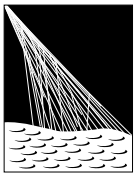
GAP-2022-0??

New Baseline Algorithm for UB Traces



Tobias Schulz, Markus Roth,
David Schmidt, and Darko Veberič

IAP, Karlsruhe Institute of Technology KIT, Germany

July 2022

Abstract

In order to acquire an unbiased signal from a time trace, it is necessary to correctly identify and subtract a baseline. After significant signal, the output of photomultipliers is visibly reduced, resulting in an undershoot of the baseline that recovers exponentially with a characteristic decay constant. Accidental muons or late air shower components result in signal contributions in the traces that may complicate the estimation of the baseline. The current baseline algorithm for SD time traces implemented in Offline is removing signal that is mistakenly determined to be baseline. Here, we present a baseline finding algorithm for UB traces that is robust to sporadic early and late signal contributions in the trace and that accounts for undershoot.

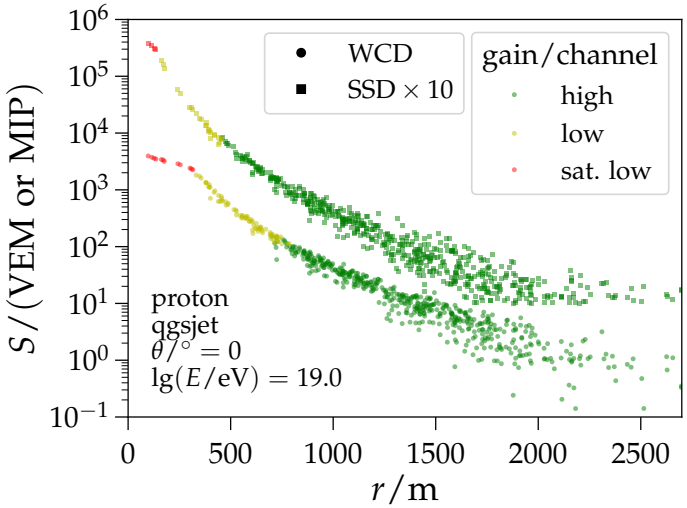


Figure 1: Lateral distribution of WCD and SSD measurements. The gains of both detectors saturate at different points, as depicted by the transition between colors.

1 Introduction

The identification and calibration of detector signals is crucial for minimizing systematic uncertainties in measurements. A constant offset, called the baseline, which is generated by the electronics, has to be determined to properly estimate the size of the signals. To measure the signal produced by the particles that enter the detectors, photomultiplier tubes (PMTs) are used to collect the emitted Cherenkov or scintillation light. The PMTs have one anode output and one at the last dynode, that is amplified by a factor of 32. The analog pulses are read out in one low- and one high-gain channel and sampled with a Flash Analog to Digital Converter (FADC) to produce time traces. For the UB, a 10 bit FADC that allows for a maximum of 1023 adc (ADC counts) per bin is used. The low-gain channel (LG) is connected to the anode and the high-gain channel (HG) to the last dynode, to cover a wide dynamic range [1]. For the UUB, both gains are derived from the anode signal, with the HG amplified by a factor of 32. The resolution is increased by using a 12 bit FADC [2]. For high energy events or stations close to the core, the signal gets too large and the high-gain channel may saturate. In this case, the signal is estimated from the low-gain channel. Various studies have reported systematic biases that might be due to a bias in the ratio between the signal of the high-gain channels and the signal of the low-gain channels [3, 4]. With the AugerPrime upgrade, the new SSDs are used in addition to the measurements of the WCD. Both detectors have differing signal thresholds for saturation, thus the transition point of which gain is used, is different. Fig. 1 shows a simulation of a lateral distribution of WCD and SSD measurements for the estimation of the shower size S_{1000} and thus the energy of the shower. The existence of a systematic bias in the ratio of the signal from the two gains would imply a shift of the signal used when fitting the lateral distribution at the transition point from HG to LG. It is therefore of importance to understand and minimize this systematic bias.

The following analysis is performed on data that was recorded by WCDs using a UB. To get an estimate of the systematic bias between the HG and LG, a data set of 770 664 events from years 2019 to 2021 of the full array with a total of 7 630 454 traces is analyzed. For each PMT, the HG and LG trace is extracted. The maximum bin entry of the HG channel gives an estimate of how close the trace is to the saturation of the HG channel. The range of ADC counts (adc) has a maximum of 1023 per time bin. Signals above this value are digitized as 1023 and the trace is thus saturated. If a HG trace reaches 1023 adc in a time bin, the LG trace is used instead. The maximum time-bin entry of the full HG trace is denoted with $\mathrm{ADC}_{\mathrm{max}}$. At $\mathrm{ADC}_{\mathrm{max}} \equiv 1023$, the HG trace is saturated and the LG is used. The ratio of the calculated signal from the HG S_{HG} and LG S_{LG} over the

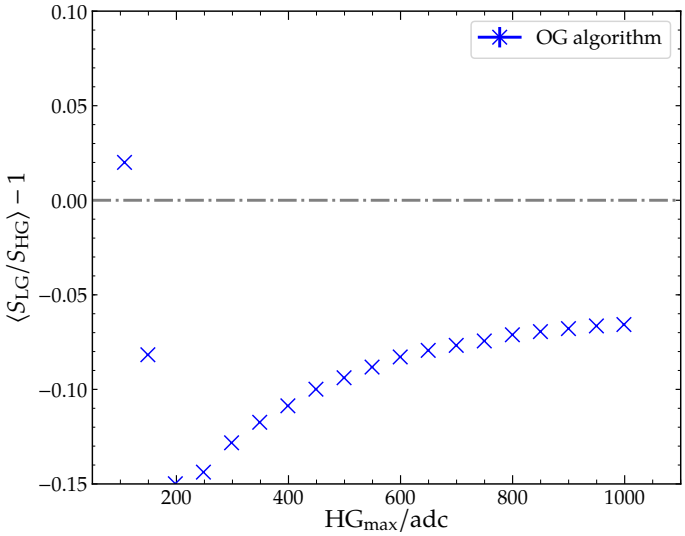


Figure 2: The ratio between signals from LG and HG traces as a function of the maximum bin entry in the HG channel, which provides an indication of how close the HG channel is to saturation. The algorithm used to determine the baselines of the time traces is referred to as the OG algorithm here. At low counts, the signals in the LG channel are too small to give reliable results, but at larger values and at the transition point from HG channel to the LG channel, the LG signals are on average 6.5% smaller than the HG signals.

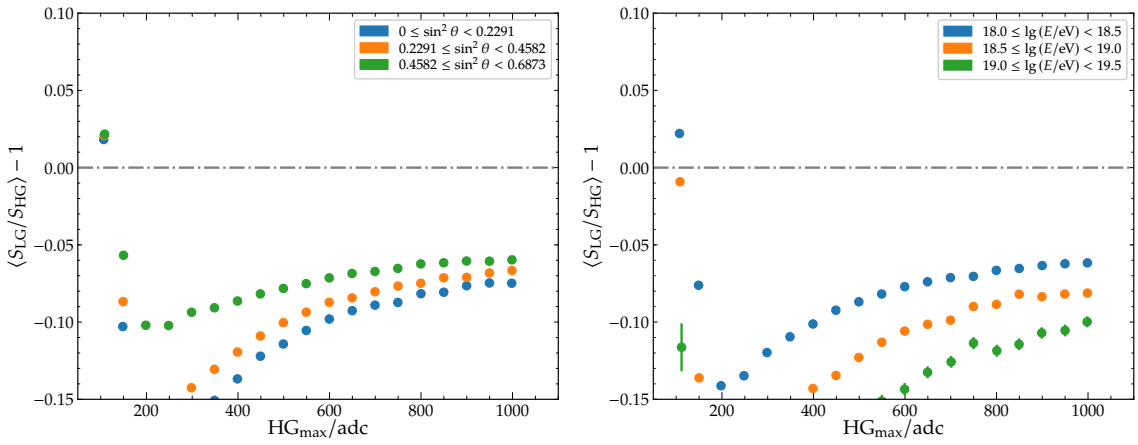


Figure 3: The ratio between the signals calculated from the low- and high-gain traces binned in zenith angle and energy. For showers with larger energies and smaller zenith angles, the bias is the largest. The difference in the bias between zenith angles ranges is smaller in comparison to the difference in the bias between the energy bins.

maximum bin entry in the HG channel ADC_{max} is shown in Fig. 2. For low ADC_{max} ranging up to a few hundred adc, the signal is too small to perform a useful estimate of the signal from the LG channel. At the transition from HG to LG, there is a negative systematic bias of about 6.5%. For the UB, this bias can have its origin in either the dynode-to-anode-ratio (D/A ratio, gain ratio) or in the calculation of the baselines from the traces.

The ratio of the LG and HG signals for different zenith and energy bins is shown in Fig. 3. The three bins in zenith angle cover a range of 0° to 56° and are uniform in $\sin^2\theta$. At the transition from HG to LG, the bias of the ratio is the smallest at approximately −6% for almost vertical showers. For the more inclined showers up to 56°, the bias at the transition is the largest with

about −7.5%. In Fig. 3-left, the ratio is plotted in 3 energy bins, ranging from $\lg(E/\mathrm{eV}) = 18.0$ up to an energy of $\lg(E/\mathrm{eV}) = 19.5$. The difference between the 3 energy bins is larger than for the zenith bins at the transition point from HG to LG. For showers with a reconstructed energy between $\lg(E/\mathrm{eV}) = 18.0$ and $\lg(E/\mathrm{eV}) = 18.5$, the bias is the smallest at approximately −6%. In the largest energy bin, the bias increases up to almost −10%. These differences can be attributed to the shape of the traces that is different for the energy and zenith bins. The electromagnetic component of showers increases with decreasing zenith angles, thus broad signal shapes appear more often in the traces, that are filtered out by the OG algorithm. Similarly broad signal shapes appear in showers with larger energy that might be removed by the OG algorithm and thus result in a larger bias of the ratio between the HG and the LG traces.

2 Current baseline algorithm

The baseline finder looks for flat segments in the trace and interpolates a baseline in between them. It has been shown that this procedure, in its current implementation, accidentally identifies parts of the signal as baseline pieces and thus gives an overestimation of the baseline. A correction procedure by Ronald Bruijn, has been implemented in the CDAS and $\overline{\text{Off}}\underline{\text{line}}$ reconstruction but is not used by default [5]. The plots in Fig. 4 show examples of time traces and the estimation of their baselines, calculated by the `BaselineFinderOG` module of $\overline{\text{Off}}\underline{\text{line}}$. The baseline algorithm used for the calculation will be now referred to as "OG algorithm". The first two traces of Fig. 4 show common examples of the misidentification of flat pieces. While the algorithm finds large flat pieces at the front and end of the trace, it also can misidentify relatively flat signal contributions as a flat piece. These misidentified flat baseline pieces result in multiple random jumps of the final, interpolated baseline, as shown in the first plot, or in single jumps, as shown in the second plot. In more extreme cases, the algorithm tries to identify flat pieces on extended tails of traces as in the third plot or on small, late signals as in the last plot of Fig. 4. This results in the unwanted removal of these signals. It is therefore of interest to find a baseline model that is based on physical considerations and does not remove relatively flat signal contributions.

3 New baseline algorithm

3.1 Physical trace model

PMT traces can exhibit various features that should be taken into consideration in the development of a baseline finding algorithm. Therefore, a physical trace model was developed as a first step. An example trace of one PMT of a WCD is shown in Fig. 5. In the first 40 time bins of the trace, a small peak is visible. This peak is very likely caused by an "accidental" particle entering the tank, which did not originate from the actual shower. Next follows a "flat" segment, where the baseline can be roughly estimated by eye to about 58 adc. At around the time bin 240 the main part of the signal of the event is visible. After the main signal, at around the time bin 340, smaller peaks from shower components occurring later are visible. The baseline after the signal is lowered by approximately 1 adc, due to the undershoot of the PMT after a large signal. It has been shown that the recovery of the undershoot takes about 300 µs, which results in a systematic overshoot of the trace. This overshoot then slowly decreases over a time span of up to 1 to 1.5 ms [6]. The total trace length is 19.2 µs and thus, the baseline after the undershoot can be assumed constant. The undershoot ΔB can be calculated as the difference between the baseline at the end and at the beginning of the trace,

$$\Delta B = B_{\text{end}} - B_{\text{front}}. \tag{1}$$

The undershoot linearly relates to the total charge q_{tot} of the trace. Fig. 6 shows the linear relationship, which is different for HG and LG traces. In considering this trace, one can define a few requirements for the new baseline algorithm.

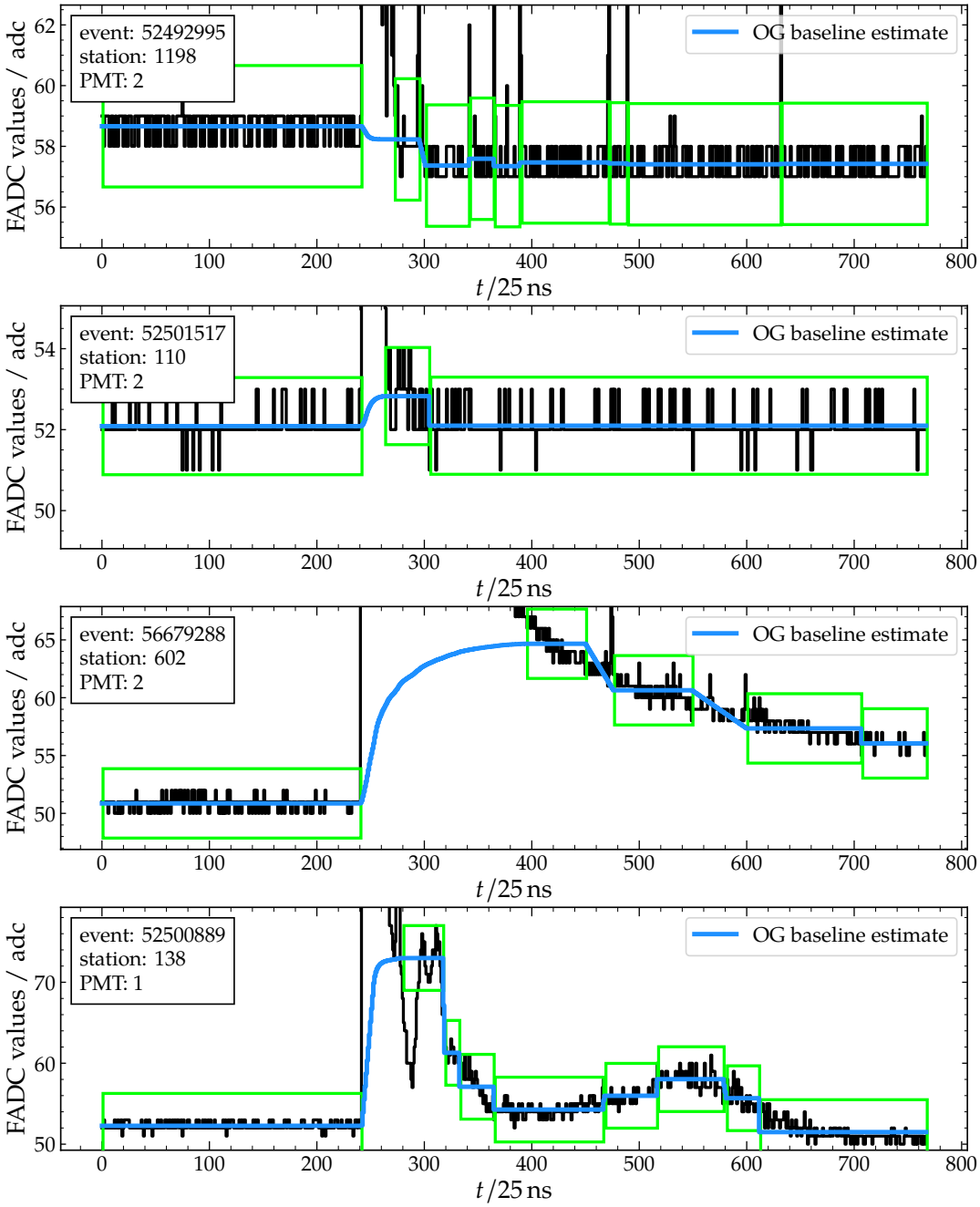


Figure 4: Example traces with the baseline estimate of the `BaselineFinderOG` module of Offline. The algorithm tries to find flat baseline pieces, which are marked with green boxes, and interpolates in between them. Part of the signal might be falsely identified as a flat piece and is consequently removed from the final trace.

1. The calculation of the baseline should be stable against any signal contributions, as for example from accidental particles.

2. Two separate baseline estimates at the beginning and the end of the trace are required, to account for a possible undershoot.

A new, three step algorithm will be discussed in the following sections in more detail. In the

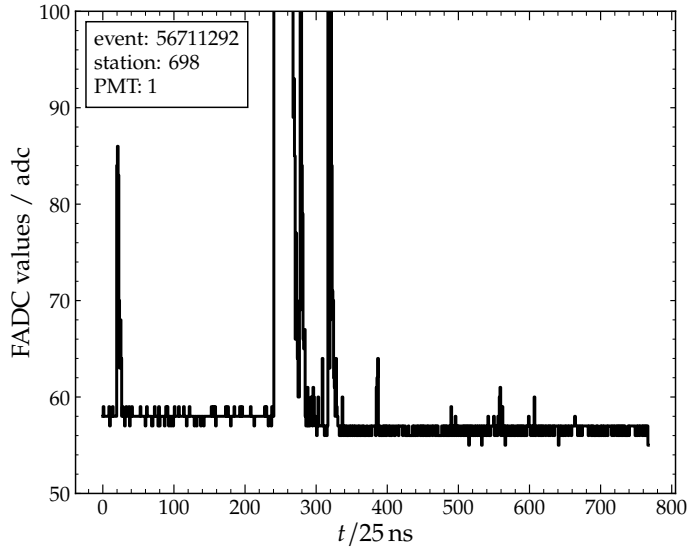


Figure 5: Zoomed in example trace of an event. In the first 40 bins, a peak from an accidental particle is visible. After the main part of the signal, at around 340 bins, the baseline of the trace is lower than during the first 200 bins of the trace.

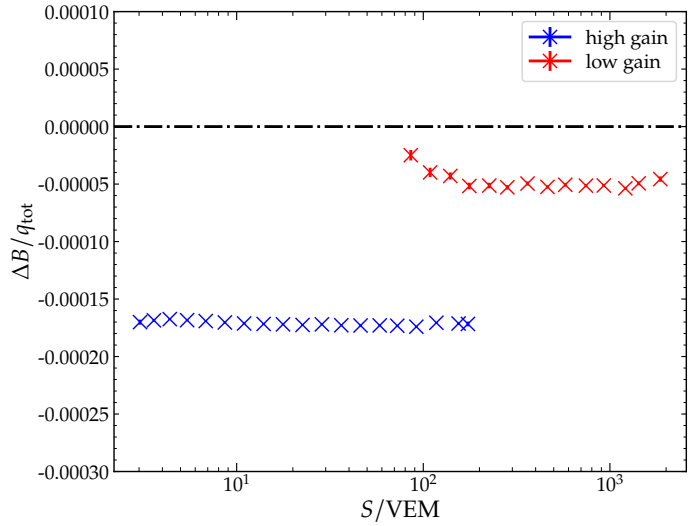


Figure 6: Relationship of difference ΔB between front and end baseline estimates and the approximate total charge q_{tot}. As the signal in the HG and LG increases, a linear relationship of the ratio between ΔB and q_{tot} becomes evident.

first step, an estimate of the baseline at the beginning and the end of the trace is derived. The algorithm that is used to derive these two robust estimates is adapted from the baseline estimation of the FUNK experiment [7]. From these two estimates, it is decided in the second step which interpolation method between the two estimates is best suited. In the last step, the interpolation between the two estimates is executed to obtain an estimation of the full baseline. The implementation of this new algorithm is done in a new module in Offline, named `BaselineFinderKG`. This new algorithm will be referred to as the "KG algorithm".

Table 1: Continuous Napoli library of CORSIKA simulations with different combinations of parameters. For each combination, 500 events were reconstructed.

primary	p, Fe
hadronic interaction model	EPOS-LHC
lg(E/eV)	18.5 - 19.0, 19.0 - 19.5, 19.5 - 20.0
θ/°	0 - 65

3.2 Front and end baseline estimation

As an initial step, two separate estimates of the baseline are derived. The first baseline estimate B_{front} is performed is done at the beginning of the trace. It is defined as

$$B_{\text{front}} = \frac{1}{N} \sum_{i=1}^{N} T_i, \tag{2}$$

where N is the length of the trace segment used to calculate the baseline. Due to the undershoot, the PMT output is reduced after a large signal and the baseline has to be reevaluated. Since the recovery time of the undershoot significantly exceeds the total trace length, the baseline can be assumed constant after the undershoot. The second baseline estimate can thus be evaluated at the end of the trace and is denoted as

$$B_{\text{end}} = \frac{1}{N} \sum_{i=L-N}^{N} T_i, \tag{3}$$

where $L = 768$ is the trace length of the UB stations [1]. The baseline can be estimated by calculating the mean of a segment of the trace that does not include any signal. The length N of this trace window, from which the baseline is calculated, has to be determined first. Optimally, the length of the trace window is chosen to be as short as possible to exclude possible signal contributions from accidental particles. To determine the optimal length, a set of 3000 simulated events with 98 457 traces is used. An overview of the simulated library is given in Table 1.

Even with a short trace window, some signals can still be included in the baseline calculation and the baseline will be overestimated when calculating the mean. Thus, an additional truncation is applied to the trace to exclude such signals from accidental particles. First, the mode m and the standard deviation σ of the trace window with a fixed number of bins are calculated. Then the trace is truncated by excluding all bins that have an ADC count larger or smaller than $m \pm 2\sigma$. To avoid the exclusion of all bins and since the ADC values are discrete integers, the $m \pm 2\sigma$ is floored or rounded up, to always include the next smaller value at downwards fluctuations, or larger value at upwards fluctuations. The standard deviation of the truncated trace is calculated again and the truncation with $\pm 2\sigma$ relative to the mode is repeated. From the final truncated trace piece, a mean is calculated as an estimate of the baseline. The estimated baseline B_i is compared to the true, simulated baseline B_{mc} and the bias $(B_i - B_{\text{mc}})$ and the resolution $\sigma(B_i - B_{\text{mc}})$ can be calculated. For different trace window lengths, the procedure is repeated, and the results are shown in Fig. 7 for the HG and in Fig. 8 for the LG. Due to the truncation, the signal from accidental particles is excluded and the bias of the front baseline is minimized. For a trace window size larger 200 time bins, the signal of the shower can already be included in the window, which can not be fully filtered out by the truncation. As a result, the bias of B_{front} increases. The baseline estimation at the end of the trace is almost bias-free as well. This is due to small signal contributions of the shower, that can not be entirely filtered with the truncation. By increasing the trace window B_{end}, the bias increases as more signal from the shower is swept up.

[1] For UUB $L = 2048$ and for Cyclon boards the trace length is only $L = 760$ bins [8]

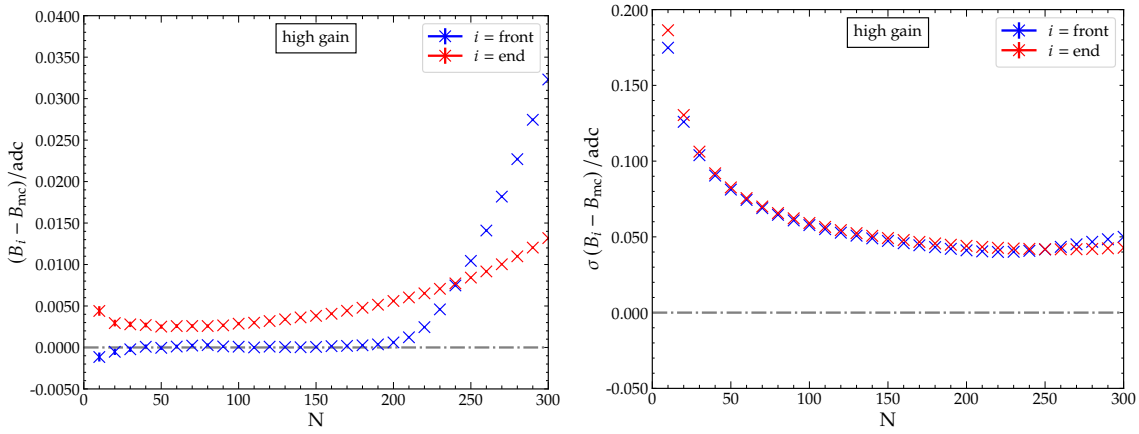


Figure 7: Bias and its resolution of the calculated baseline for various lengths of the trace window for HG traces. The front baseline, which is depicted in blue, gets heavily biased by the shower signals starting at around time bin 200. The bias of the end baseline, shown in red, increases with larger window sizes, as more late shower signals are included. By including more bins in the window, the resolution improves until shower signals start to be included in the window. *Left:* bla bla

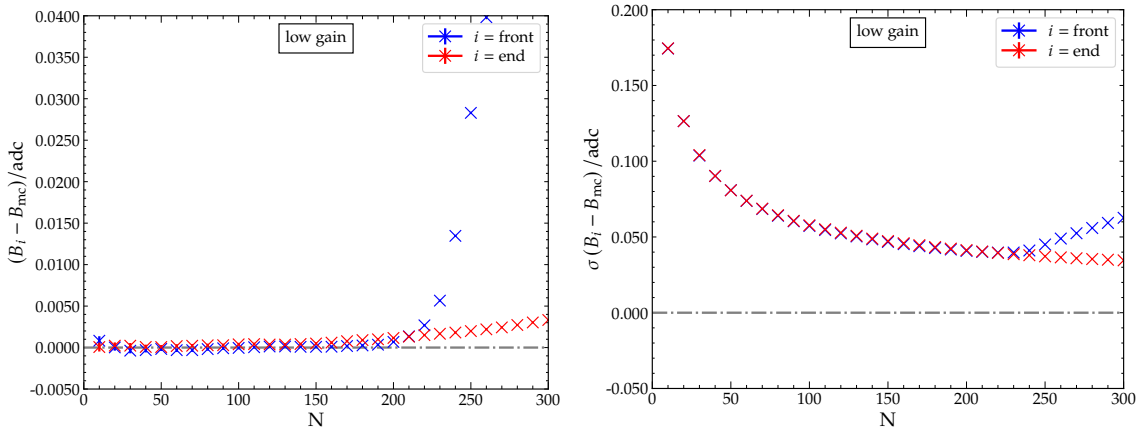


Figure 8: Bias and its resolution of the calculated baseline for various lengths of the trace window for LG traces. The front baseline is shown in blue, the end baseline in red. B_{front} gets heavily biased by the shower signals, starting at around time bin 200. Since the amplitudes in the low-gain traces (in adc units) are less than for their respective high-gain traces, the bias is smaller, although the resolution stays the same.

For the LG traces, this bias is reduced, since small signals can not be resolved in the trace. The resolution of the baseline estimates improves with larger trace windows, as one would expect. As soon as the shower signal is included in the window, the resolution worsens. From this analysis, it is possible to choose a trace window between 10 and approximately 180 bins, without a significant increase in the bias.

Since the criterion is to use as few bins as possible a trace window of 100 time bins is considered a good choice. The resolution of the baseline estimates with a trace window size of 100 time bins is approximately

$$\sigma(B_{\text{front}}) \approx 0.058\,\text{adc} \qquad \text{and} \qquad \sigma(B_{\text{end}}) \approx 0.059\,\text{adc}, \tag{4}$$

for the high-gain channel and

$$\sigma(B_{\text{front}}) \approx 0.057\,\text{adc} \qquad \text{and} \qquad \sigma(B_{\text{end}}) \approx 0.058\,\text{adc}, \tag{5}$$

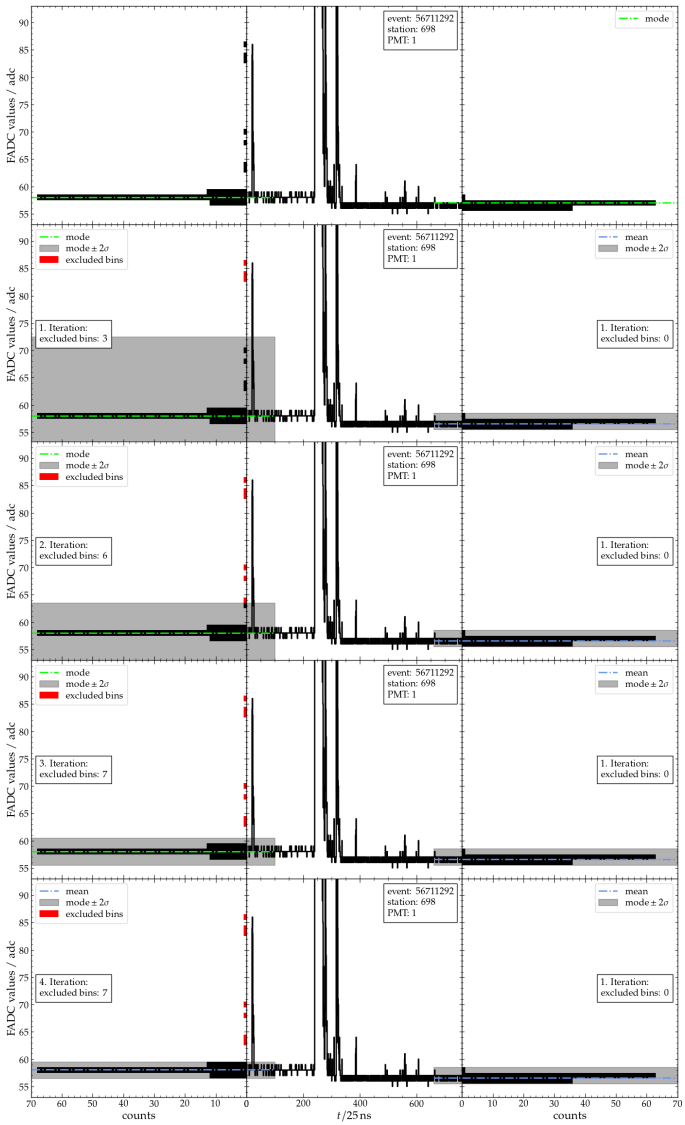


Figure 9: Example of the truncation procedure for the trace of Fig. 5. The front and end trace windows have a length of 100 time bins.

for the low-gain channel. With the trace window set to 100 time bins, the front and end baseline estimates are calculated in a way similar to the first rough determination with the simulated traces.

Fig. 9 shows the step-by-step procedure of the front baseline estimate for the example trace from Fig. 5. In the initial step, the mode m and standard deviation σ of the trace window, consisting of the first 100 time bins for the front baseline and the last 100 time bins for the end baseline, is calculated. The thresholds for the truncation of the trace are set relative to the mode to

$$m \pm 2\sigma. \tag{6}$$

The trace can only have discrete integer values and thus the resulting upper threshold is rounded upwards to include the next adc value and the lower threshold is floored similarly. All bins with adc values larger or smaller than the threshold are then excluded from the trace window. The standard deviation is calculated again and a new threshold is calculated in the same way as in

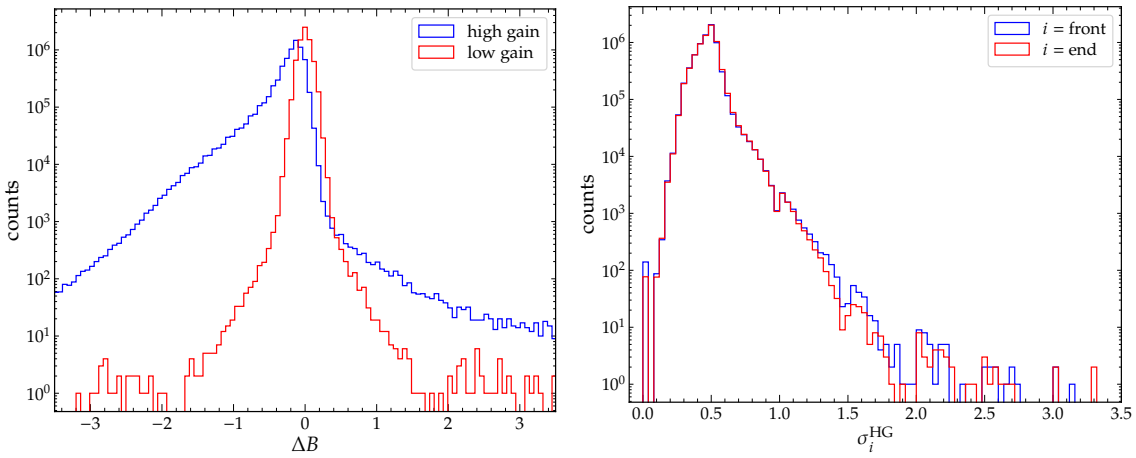


Figure 10: Anomalous traces can be found, based on their ΔB as well as σ_{front} and σ_{end}. A negative ΔB is expected. A large positive ΔB or large values of σ_{front} and σ_{end} can indicate possible problems such as malfunctioning PMTs. *Left:* Distribution of the difference between the front and end baseline estimate for high-gain and low-gain traces. The tail at negative values of ΔB corresponds to the undershoot. *Right:* Distribution of the calculated standard deviations of the trace windows of the front baseline estimate (blue) and end baseline estimate (red) for high-gain traces.

Eq. (6). This procedure is repeated as long as time bins are excluded. As soon as no more time bins are excluded, the process stops and the mean is calculated from the remaining time bins as the resulting baseline estimate. If the number of remaining time bins $n_{\mathrm{front,end}}$ for the baseline estimation is lower than 40, no baseline is calculated and the trace will not be used.

3.3 Decision of interpolation method

After a front and end baseline estimate have been determined, the baseline difference ΔB can be calculated as

$$\Delta B = B_{\mathrm{end}} - B_{\mathrm{front}}. \tag{7}$$

The left plot of Fig. 10 shows the distribution of ΔB of the traces from the SD data set from years 2019 to 2021 that did not saturate their corresponding channel. The high-gain traces show a long tail of the distribution to negative ΔB values, which corresponds to an end baseline that is below the front baseline. This is an expected behavior due to the undershoot that occurs at large signals. The tail is more prominent for high-gain traces. The right plot of Fig. 10 shows the distribution of the standard deviations σ_{front} and σ_{end}, calculated from the trace windows of the high-gain traces. The median of the distributions is approximately at a value of 0.49 for both the front and end trace window. Due to the conversion of the analog signals to discrete adc values, some traces may have a $\sigma_{\mathrm{front,end}}$ of zero since minimal fluctuations can not be resolved.

In Fig. 11 the effects of the discretization from an analog to a digital signal are shown. Multiple traces with the size of 1 000 000 time bins are created with varying analog means μ and standard deviations σ_{analog}. Each trace is then rounded to integer values and the mean $\langle T \rangle$ and standard deviation $\sigma(T)$ of each rounded trace is calculated. The calculated mean is shown as a function of the analog fluctuation σ_{analog} in Fig. 11-left. If the fluctuations of the analog values are sufficiently large, the mean of the rounded trace will correspond to the true mean of the trace. However, with decreasing σ_{analog}, the mean of the trace will take the values of the half integers due to the rounding of the values. The standard deviation of the traces is plotted as a function of σ_{analog} in Fig. 11-right. If the true analog mean value lies close to the threshold of rounding upwards or downwards, even a very small σ_{analog} will cause the rounded trace to flip between both integer values, leading to a standard deviation of exactly 1/2. For a μ that is closer to an

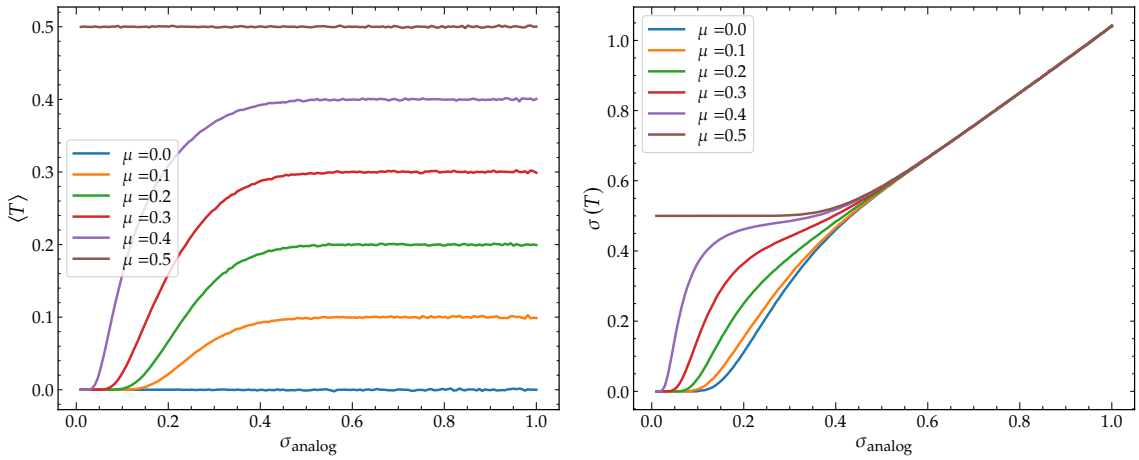


Figure 11: The analog signal of traces has to be converted to a digital signal, rounding traces to integers. Due to this process, the mean and the error of the mean of a baseline estimate change to the true mean, if the fluctuations of the analog trace are below a threshold of $\sigma_{\text{analog}} = 0.5$. To prevent the error becoming artificially small, it is chosen according to Eq. (8). *Left:* Average of rounded traces for varying analog fluctuations and mean values. Due to rounding of the traces, the mean of a trace is biased if σ_{analog} is small. *Right:* Standard deviation of the traces. This value can be compared to the error of the mean of the estimated baselines. Due to the rounding of the traces, the error changes depending on the value of the true mean μ of the trace if the fluctuations are smaller than $\sigma_{\text{analog}} = 0.5$.

integer value, small fluctuations cannot cause flipping between the two integers, the fluctuations are suppressed, and the standard deviation drops to zero. To prevent the error of the mean of the baseline estimates artificially decreasing to zero due to this discretization effect, an error of

$$\sigma_{b_{\text{front,end}}} = \frac{\max(0.5, \sigma_{\text{front,end}})}{\sqrt{n_{\text{front, end}}}}, \tag{8}$$

is used. Outliers that have either a large, positive ΔB or a large $\sigma_{\text{front,end}}$ can be investigated in more detail by analyzing the distributions from Fig. 10. Fig. 12 shows two example traces with a large σ_{front} and σ_{end}. The reasons behind the abnormal shapes of these traces might be either related to issues in the electronics or lightning conditions.

The baseline difference ΔB is used to determine if an undershoot or an anomalous overshoot is present in the trace. For this, the error of the baseline difference $\sigma_{\Delta B}$, which is dependent on the error of the front and end trace window is calculated as

$$\sigma^2_{\Delta B} = \sigma^2_{b_{\text{front}}} + \sigma^2_{b_{\text{end}}}. \tag{9}$$

Different interpolation methods between B_{front} and B_{end} can now be chosen depending on ΔB and $\sigma_{\Delta B}$.

3.3.1 Rejection of anomalous upward fluctuations $\Delta B \geq +5\sigma_{\Delta B}$

If the end baseline is estimated to be more than $5\sigma_{\Delta B}$ larger than the front baseline, the trace is rejected as an anomalous trace. Aside from random fluctuations of B_{front} and B_{end}, there is no reason the baseline should increase after a signal. A significant increase of the baseline indicates possible errors in the electronics and the resulting traces should not be considered for the estimation of a signal. In Fig. 13, two example traces with $\Delta B > +5\sigma_{\Delta B}$ are shown.

In both cases, the trace seems to have a long tail that decreases slowly over time but does not reach the original baseline.

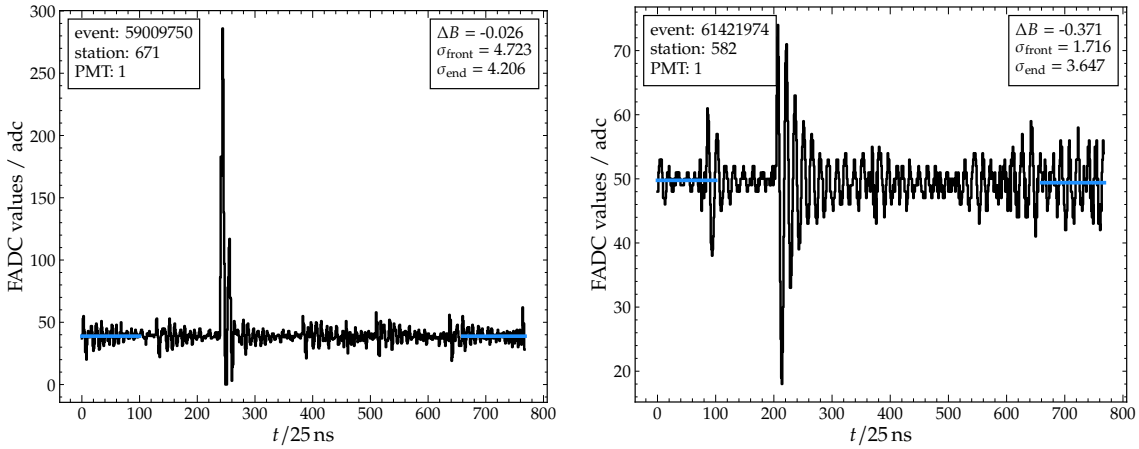


Figure 12: A large σ_{front} or σ_{end} can be caused by various reasons, as for example by lightning events or broken PMTs. *Left:* Noisy trace due to electronics. *Right:* Trace of a lightning event.

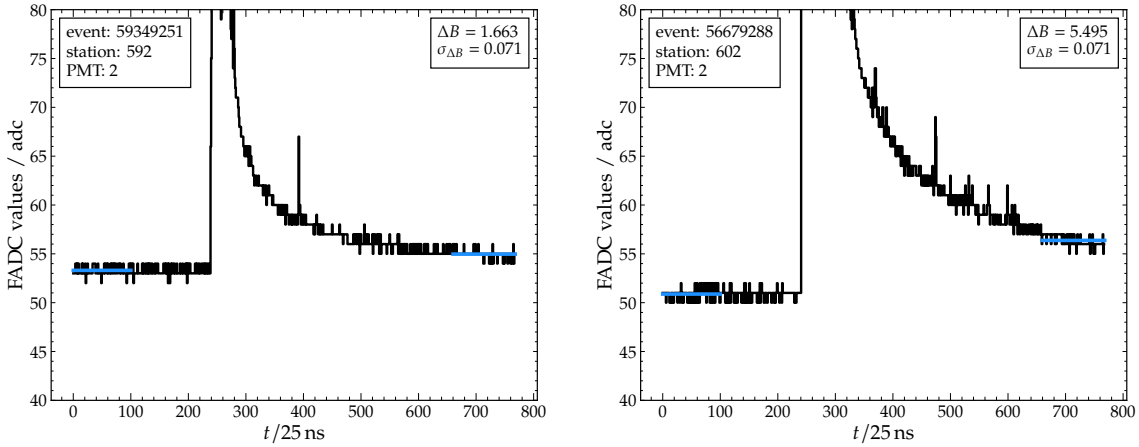


Figure 13: Anomalous traces that have an overshoot can be filtered out by applying a threshold of $5\sigma_{\Delta B}$ as maximum positive ΔB. *Left:* Station 592 with $B_{\text{end}} > B_{\text{front}}$. *Right:* Station 602 with $B_{\text{end}} > B_{\text{front}}$.

3.3.2 Constant approximation for small upward fluctuations $+5\sigma_{\Delta B} > \Delta B \geq 0$

An increase of the baseline up to $+5\sigma_{\Delta B}$ is assumed to be random fluctuations. A reevaluation of the baseline based on the full trace is performed. The truncation method used for the front and end baseline estimates is applied to the full range of the trace. The resulting estimate is chosen as a robust constant for the full baseline. Fig. 14-left shows an example trace where the end baseline estimate is larger than the front baseline estimate but still smaller than the allowed $5\sigma_{\text{Delta B}}$ upwards fluctuation.

3.3.3 Step-function approximation for small downward fluctuations $0 > \Delta B \geq -1\sigma_{\Delta B}$

Similar to the positive fluctuations the baseline at the end of the trace can be smaller than the front baseline. However, traces are expected to have an undershoot after a signal. Instead of calculating a robust constant, a step function is used to account for the undershoot. B_{front} is used as the baseline from the start of the trace to the bin with the maximum trace value. Then the baseline is set to B_{end} until the end of the trace. An example trace with a step-function as baseline interpolation is shown in Fig. 14-right.

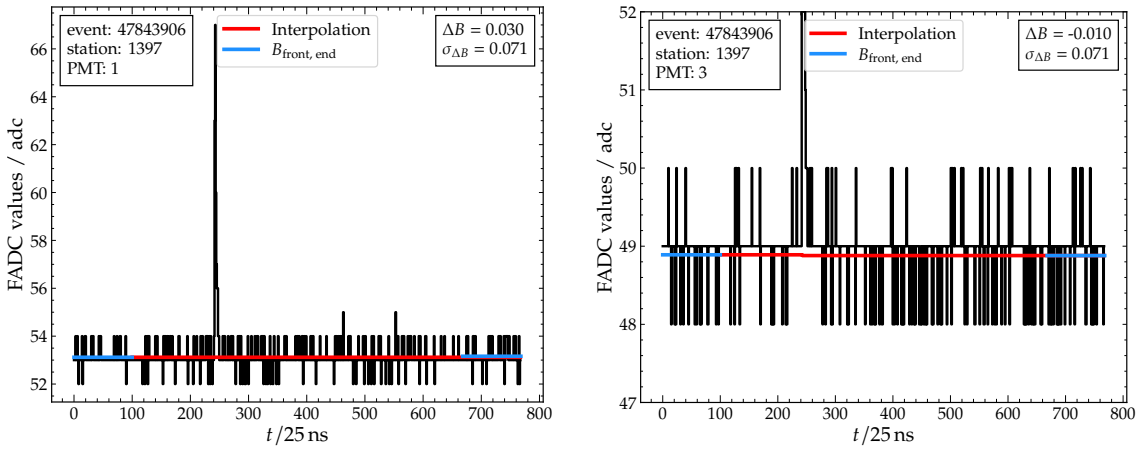


Figure 14: If the undershoot is small compared to the error of the baseline difference $\sigma_{\Delta B}$, two cases are distinguished. For upwards fluctuations up to $+5\sigma_{\Delta B}$, a robust constant is evaluated by applying the truncation procedure on the full trace (left). A step-function is used, if there are small downwards fluctuations down to $-1\sigma_{\Delta B}$ (right).

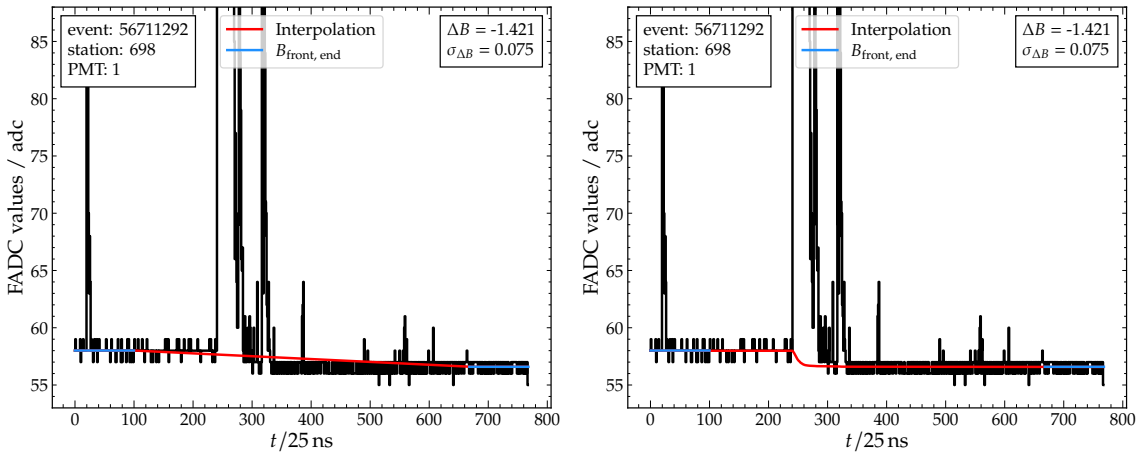


Figure 15: If the undershoot exceeds $-1\sigma_{\Delta B}$, a combination of two interpolation methods is used to find an appropriate baseline. In the first step, the baseline is interpolated linearly in time between the front and end baseline estimates (left). Using this first approximation, a total charge of the trace can be calculated and the baseline is reevaluated, using a linear interpolation in charge (right).

3.3.4 Charge-linear approximation for large undershoots $-1\sigma_{\Delta B} > \Delta B$

If B_{end} is smaller than $-1\sigma_{\Delta B}$ compared to B_{front}, a more refined procedure is used to model the baseline. The baseline is interpolated linearly in time between the front and end baseline estimate, as shown in Fig. 15-left. Afterwards, the baseline is linearly interpolated relative to the total charge of the trace in an iterative procedure. Fig. 15-right shows an example of the first iteration of the charge-linear interpolation. The interpolation is only executed, if the bin with the most ADC counts exceeds a threshold of 50 adc relative to the front baseline estimate. If the difference of the maximum of the trace and B_{front} is below 50 adc, the previously introduced step function is used to calculate the baseline. A more detailed explanation of this charge linear interpolation is explained in Section 3.4.

3.4 Charge-linear interpolation

The last step of the algorithm is to find an appropriate interpolation of the baseline between the front and end baseline estimates B_{front} and B_{end}. The baseline is set to be either a robust constant, in case of upwards fluctuations of the end baseline or a step function, for small downwards fluctuations. For significant undershoot an appropriate interpolation between B_{front} and B_{end} is needed. As a first step the baseline b_i is interpolated linearly in time between the end point b_{start} of B_{front} and the start point b_{end} of B_{end}

$$b_i = B_{\text{front}} - (i - b_{\text{front}}) \frac{\Delta B}{n_{\text{mid}}}. \tag{10}$$

i denotes the current bin, $n_{\text{mid}} = n - (b_{\text{front}} + b_{\text{end}})$ is the the length of the interpolated trace and n is the total trace length. By subtracting the baseline from the trace, an approximate total charge q_{tot} can be calculated by integrating the full trace T_i

$$q_{\text{tot}} = \sum_{i=b_{\text{front}}}^{b_{\text{end}}} T_i. \tag{11}$$

As previously shown in Fig. 6 the undershoot per charge is constant. After the linear interpolation between the baseline estimates, a new baseline estimate is calculated, which is linearly dependent on the charge

$$b_i = B_{\text{front}}(1 - \epsilon_i) + B_{\text{end}}\, \epsilon_i, \tag{12}$$

with ϵ defined as the fraction of the cumulative current charge

$$\epsilon_i = \frac{\sum_{j=b_{\text{front}}}^{i} T_j}{q_{\text{tot}}}. \tag{13}$$

The total charge of the trace is recalculated using the new baseline estimation and the procedure is repeated again. For the test data set, this procedure is repeated 30 times. If q_{tot} is evaluated negative during the procedure the iterations are stopped and the trace will not be used further. Fig. 16 shows the absolute difference between the calculated total charge $q_{\text{tot}}(i)$ after the i-th iteration and total charge of the previous iteration $q_{\text{tot}}(i-1)$. q_{tot} varies the most in the first two steps and the absolute difference reduces as more iterations are done.

3.5 Limitations

After a baseline has been determined a check on the baseline is performed. The mean negative amplitude $\langle I_{\text{neg}} \rangle$ of the baseline-subtracted trace is calculated as

$$\langle I_{\text{neg}} \rangle = \frac{q_{\text{neg}}}{n_{\text{mid}}}, \tag{14}$$

with q_{neg} as the sum of the negative charges of the trace

$$q_{\text{neg}} = \sum_{i=b_{\text{front}}}^{b_{\text{end}}} T_i \begin{cases} 0, & T_i \geq 0 \\ T_i, & T_i < 0 \end{cases}. \tag{15}$$

If the estimated baseline matches the true baseline, $\langle I_{\text{neg}} \rangle$ should not exceed the root-mean-square (RMS) value of the trace. A histogram of the mean negative amplitudes of 443 011 traces is shown in Fig. 17 with a threshold of 0.5. About 99.9% of these traces have a mean negative amplitude smaller than this threshold. For a trace without any signal the mean negative amplitude should be of the order of the product of the RMS of the trace and the trace length. The RMS of the UB traces is around $\sigma \approx 0.5$ as can be seen in Fig. 10-right. Four example traces with a mean

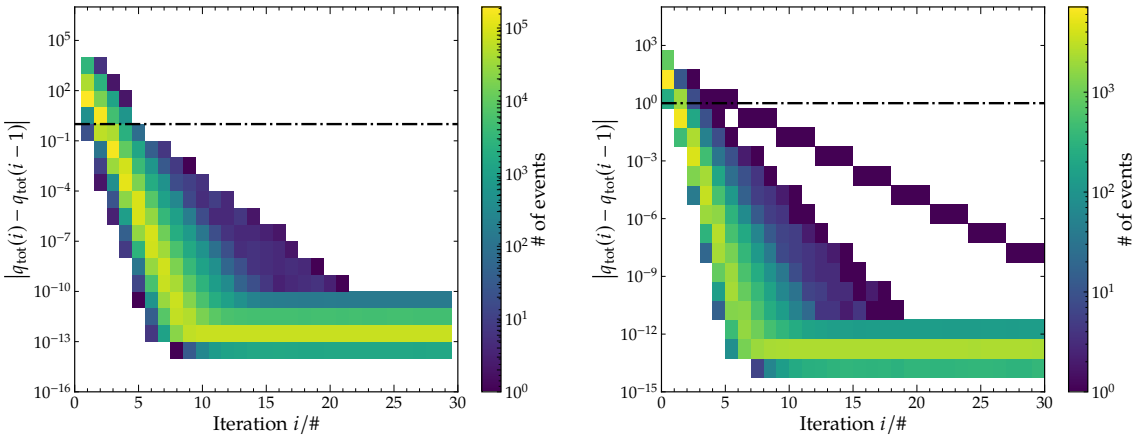


Figure 16: The absolute difference between the total charge q_{tot} of the i-th iteration and the previous iteration is decreasing with each step. After about 5 iterations, the absolute difference becomes smaller than 1 and the procedure can be stopped. *Left:* Total charge q_{tot} for each iteration of high-gain traces. *Right:* Total charge q_{tot} for each iteration of low-gain traces.

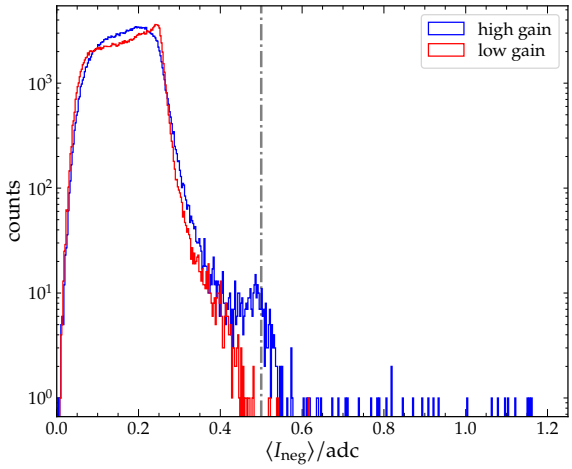


Figure 17: The mean negative amplitude $\langle I_{\mathrm{neg}} \rangle$ of the baseline-subtracted trace can be used for a fast check of the calculated baseline. If $\langle I_{\mathrm{neg}} \rangle$ exceeds a threshold of 0.5, the trace might have an anomalous shape.

negative amplitude smaller than this threshold are shown in Fig. 18. The first figure shows an anomalous trace with a visible undershoot recovery, which should not be present in UB traces. In the second figure a trace with an oscillating baseline is shown. An example of a malfunctioning PMT is depicted in the third figure. After the signal strong oscillations in the trace can be seen. The last trace of Fig. 18 shows anomalous drops of the trace. These traces do not correspond to the physical trace model for UB traces that was introduced in Section 3.1 and should not be used for further analysis.

4 Evaluation with Simulated Data

The results of the KG baseline algorithm are compared with the OG baseline algorithm that is currently used. In Offline, non-saturated traces T_i are simulated and extracted. These traces have no undershoot, baseline or noise and can be considered as pure Monte Carlo traces. From these traces, the total charge q_{tot}, as well as the true signal S_{mc} is calculated. Afterwards an artificial

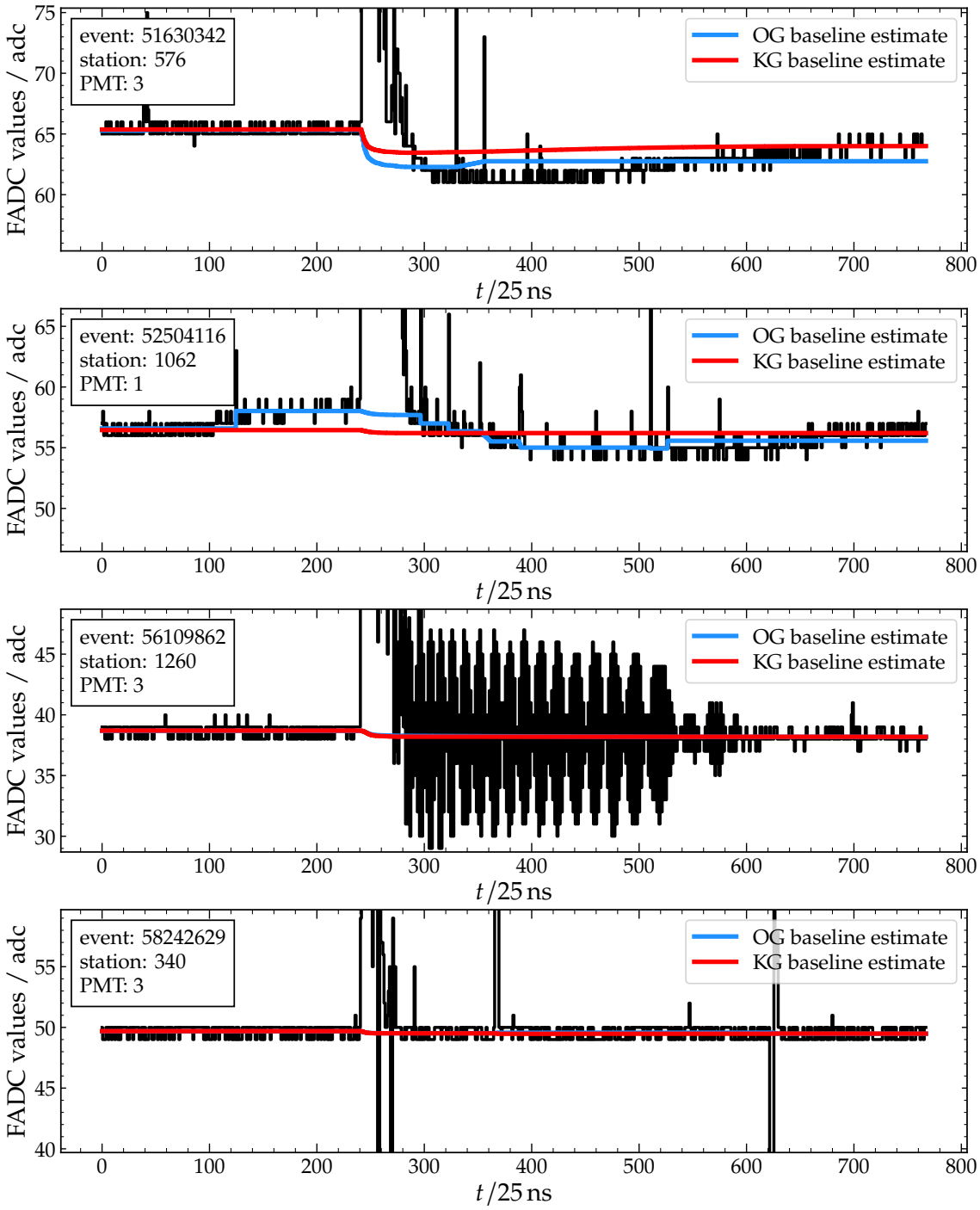


Figure 18: Examples of anomalous trace shapes. The mean negative amplitude $\langle I_{\text{neg}} \rangle$ of these traces exceeds the threshold of 0.5 and thus can be identified as anomalous. *From top to bottom:* Too-fast undershoot recovery. Oscillating baseline with long wave. Malfunctioning PMT with high-frequency oscillations. Baseline drops.???

undershoot is added to the trace T_i, to receive the new trace T_i'

$$T_i' = T_i - kq_i, \tag{16}$$

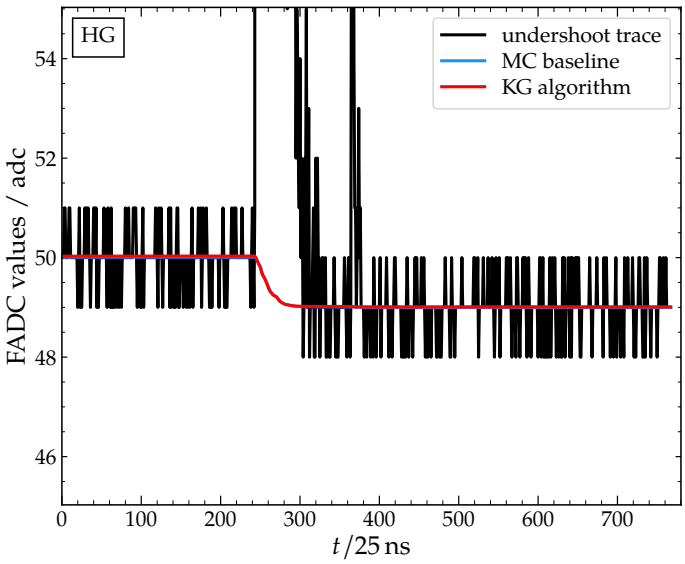


Figure 19: Simulated trace with artificial noise and undershoot. The estimated baseline from the KG algorithm (red) matches the simulated, true baseline (blue) very well.

with a constant factor k and

$$q_i = \sum_j = 0^i T_j. \tag{17}$$

The factor k can be determined from the previously described linearity between the undershoot ΔB and the total charge q_{tot}. From Fig. 6 this factor can be read off to be approximately $1.7{\times}10^4$ for HG traces and $-0.5{\times}10^{-4}$ for LG traces. A true baseline b_{mc} is added to the new trace t'_i and artificial random noise with a standard deviation of 0.5. In the last step, the values of the trace are rounded to integer values to resemble a trace of how it would be received from the electronic boards. An example simulated trace is shown in Fig. 19 together with the true baseline in blue and the estimated baseline from the KG baseline algorithm in red. The KG algorithm matches the true baseline in this example almost perfectly. This procedure is applied to all traces of the simulated library in Table 1. The signal S is determined using the OG baseline algorithm and the KG algorithm and compared to the true signal S_{mc}. A bias relative to S_{mc} as well as the uncertainty σ is then calculated. The results are shown in Fig. 20-left for the OG baseline algorithm and in Fig. 20-right for the KG algorithm. The HG data points include only traces that are below the saturation threshold of 1023 adc. For the LG data points all traces where the HG is saturated are used. It is thus possible to determine the region of transition from HG to LG, which ranges from around 30 VEM up to 200 VEM.

For both versions of the baseline algorithm, the relative bias of the HG is below 1%. The bias of the LG for the OG algorithm however ranges from a negative bias of around 0.5% up to 5%. While at large signals at around 700 VEM the bias is rather small, it has a wide range at the transition region from HG to LG. In the lower plots of Fig. 20 the signal uncertainty is shown next to the uncertainties from the baseline predictions. The uncertainty of the baseline algorithm decreases with increasing signal. While for both algorithms the uncertainty is below the signal uncertainty, the uncertainty of the OG algorithm for the LG signals at around 200 VEM is as large as the signal uncertainty. Compared with the KG baseline algorithm, the relative bias of the LG traces ranges from a maximum of around 1% at the transition region up to almost no bias for larger signals. The procedure thus can give an improved estimate of the true baseline, compared to the OG algorithm.

The thresholds for deciding which interpolation method is used are not equally distributed. This can be reasoned by the physical trace model, introduced earlier. A trace should not show a

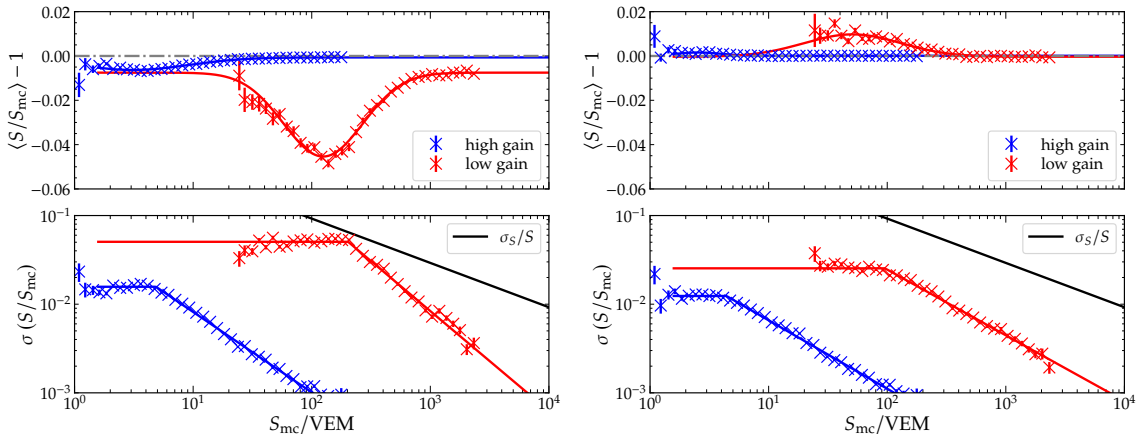


Figure 20: Comparison between the estimated signals, using the OG baseline algorithm (left) and the KG baseline algorithm (right). The upper plots show the bias relative to the true signal S_{mc}, fitted to a Gauss function. The lower plots show the uncertainty of the estimated signal, fitted to a rational function, compared to the signal uncertainty model used in $\overline{\text{Off}}\underline{\text{line}}$.

significant overshoot, thus a maximum of $+5\sigma_{\Delta B}$ is used to include as many traces as possible when calculating a robust constant. An undershoot however is expected and thus the step-function is used to describe the traces and the charge-linear interpolation is already used at a lower threshold of $-1\sigma_{\Delta B}$.

5 Impacts on Reconstruction

The estimated baselines from the OG algorithm can now be compared with the baselines estimated by the KG algorithm, using the data set of events between years 2019 and 2021 from Section 1. Fig. 21 shows the previously shown traces and $\overline{\text{Off}}\underline{\text{line}}$ baselines from Fig. 4 and the baselines of the KG algorithm. For the baseline at the beginning of the trace, both algorithms produce similar estimates. While in the $\overline{\text{Off}}\underline{\text{line}}$ baseline finder small humps in the baseline are present due to the identification of flat pieces, these humps vanish with the charge-linear interpolation, as in the first and last figure, or by the constant baseline estimate, as shown in the second figure of Fig. 21. In the third figure only a front and end baseline estimate are calculated. Since the end baseline is significantly larger than the front baseline, the trace is rejected as an anomalous trace.

Fig. 22 shows the comparison of the ratio between LG and HG for both algorithms for 3 years of SD data, as used in Fig. 2. At the transition point from HG to LG, the bias of the KG algorithm improves to -1.5% at the transition from LG to HG, compared to the bias of -6.5% with the OG algorithm. However, this negative bias is in contrast to the expected positive bias of about 1% in the transition region from the simulations.

In Fig. 23 the ratios between LG and HG, binned in zenith angle and energy, are compared between both algorithms. The ratio calculated with the current baseline algorithm is dependent on energy and zenith angle. Using the KG baseline algorithm, this dependence almost vanishes. This indicates, that the remaining bias might originate from a bias of the (D/A)-ratio, which should be investigated further.

A comparison between the reconstructed S_{1000} with both algorithms is shown in Fig. 24 as a function of S^{OG}_{1000}. The KG algorithm gives larger results for S_{1000} than the OG algorithm, that increase with larger S^{OG}_{1000} up to 5% on average. Although S_{1000} is used for the energy estimation, a KG energy calibration has to be done first, before conclusions about the shift of the energy can be done.

The absolute difference between the start and stop times is shown in Fig. 25. The start times

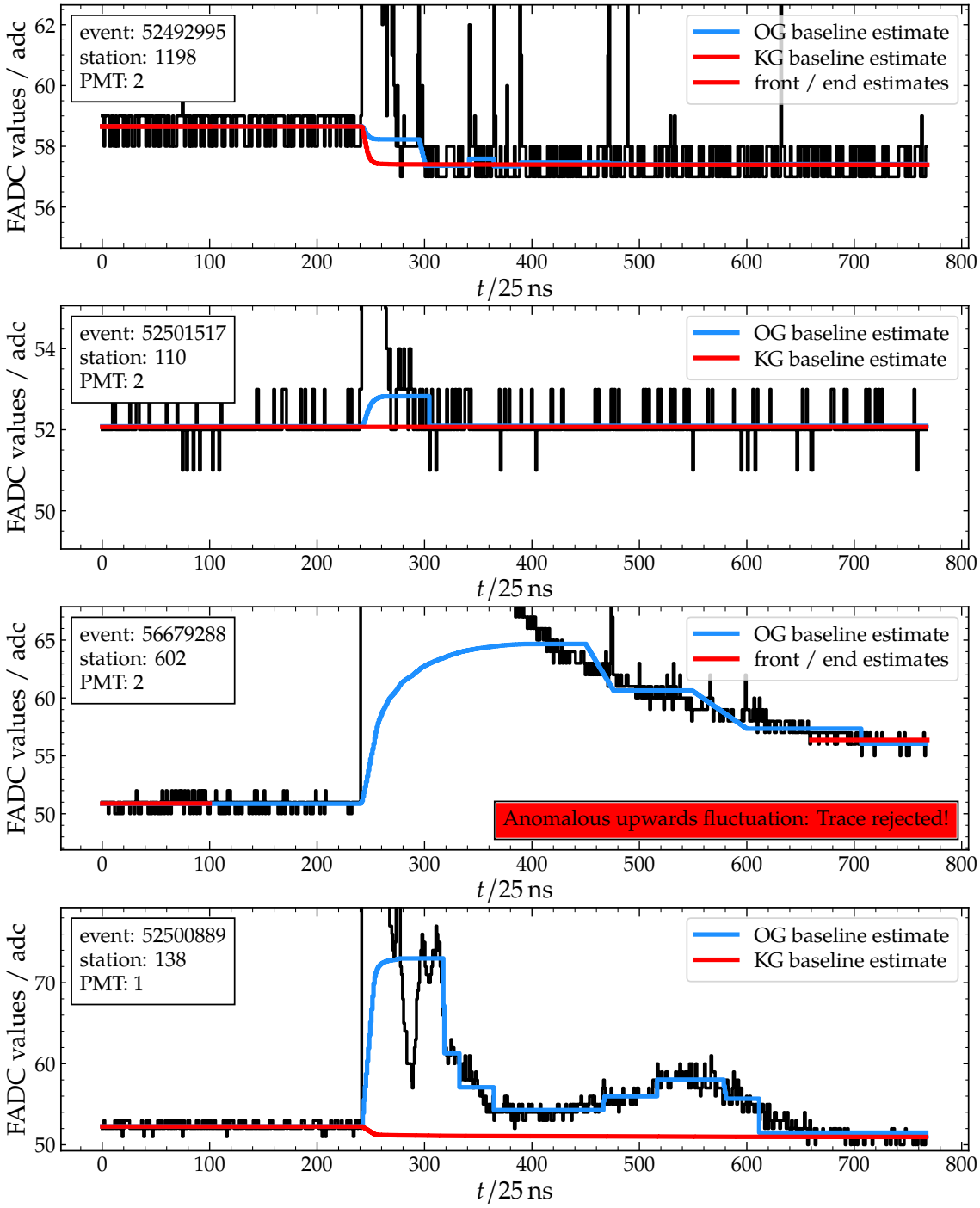


Figure 21: Example traces with the baseline estimate of the `BaselineFinderOG` module of Offline and the KG `BaselineFinderKG` module. At the beginning of the trace the baseline estimates are similar, but start to deviate as soon as the signal starts. *From top to bottom:* Charge-linear interpolation. Estimation of a robust, constant baseline. Rejection due to anomalous upwards fluctuation. Charge-linear interpolation.

do not change for most of the events except for a few outliers, that are more frequent at smaller signals. The stop times however have a broader distribution of outliers compared to the start

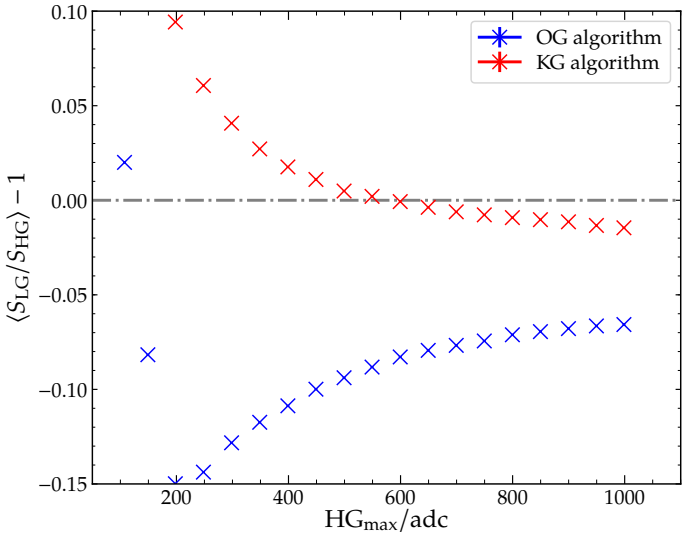


Figure 22: Compared to the OG algorithm implemented in Offline (blue), the KG algorithm (red) improves the systematic error between the LG and HG signals. At the transition point from HG to LG at $ADC_{max} = 1023$, a negative bias of approximately 1.5% remains.

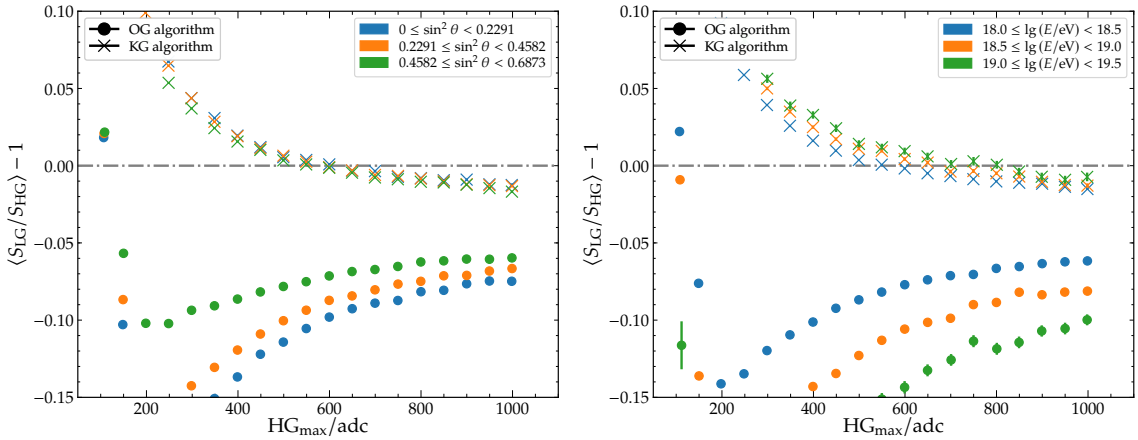


Figure 23: Comparison of the low-gain-to-high-gain ratio with different binning for the OG algorithm (circles) and the KG algorithm (crosses). While for the OG algorithm, there is a strong dependence on energy and zenith angles visible, this dependency almost vanishes when using the KG algorithm.

times. At larger signals the average of the absolute difference increases up to 50 bins. This can be attributed to the inclusion of small, late signal contributions that were previously filtered out by the OG baseline algorithm.

6 Conclusions

We have shown that a large part of the systematic bias in the ratio of LG and HG signals origins from the estimation of the baseline of the time traces. The OG baseline algorithm tries to find flat signal pieces and connects them as an estimate. Even with a correction procedure the estimated signal of the LG is deviating from the true signal. We introduced a physically motivated KG baseline algorithm, that accounts for undershoot after large signals and is robust against accidental

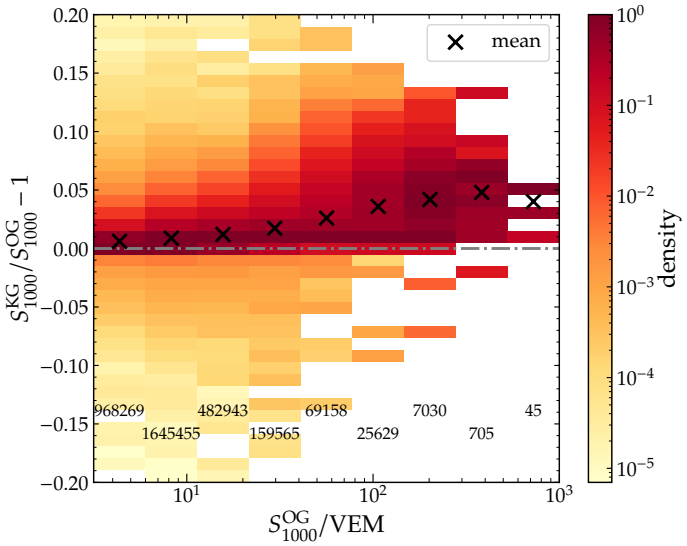


Figure 24: Comparison between the reconstructed S_{1000}, using the OG and the KG algorithm. The color of each column is normalized to the bin with the most entries. The total number of entries of each column is given at the bottom of the plot. With increasing values of S_{1000}, the KG algorithm leads to a higher reconstruction of S_{1000} than the OG algorithm.

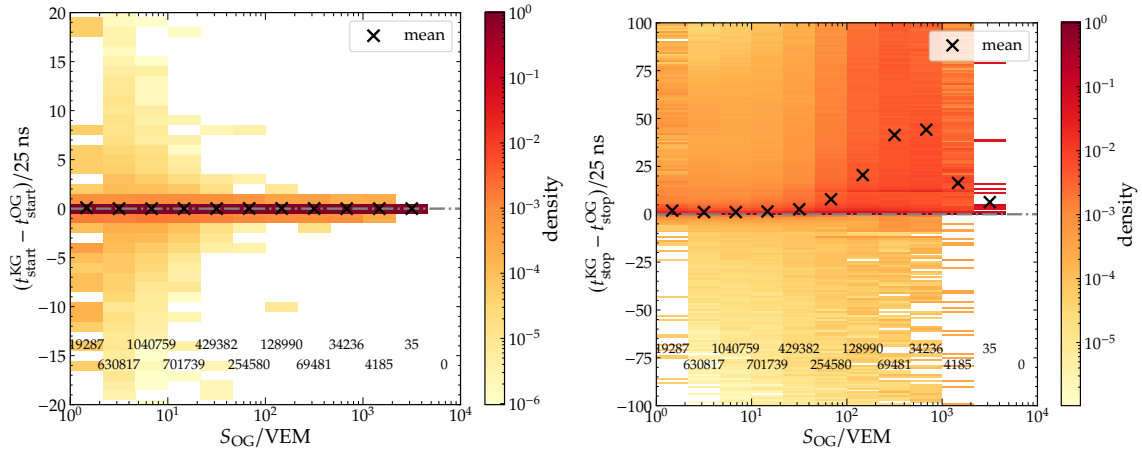


Figure 25: At large station signals, the start times do not significantly change between both algorithms. The color of each column is normalized to the bin with the most entries. The total number of entries of each column is given at the bottom of the plot. The stop times increase at larger station signals by up to 50 bins on average.

signal contributions outside of the signal region. By determining two robust estimates of the baseline at the beginning and the end of the trace, it can be checked for anomalies or bad traces, that should not be included in reconstructions. Depending on the relative difference of both estimates, the significance of the undershoot can be determined and an interpolation method for the baseline can be decided. The KG algorithm has been shown to work extremely well on simulated traces and also shows a significant improvement in the ratio between LG and HG signals. This algorithm is now included in a new `BaselineFinderKG` module of Offline.

References

[1] Pierre Auger collaboration, *The front-end electronics for the Pierre Auger Observatory surface array*, *IEEE Trans. Nucl. Sci.* **51** (2004) 413.

[2] Pierre Auger collaboration, *New Electronics for the Surface Detectors of the Pierre Auger Observatory*, *PoS* **ICRC2019** (2021) 370.

[3] P. Sánchez Lucas, *The ⟨Δ⟩ Method: An estimator for the mass composition of ultra-high-energy cosmic rays*, Ph.D. thesis, U. Granada (main), 2016.

[4] Q. Luce and I. Lhenry-Yvon, "Saturation and estimation of the signal on high and low-gain channel." 2017.

[5] Q. Luce and I. Lhenry-Yvon, "A modified baseline estimation and signal selection in fadc traces of sd detectors within cdas." 2016.

[6] B. Genolini, T. Nguyen Trung and J. Pouthas, "Base line stability of the surface detector pmt base." 2003.

[7] FUNK Experiment collaboration, *Limits from the Funk Experiment on the Mixing Strength of Hidden-Photon Dark Matter in the Visible and Near-Ultraviolet Wavelength Range*, *Phys. Rev. D* **102** (2020) 042001 [2003.13144].

[8] Z. Szadkowski and K.-H. Kampert, "Analysis of the sd-pld firmware and implications to physics data." 2006.

A Quick overview (tl;dr)

I **Robust baseline estimation**

1 Calculate the mode m of a given trace window.

2 Calculate mean and the standard deviation σ.

3 Relative to the mode use 2σ as truncation criteria. Additionally round it up/down to include bins.

- If bins are excluded, repeat step 2 and 3, until no more bins are excluded.
- If number of remaining bins $n < 40$ reject the trace

4 Calculate mean b of truncated trace window, as well as σ.

II **Decision on interpolation methods**

Calculate undershoot $\Delta B = b_{\text{end}} - b_{\text{front}}$ and decide on interpolation, if

- $\Delta B \geq 5\sigma_{\Delta B}$: reject the trace.
- $5\sigma_{\Delta B} > \Delta B \geq 0$: make a robust baseline estimate (Procedure 1) on full trace.
- $0 > \Delta B \geq 1\sigma_{\Delta B}$: use a step function with the transition $B_{\text{front}} \rightarrow B_{\text{end}}$ at the trace maximum.
- $1\sigma_{\Delta B} > \Delta B$: use charge-linear interpolation between the front and end estimates, if the number of max bins is larger than 50, otherwise step function.

III **Charge-linear interpolation**

1 Linearly interpolate the baseline (in time) between the end of the front baseline estimate b_{front} and the start of the end baseline estimate b_{end}.

2 Subtract the baseline from the trace to obtain a baseline-subtracted trace T_i.

3 Integrate the trace T_i between b_{front} and b_{end} to obtain an estimate of the total charge q_{tot}.

4 Calculate baseline $B_{\text{new}} = B_{\text{front}}(1 - w) + B_{\text{end}}\, w$ and $w = q_{\text{part}}/q_{\text{tot}}$ were q_{part} is the partial sum and q_{tot} the total sum of the trace T.

5 Repeat five times starting from step 2.

6 If q_{tot} becomes negative the baseline algorithm fails, trace might be anomalous $\rightarrow$ reject