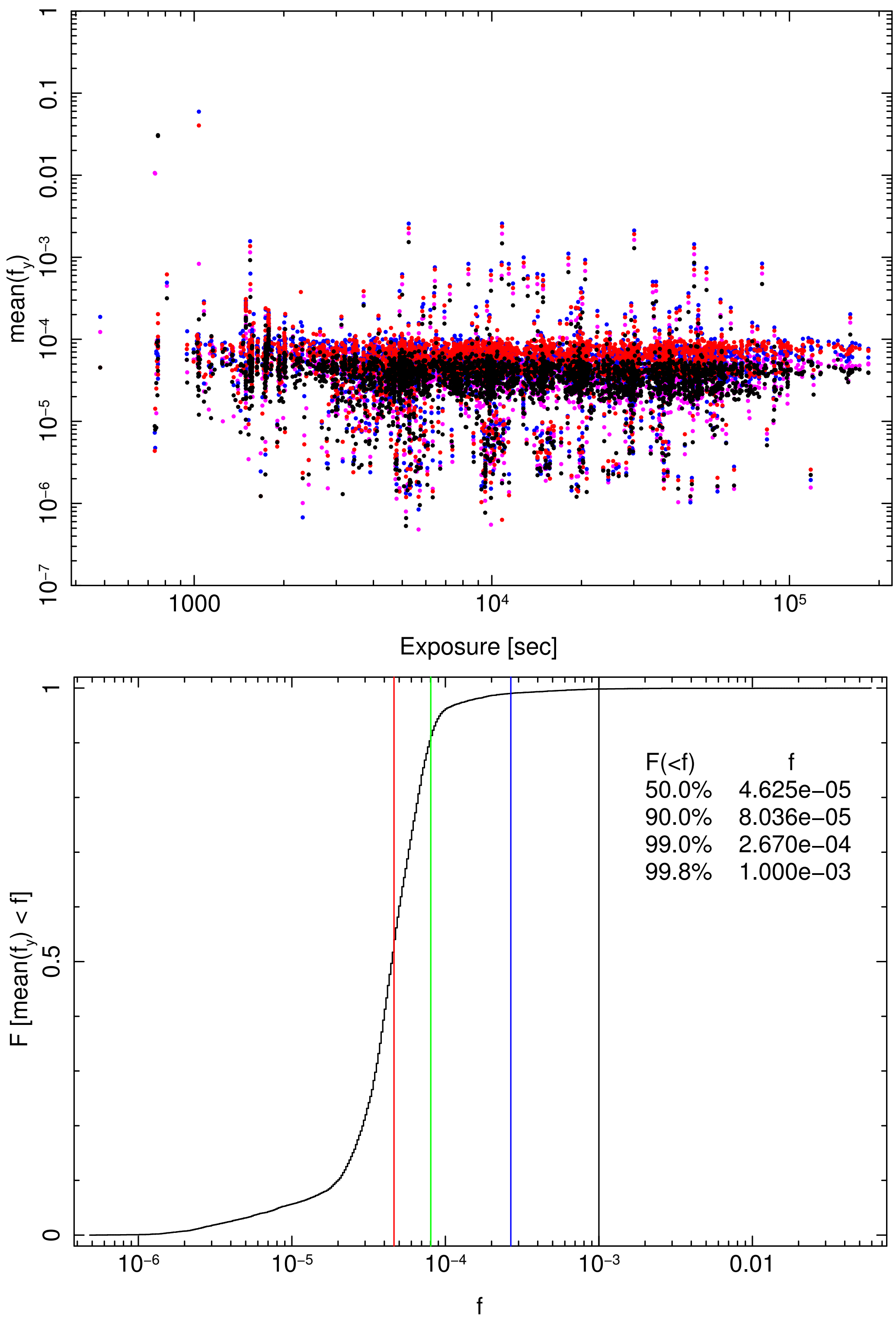

Figure 1: Upper panel: $\mathrm{mean}(f_y)$ vs. ACIS−8 exposure time for 3722 imaging mode obsids. Each point corresponds to a single node from a single obsid; different colors correspond to different nodes. Lower panel: Cumulative distribution of the quantity $\mathrm{mean}(f_y)$. Vertical lines indicate where $F = 0.5$ (red), $F = 0.9$ (green), $F = 0.99$ (blue), and $f = 10^{-3}$ (black).

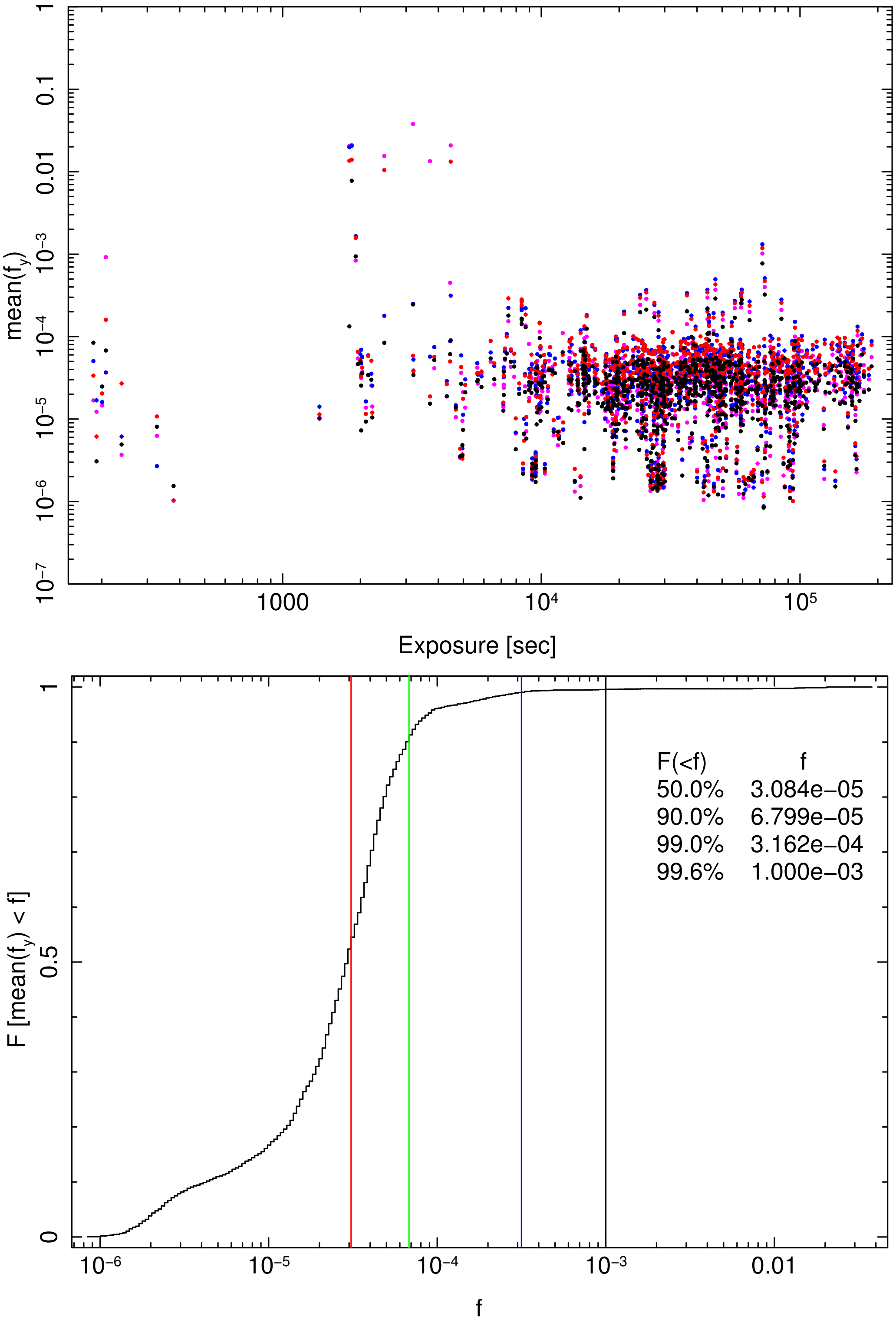

Figure 2: Upper panel: $\mathrm{mean}(f_y)$ vs. ACIS−8 exposure time for 961 grating obsids. Each point corresponds to a single node from a single obsid; different colors correspond to different nodes. Lower panel: Cumulative distribution of the quantity $\mathrm{mean}(f_y)$. Vertical lines indicate where $F = 0.5$ (red), $F = 0.9$ (green), $F = 0.99$ (blue), and $f = 10^{-3}$ (black)

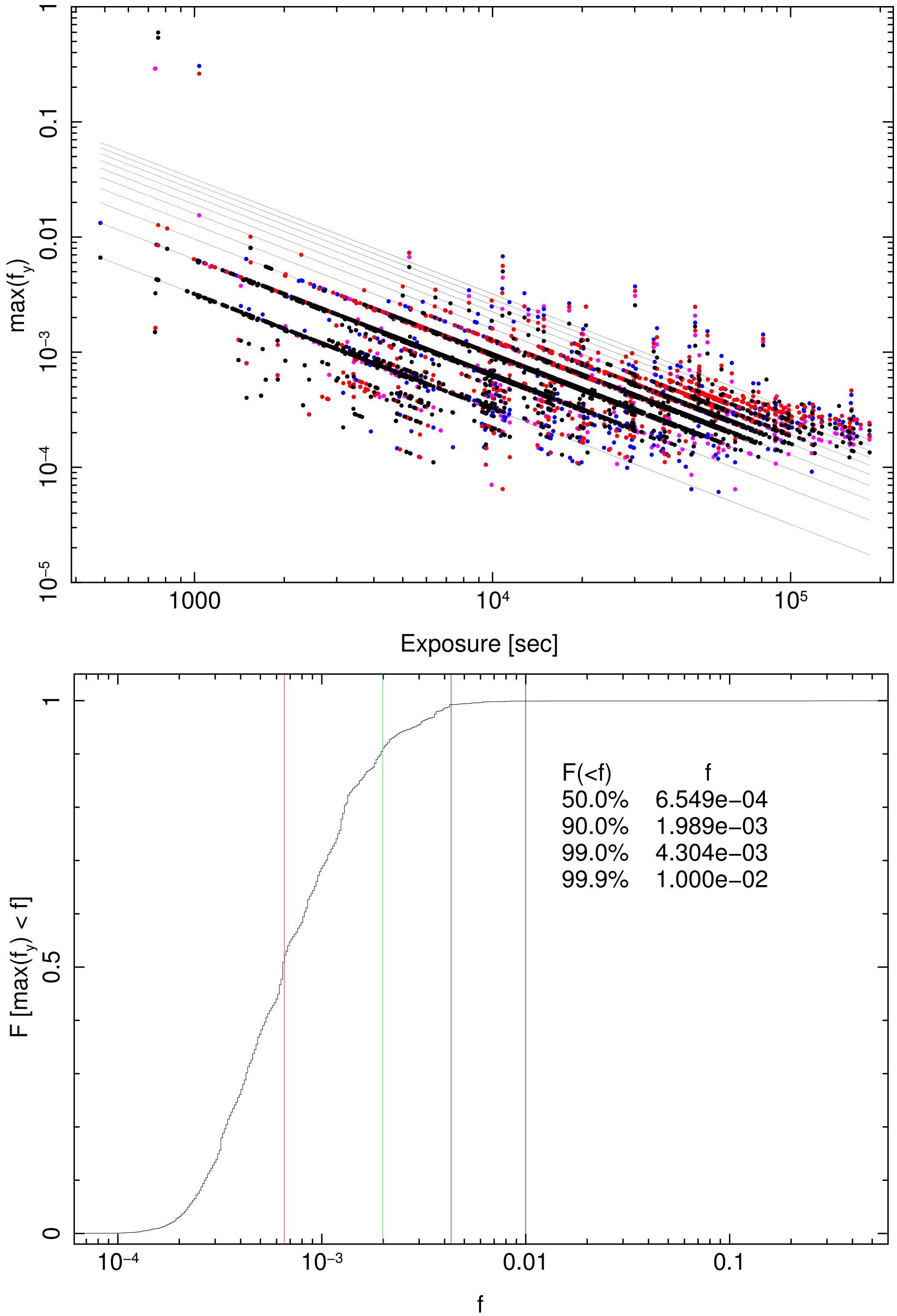

Figure 3: Upper panel: $\max(f_y)$ vs. ACIS−8 exposure time for 3722 imaging mode obsids. Each point corresponds to a single node from a single obsid; different colors correspond to different nodes. The grey diagonal lines correspond to a loss fraction of $3.2m/T$, where T is the ACIS−8 exposure time in seconds and $m = 1, 2, \ldots, 10$ is the maximum number of frames discarded from a single row. Lower panel: Cumulative distribution of the quantity $\max(f_y)$. Vertical lines indicate where $F = 0.5$ (red), $F = 0.9$ (green), $F = 0.99$ (blue), and $f = 0.01$ (black).

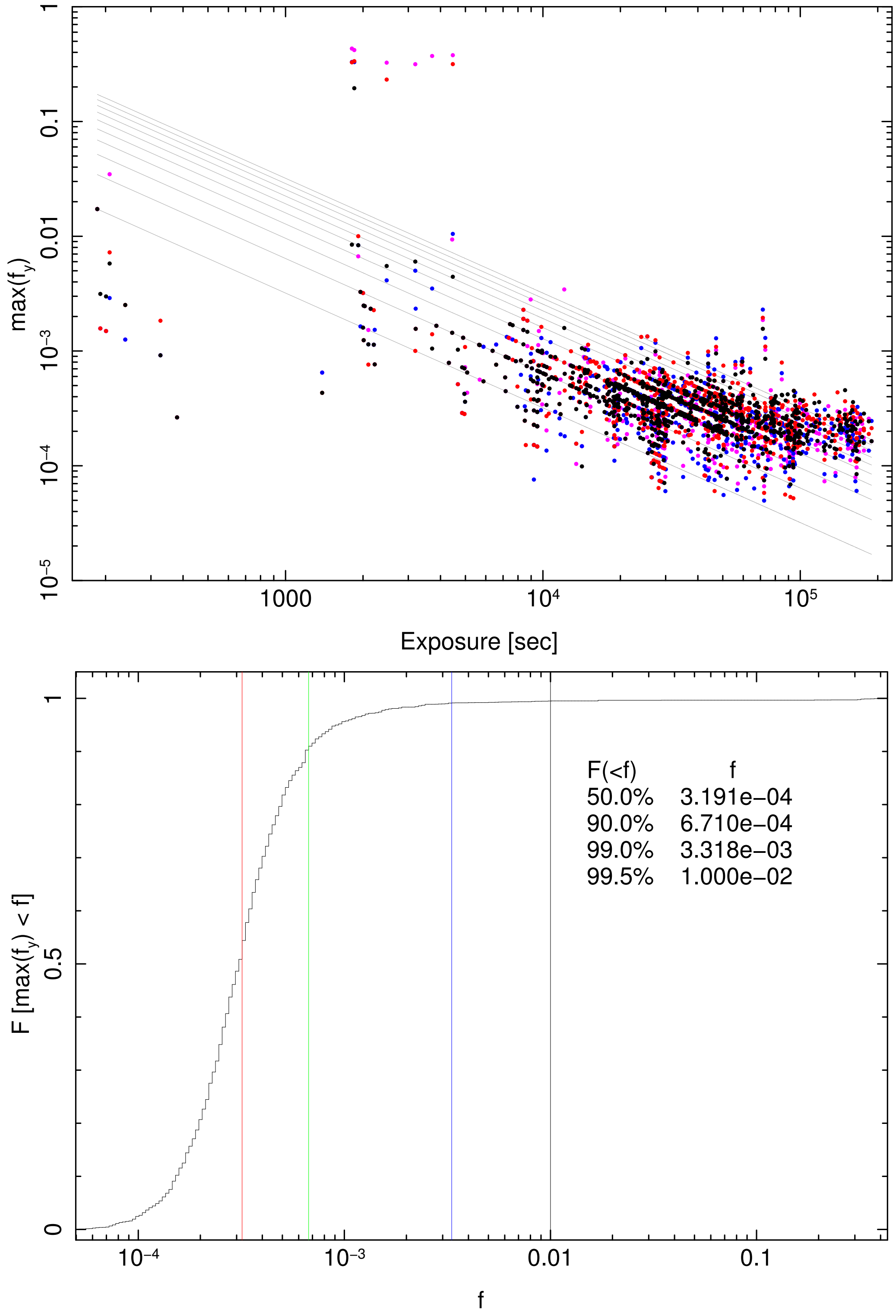

Figure 4: Upper panel: $\max(f_y)$ vs. ACIS-8 exposure time for 961 grating obsids. Each point corresponds to a single node from a single obsid; different colors correspond to different nodes. The grey diagonal lines correspond to a loss fraction of $3.2m/T$, where T is the ACIS-8 exposure time in seconds and $m = 1, 2, \ldots, 10$ is the maximum number of frames discarded from a single row. Lower panel: Cumulative distribution of the quantity $\max(f_y)$. Vertical lines indicate where $F = 0.5$ (red), $F = 0.9$ (green), $F = 0.99$ (blue), and $f = 0.01$ (black).

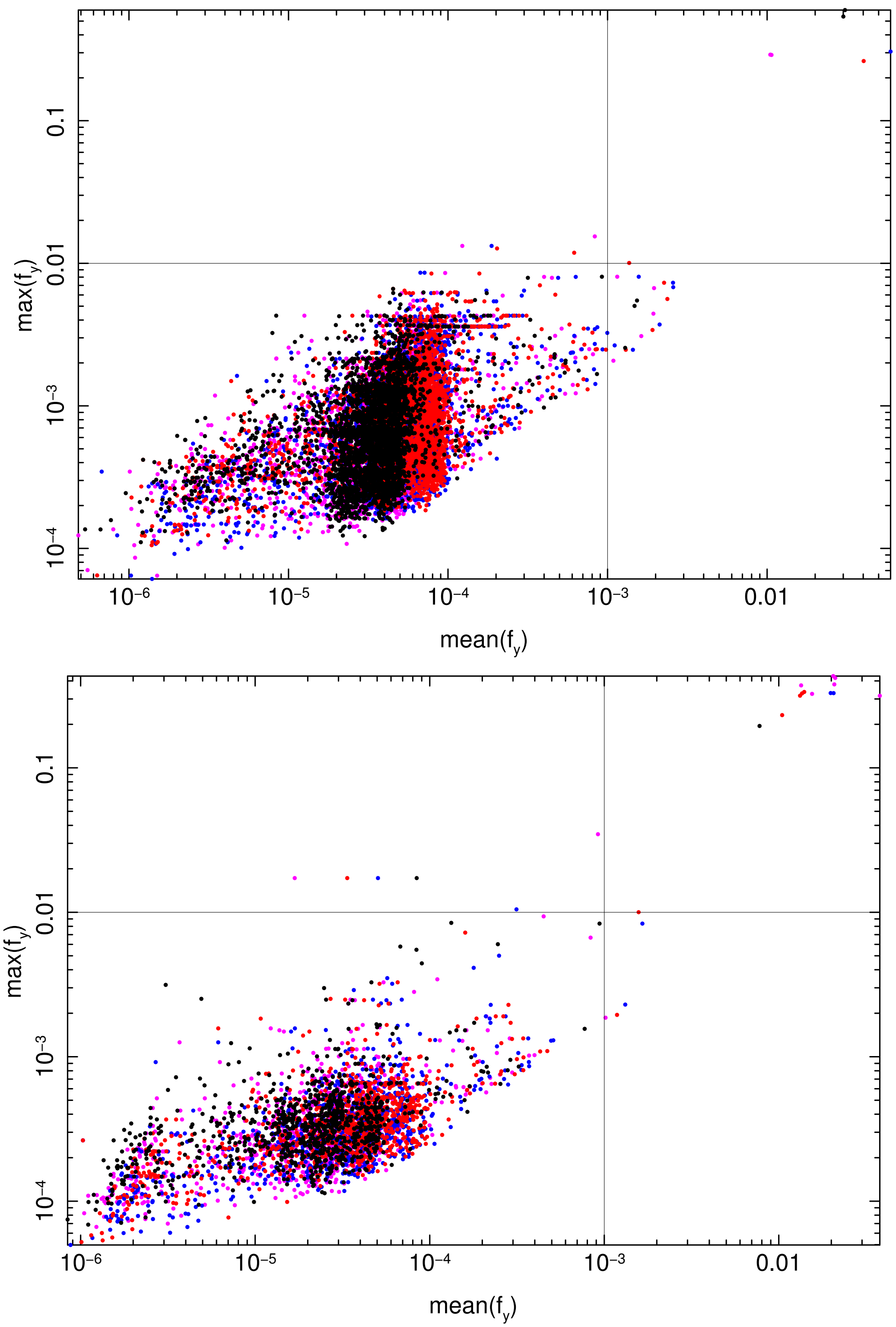

Figure 5: $\max(f_y)$ vs. $\mathrm{mean}(f_y)$ for 3722 imaging obsids (upper panel) and 961 grating obsids (lower panel). Each point corresponds to a single node from a single obsid; different colors correspond to different nodes. Horizontal and vertical lines indicate the suggested threshold criteria, $\max(f_y) = 0.01$ and $\mathrm{mean}(f_y) = 10^{-3}$.

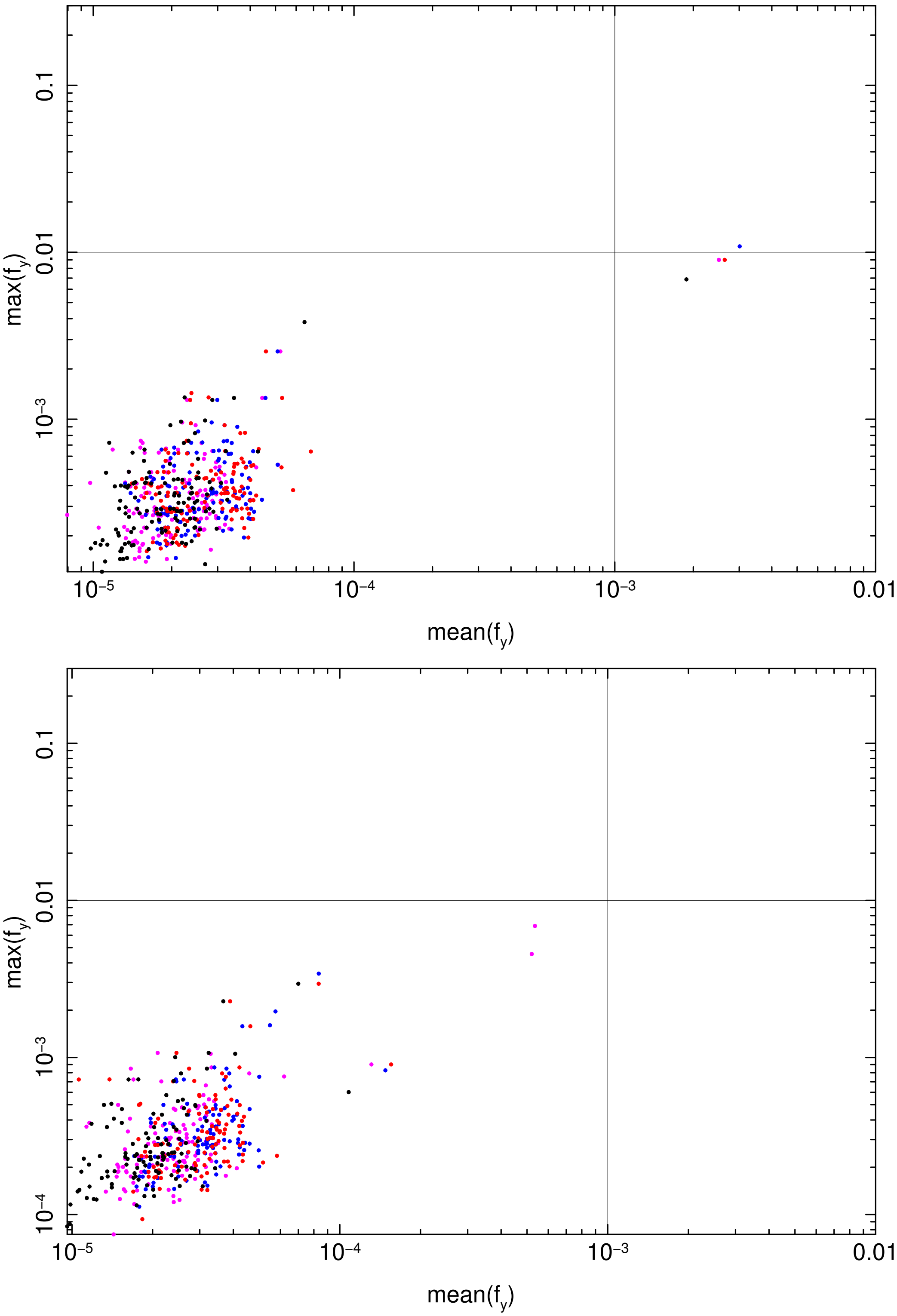

Figure 6: $\max(f_y)$ vs. $\mathrm{mean}(f_y)$ for 125 CC mode imaging obsids (upper panel) and 110 CC mode grating obsids (lower panel). Each point corresponds to a single node from a single obsid; different colors correspond to different nodes. Horizontal and vertical lines indicate the suggested threshold criteria, $\max(f_y) = 0.01$ and $\mathrm{mean}(f_y) = 10^{-3}$.