PROCEEDINGS OF SCIENCE

Using pulse-shape information for reconstructing cosmic-ray air showers and validating antenna responses with LOFAR and SKA

Nikolaos Karastathis,[f,*] **S. Buitink,**[a,b] **A. Corstanje,**[a,b] **M. Desmet,**[a] **H. Falcke,**[b,d,e] **B. M. Hare,**[d] **J. R. Hörandel,**[a,b,e] **T. Huege,**[a,f] **V. B. Jhansi,**[c] **G. K. Krampah,**[a] **P. Mitra,**[a] **K. Mulrey,**[b,e] **B. Neijzen,**[d] **A. Nelles,**[g,h] **H. Pandya,**[a] **O. Scholten,**[i,j] **K. Terveer,**[h] **S. Thoudam,**[c] **G. Trinh**[k] **and S. ter Veen**[d]

[a] *Vrije Universiteit Brussel, Astrophysical Institute, Pleinlaan 2, 1050 Brussels, Belgium*
[b] *Department of Astrophysics/IMAPP, Radboud University Nijmegen, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands*
[c] *Department of Physics, Khalifa University, P.O. Box 127788, Abu Dhabi, United Arab Emirates*
[d] *Netherlands Institute for Radio Astronomy (ASTRON), Postbus 2, 7990 AA Dwingeloo, The Netherlands*
[e] *Nikhef, Science Park Amsterdam, 1098 XG Amsterdam, The Netherlands*
[f] *Institut für Astroteilchenphysik, Karlsruhe Institute of Technology (KIT), P.O. Box 3640, 76021 Karlsruhe, Germany*
[g] *Deutsches Elektronen-Synchrotron DESY, Platanenallee 6, 15738 Zeuthen, Germany*
[h] *ECAP, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany*
[i] *Interuniversity Institute for High-Energy, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*
[j] *University of Groningen, Kapteyn Astronomical Institute, Groningen, 9747 AD, Netherlands*
[k] *Department of Physics, School of Education, Can Tho University Campus II, 3/2 Street, Ninh Kieu District, Can Tho City, Vietnam*

E-mail: nikolaos.karastathis@kit.edu

The Low Frequency Array (LOFAR) is capable of measuring extensive air showers through their radio emission in the frequency range of 30–80 MHz, while the Square Kilometer Array (SKA) will be able to expand this range to 50–350 MHz. A very important characteristic of cosmic rays is the mass of the primary particle, which is associated with the atmospheric depth of the shower maximum ($X_{\max}$). The standard $X_{\max}$ reconstruction procedure with LOFAR involves the use of a library of CORSIKA/CoREAS simulations for a specific measured event and uses the energy deposited to the ground in terms of radio fluence. In this study, to extract information about shower development, not only the energy fluence is considered but the possibility of using information from the pulse shape is investigated in both frequency ranges (30–80 MHz and 50–350 MHz). The study of the pulse shape through the pulse agreement of measured data and simulations also provides a way to diagnose the proper functioning of individual LOFAR dipoles.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

*Speaker

PoS(ICRC2023)487

1. Introduction

The radio detection technique of extensive air showers has undergone an impressive renaissance over the past 20 years [1] and has become a technique competitive with particle and fluorescence detection. The Low Frequency Array (LOFAR) [2] is a digital radio telescope spread all over 24 European countries, which is capable of measuring the radio emission of extensive air showers (EAS) in the frequency range of 30–80 MHz. Its main feature is that it consists of a dense core in the Netherlands which allows for high quality data collection. The Square Kilometer Array (SKA) [3] will be an even denser antenna array, which aims for an even better resolution and will be able to measure EAS in the broader frequency range of 50–350 MHz. In order to be able to characterize the primary cosmic ray particle we need to know its mass. This can be associated with the atmospheric depth of the shower maximum ($X_{\max}$). To analyze experimental data and reconstruct the main properties of the shower and of the primary particle, the community relies heavily on software that simulates the radio emission [4], [5], [6], [7]. For LOFAR, there has been developed a reconstruction technique for $X_{\max}$ which makes use of a library of CORSIKA/CoREAS simulations and has achieved a resolution of 17 g/cm^2 [8]. An updated study has managed to lower the resolution to the 7 to 9 g/cm^2 range [9] for a single event.

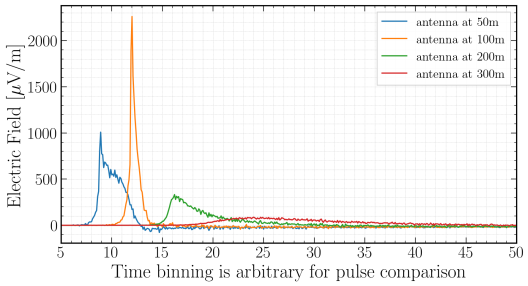

Figure 1: The signal pulse shape varies for different distances from the shower core.

This $X_{\max}$ reconstruction technique uses the energy deposited on the ground, or fluence, from the radio emission. It tries to match the total fluence of each individual antenna between measured data and libraries of CORSIKA/CoREAS simulations. In order to calculate the fluence though, only the signal strength of each pulse is considered, throwing away potentially significant information that could improve the resolution. We know that there is information also in the pulse shape we detect that is associated to $X_{\max}$ and distance from the shower core (Figure 1). A pulse can be wider or narrower (and the frequency spectra steeper or flatter respectively) depending on $X_{\max}$ and the distance of the antenna from the shower core. Close to the shower core the pulse is wider and as one moves closer to the Cherenkov ring the pulse reaches its maximum amplitude and gets narrower. In the same fashion, the frequency spectra close to the shower core fall steeply and close to the Cherenkov ring become flat. Moving away from the Cherenkov ring the signal starts to get weaker, the pulse becomes wider and the frequency spectra start to fall steeply again. These observations are the main motivation of this study. Our goal is to associate the pulse shape with $X_{\max}$ and devise a thorough reconstruction procedure.

The rest of the text is structured as follows: in section 2 the method we associate $X_{\max}$ with the pulse shape is presented along with an example study based on simulations. In section 3 the same method is expanded through the inclusion of simulated noise and the procedure we followed to add it to our simulations is explained. Finally, in section 4 we perform a statistical study using $\sim$ 5000 simulations in an attempt to benchmark the performance of our method with and without the presence of noise.

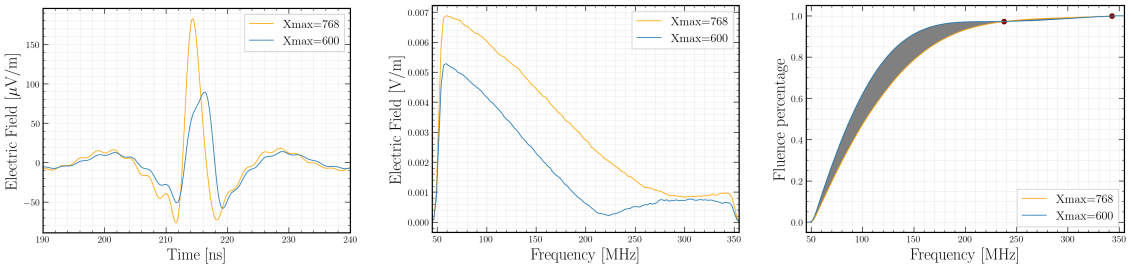

Figure 2: Two pulses at the same antenna (50 m from the shower core) with different $X_{\max}$ values in the time domain (left). The corresponding frequency spectra (middle) and the *fluence percentage* (right).

2. Presenting the method to reconstruct $X_{\max}$

The goal of this method is to associate $X_{\max}$ with the pulse shape and eventually come up with an $X_{\max}$ reconstruction scheme. For this we used existing CORSIKA/CoREAS simulations that have been already used for $X_{\max}$ reconstruction of detected events from the LOFAR radio telescope. We pick $\sim$ 30 simulations of one random event whose main characteristic is that the setup is identical, except for the random number seed. In more detail, these simulations have the same primary cosmic ray particle, same energy, same zenith and azimuth angle, same particle energy cuts and the difference in the seed affects solely the air shower development in the atmosphere, hence $X_{\max}$. For the antenna array all simulations use a star-shaped pattern of 160 antennas (sampling period is set to 0.1 ns) located at the ground, in 20 concentric rings spaced equally from 25–500 m from the shower axis with 8 antennas distributed azimuthally in each ring. With this setup we manage to obtain signal pulses from 160 antennas for showers whose only difference is $X_{\max}$. For instance, for a random antenna of the star-shaped grid we have $\sim$ 30 pulses for showers which differ only in $X_{\max}$. For every antenna location we want to compare all these $\sim$ 30 pulses among each other in order to find a quantifiable metric for $X_{\max}$ information.

We first investigated the pulse shape information in the time domain using the correlation factor of the pulses, but quickly realized that the values were very close to 1 even in simulations which means that such a method would easily break in real world scenarios in the presence of noise. We next moved to investigate the pulse shape in the frequency domain in the range of 50–350 MHz as this is the range that SKA will operate. We also did a simulation study in the range of 30–80 MHz that LOFAR operates, although we did not extend it to study measured data as we know that the LOFAR antennas are highly resonant in a very small window around 60 MHz. The pulse shape information in the frequency domain can be encapsulated in the slope of the spectra, so one could argue that taking the ratio of 2 spectra with different $X_{\max}$ might be enough. Such an idea, although interesting would seriously be affected by noise so we moved further to a more robust metric. Starting from the frequency spectra we take the square of the amplitude spectrum of each frequency bin and then calculate its cumulative sum. Finally, we normalize all the bins of the cumulative sum with respect to the total fluence of the antenna, in order to encapsulate the rate in which energy is deposited to the antenna independent of the amplitude. We call this *fluence percentage* (f), as at the start of the frequency band in which we filter our pulses, 0% of the total fluence is deposited in the antenna, while at the end of the frequency band 100% of the total fluence has been deposited

in the antenna. For a random bin j, the *fluence percentage* can be summarized with the following formula:

$$f_j = \frac{\sum_{i=1}^{j} A_i^2(\nu)}{\sum_{i=1}^{N} A_i^2(\nu)} \tag{1}$$

With A we denote the amplitude spectra and with ν the frequency. It is worth noticing that the denominator of eq.(1) multiplied by the constant factors $\epsilon_0 c \Delta t$, where ϵ_0 is the vacuum permittivity, c is the speed of light in vacuum and Δt is the finite time binning is the *total fluence* with respect to which we normalize every bin j. These constant factors appear also in the numerator of eq.(1) so they eventually cancel out.

This can all be illustrated through an example in the 50–350 MHz range, that is shown in Figure 2. Starting from left to right, we plot the pulses of the dominant polarization from 2 simulations in the time domain for an antenna at 50 m from the shower core. The orange pulse is the one detected from that antenna from an EAS whose $X_{\max}$ is $\sim 768\,\mathrm{g/cm^2}$, while the blue pulse is the one detected from the same antenna from an EAS that differs only in $X_{\max}$ with the previous one, which is $\sim 600\,\mathrm{g/cm^2}$. We can see that that the 2 pulses differ in amplitude but also in pulse shape with the pulse with the smaller $X_{\max}$ being wider. Moving to the frequency spectra in the middle of Figure 2, we observe a difference in amplitude but also a difference in the slope with the smaller $X_{\max}$ spectra falling steeper and producing a "dip" at ~ 220 MHz. Finally, in the right of Figure 2 we plot the *fluence percentage* where we can observe that the rate with which energy is deposited on the same antenna is different for different $X_{\max}$. We quantify this difference in one number. This is the "difference in area" between the 2 curves. The hypothesis is that the smaller the difference in this area the closer the 2 pulses are in terms of $X_{\max}$, and if the difference is 0, then they share the same value for $X_{\max}$.

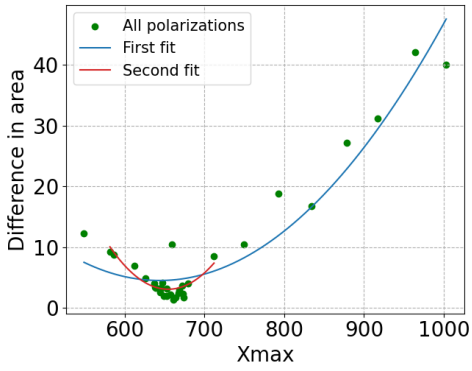

Figure 3: $X_{\max}$ reconstruction scheme using the parabola fitting procedure.

We test this hypothesis in the process explained below. We start from a library of $\sim$ 30 simulations that differ only in $X_{\max}$ and we pick one and treat it as *reference*, i.e. we use the procedure explained above to compare this specific simulation to the rest of the simulations. The goal is to approximate this reference $X_{\max}$ by calculating the "difference in area" for all antenna positions and all polarizations of the reference $X_{\max}$ versus the $X_{\max}$ of the rest of the simulations in the library. The antenna positions we considered were the ones inside the radius of 250 m of the star-shaped pattern antenna array because in larger radii the signal is usually very weak. For every $X_{\max}$ simulation, the "difference in area" for all antenna positions for each polarization is added together in an attempt to express the total "difference in area" in one number. To every $X_{\max}$ now corresponds a number, and the smaller this number is, the closer this $X_{\max}$ is to the reference $X_{\max}$ according to our hypothesis.

This behavior can be seen in Figure 3 were the $X_{\max}$ of the reference simulation is $\sim 661\,\mathrm{g/cm^2}$. At this stage, we perform a parabola fit in order to find its minimum as an indication on where

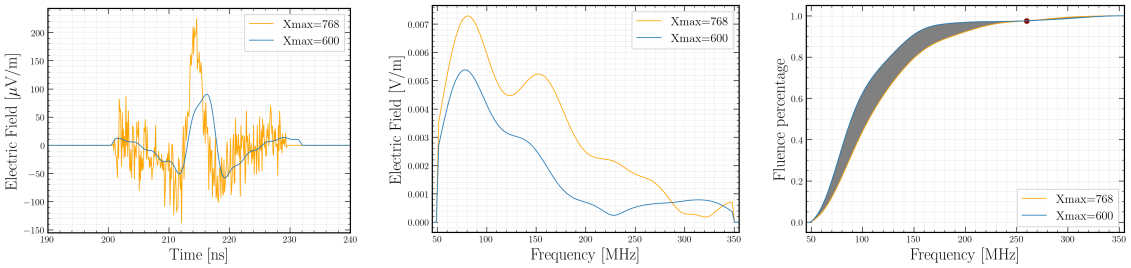

Figure 4: Artificial noise of the order of 10% of the amplitude added to the pulses of 2. The time window used is 30 ns.

the reference $X_{\max}$ might be. Then we create a "grammage window" of $\pm 75\,\mathrm{g/cm^2}$ around this minimum and use the points in this envelope to do a second parabola fit. The minimum of the second fit, we consider the reconstructed $X_{\max}$ ($X_{\max/\mathrm{reco}}$) and in this case is $\sim 656\,\mathrm{g/cm^2}$, differing $5\,\mathrm{g/cm^2}$ from the reference $X_{\max}$ we were looking for. It is worth noting, that this procedure when performed for the reference $X_{\max}$ compared to itself will give back a "difference in area" equal to zero, which is why we exclude this point in order to make the parabola fitting more fair.

This parabola fitting scheme can have great impact on the performance of our method. One could try smaller or bigger "grammage windows" which affect the final result, or try an even more sophisticated parabola fitting procedure like the one being used for standard $X_{\max}$ reconstruction in LOFAR [8] that fits data points on the lower part of an envelope. Our plan is to improve on our fitting procedure as we expect this to give significantly better results especially in the presence of noise.

3. Adding generated noise to the simulations

We continue our study using the same simulation library where we add generated noise to the reference simulation for the 50–350 MHz range. Since we do not have measured noise for the SKA site, we add the noise in the scheme described below. We find the antenna in the reference simulation with the strongest signal in its dominant polarization and use a fraction of its amplitude (5%, 10% and 20%) to set the width of the Gaussian we use to generate noise. For every antenna separately, we generate random Gaussian noise in the time domain and the width of the Gaussian is a fraction of the amplitude mentioned above. We then add bin by bin the generated noise on top of the simulated antenna pulse. This way every antenna has a different noise signature, although the noise amplitude is the same for all antennas. Our goal is to minimize noise in the frequency spectra. To achieve that, we apply a *Hanning* time window of 30 ns around the pulse's peak and the rest of the trace is set to 0 values (left plot in 4). The advantage of the *Hanning* window is that it smoothes the edges of the pulse between the zero values and the non zero values, which reduces the noise in the frequency domain. Then we transform the pulse to the frequency domain and we apply a simple band pass filter in the 50–350 MHz range.

The steps described above are shown in Figure 4, where we have used the same pulses as in Figure 2 in order to exhibit how noise alters the pulses. The Gaussian we used for the noise has a width that is 10% of the amplitude in the strongest time trace of the simulation. Noise is added

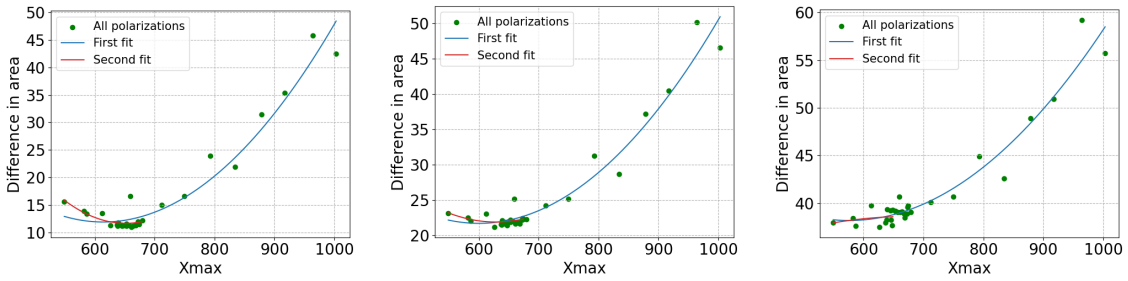

Figure 6: The X_{max} reconstruction scheme for different levels of added artificial noise. From left to right the noise is 5%, 10% and 20% of the amplitude.

only in the pulse whose $X_{\max}$ is $\sim 768\,\mathrm{g/cm^2}$, denoted with orange color in Figure 4. From left to right, the two pulses are plotted in the time domain within the 30 ns window, while values outside of this window are set to 0. The noise is unfiltered at this level. In the middle, the noise is now filtered and we see its effect in the frequency spectra denoted with the orange color. The frequency spectra in blue are identical to the ones in Figure 2 (middle), except for the fact that they are now restricted to a smaller time window of 30 ns. The importance of the choice of the time window is worth pointing out. We managed to have a small amount of noise in the frequency spectra denoted with orange, because we chose a narrow enough window. This makes the *fluence percentage* on the right feasible, by getting rid as much noise as possible and at the same time, keeping as much of the signal we can. As an example, we show how the frequency spectra would look like for a larger 100 ns window in Figure 5, which in turn introduces a larger difference in area in the *fluence percentage* plot.

In addition, we inspect the effect of different noise levels in the parabola fitting scheme. In Figure 6, from left to right we see how the fitting scheme behaves for noise of the 5%, 10% and 20% level of the amplitude as defined above respectively. For 5% $X_{\mathrm{max/reco}}$ is $\sim 651\,\mathrm{g/cm^2}$, so $10\,\mathrm{g/cm^2}$ below compared to the reference $X_{\max}$. For 10% of the amplitude, the resolution gets worse with $X_{\mathrm{max/reco}}$ being $\sim 630\,\mathrm{g/cm^2}$, which is $31\,\mathrm{g/cm^2}$ off from the reference $X_{\max}$. Finally, for 20% of the amplitude the parabola fitting scheme starts to break down providing unreliable results for these specific data points.

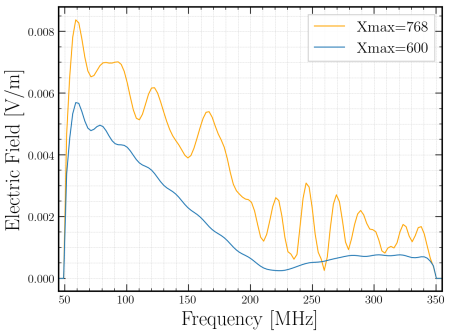

Figure 5: A wider 100 ns time window applied to the frequency spectra from Figure 4 introduces more noise to the spectra.

4. Benchmarking the method

As a final step in this study we try to benchmark this method using a large dataset of simulations. We use 146 simulated events that have been used for $X_{\max}$ reconstruction in the LOFAR analysis. All these events have primary particle energies $> 100\,\mathrm{PeV}$ and differ in zenith angles, which we have used to distinguish the events in the benchmark plots shown below. Each event consists of a number of simulations in the order of tens whose difference is $X_{\max}$, totalling to ~ 5000 simulations of iron

and proton air showers. For each event we sort in an array the simulations with respect to $X_{\max}$ in an ascending order. We pick as a reference shower the fifth shower in that array in order to have enough simulations with a smaller $X_{\max}$ than the reference one, and we perform our method comparing with all the showers in the array. For every combination of shower comparison we calculate and store the following quantity: $X_{\max/\text{reco}}-$ reference $X_{\max}$. We repeat this procedure until we pick as a reference the fifth shower from the end of this array. Once we have compared all the possible combinations of showers in that array, we define as bias the mean value of all the $X_{\max/\text{reco}}-$ reference $X_{\max}$ values we have calculated for this event. As resolution we define the standard deviation of the bias. Next, we repeat this procedure for the all the events in our library. Finally, we group together the events with respect to their zenith angle in 10^o intervals and we calculate the mean bias and mean resolution in these intervals.

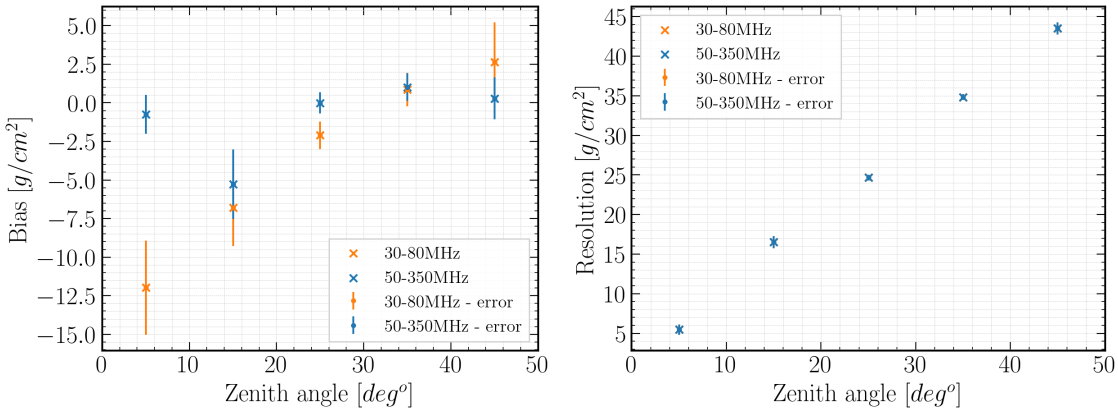

Figure 7: Benchmarking pure simulations in the 30–80 MHz and 50–350 MHz bands. On the left the bias is plotted and on the right the resolution.

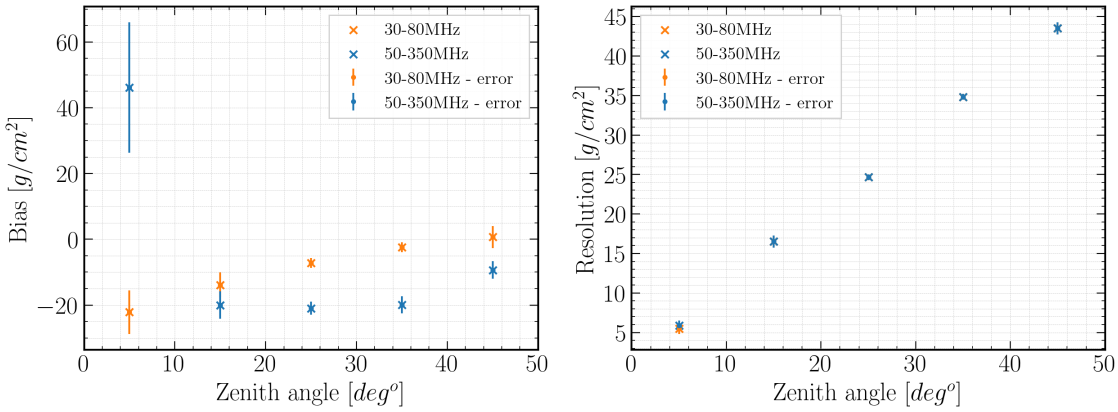

Figure 8: Benchmarking simulations with generated noise of 5% of the amplitude in the 30–80 MHz and 50–350 MHz bands. On the left the bias is plotted and on the right the resolution.

We performed this statistical study in both the 30–80 MHz and 50–350 MHz frequency band in 2 scenarios. The first one was pure simulations without noise and the second one is with noise which corresponds to 5% of the amplitude as discussed earlier. In Figure 7 we see the benchmark

without noise in both frequency bands. In the 30–80 MHz band the total mean value of the bias of all events corresponds to $-1.58\,\mathrm{g/cm^2}$ and the total resolution to $11.04\,\mathrm{g/cm^2}$, while for the 50–350 MHz band the total mean value of the bias is $-0.22\,\mathrm{g/cm^2}$ and the total resolution to $9.52\,\mathrm{g/cm^2}$. In Figure 8 we see the same benchmark in the presence of noise for both frequency bands. In the 30–80 MHz band the total mean value of the bias of all events corresponds now to $-6.12\,\mathrm{g/cm^2}$ and the total resolution to $19.66\,\mathrm{g/cm^2}$, while for the 50–350 MHz band the total mean value of the bias is $-12.11\,\mathrm{g/cm^2}$ and the total resolution to $19.14\,\mathrm{g/cm^2}$. We also used a random subset of $\sim$ 500 simulations where we studied the effect of noise in the total mean bias and total resolution. Along with a noise level of 5% of the amplitude we tried 10% and 20%. In this subset in the 30–80 MHz band, for 5% of the amplitude the total mean value of the bias of all events is $-0.77\,\mathrm{g/cm^2}$ and the total resolution is $-17.71\,\mathrm{g/cm^2}$. For 10% the same values are $-9.44\,\mathrm{g/cm^2}$ and $25.69\,\mathrm{g/cm^2}$, and for 20% are $-24.89\,\mathrm{g/cm^2}$ and $34.84\,\mathrm{g/cm^2}$ respectively. For 5% of the amplitude in the 50–350 MHz band, the total mean value of the bias is $-15.72\,\mathrm{g/cm^2}$ and the total resolution to $15.52\,\mathrm{g/cm^2}$. For 10% the same values are $-32.72\,\mathrm{g/cm^2}$ and $20.51\,\mathrm{g/cm^2}$, while for 20% are $-64.81\,\mathrm{g/cm^2}$ and $31.48\,\mathrm{g/cm^2}$.

In summary, the shape of the pulse contains information that can be associated to X_{max} and the method we presented seems promising. A more sophisticated parabola fitting procedure like the one used in [8] is expected to improve the bias. The bias is also expected to ameliorate for simulations tailor made for this specific method, as we have observed that if we have at our disposal enough simulations evenly distributed around the reference X_{max} the resolution drastically improves. Our goal is to investigate and improve further this method.

Acknowledgements

BMH is supported by ERC Grant agreement No. 101041097; AN and KT acknowledge the Verbundforschung of the German Ministry for Education and Research (BMBF). NK acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 445154105. MD is supported by the Flemish Foundation for Scientific Research (G0D2621N). ST acknowledges funding from the Abu Dhabi Award for Research Excellence (AARE19-224).

References

[1] T. Huege, Phys. Rept. **620** (2016), 1-52 [1601.07426].
[2] M. P. van Haarlem et al, [DOI: 10.1051/0004-6361/201220873] [1305.3550].
[3] T. Huege et al, [EPJ Web Conf. Volume 135, 2017] [arxiv:1608.08869].
[4] T. Huege, M. Ludwig and C. W. James, AIP Conf. Proc. **1535** (2013) 1, 128 [arxiv:1301.2132].
[5] J. Alvarez-Muñiz, W. R. Carvalho, Jr., M. Tueros and E. Zas, Astropart. Phys. **35** (2012), 287-299 [arxiv:1005.0552]
[6] D. Heck, J. Knapp, J.N. Capdevielle, G. Schatz, and T. Thouw, FZKA Report 6019, 1998.
[7] N. Karastathis, R. Prechelt, T. Huege, and J. Ammerman-Yebra, PoS (ICRC2021) 427
[8] S. Buitink et al, PhysRevD.90.082003 [arxiv:1408.7001]
[9] A. Corstanje et al, PhysRevD.103.102006 [arxiv:2103.12549]