Response to: Report of Referee 1

IceCube Collaboration 03/2020

This contribution describes an improved experimental approach to identify the single photoelecton charge distribution of the IceCube PMTs. Several examples are given of how results of this technique can be used to evaluate the performance of different technological choices as well as to describe better the detector in data analysis. However, there are several points needed to be addressed further. This contribution should be published after the authors have made the revisions mentioned in the following.

Thanks for all the constructive comments. We've updated the text to reflect the points that needed to be addressed further. The specific response to each point can be found in the rest of this document for Referee 1.

A reference for the WaveDeform (waveform unfolding process) is needed.

Thanks, WaveDeform was meant to be referenced to Ref. 22 (Energy reconstruction methods in the IceCube Neutrino Telescope).

Page 7, first paragraph: The weight factor has been set to the specific value of 28.4 in order to compensate for the different sizes of the MinBias and BeaconLaunch datasets? Please clarify.

This description should be removed in favor of simply comparing event rates rather than total number of events. The updated text does not include this paragraph, and instead shows Fig. 4 in terms of rates.

Page 7, first paragraph: What are the "late pulses from the trigger"?

The text has been updated to say: "late pulses originating from the trigger pulse."

The trigger pulse (the pulse that passes the discriminator threshold) has a small probability to produce a late pulse. A likely example of this is shown in Fig. 3 (left) actually. The second pulse in that figure may be a late-pulse originating from the trigger. The pulse selection eliminates them from the sample.

Page 8, last paragraph: How this double occupancy (6.5%) has been estimated? (by fitting the charge distributions, as in Fig. 5?). Please explain.

That's correct. It was calculated by fitting to the charge distribution for each DOM, determining the single and two PE distribution, then comparing them. We say that it is an estimate because the MPE contamination is actually depth dependent (muon rate is higher at the top of the detector compared to the bottom). However, we wanted to include a rough estimate for the readers. The text was not updated.

Page 9, Fig. 3: The term "cumulative distribution" has a specific statistical meaning. Please rephrase.

Thank you. We've replaced the term cumulative distribution with "charge distribution summed over all seasons and aggregate charge distribution."

The MinBias and BeaconLaunch histograms, shown in this Fig. 4, are normalized to the same livetimes (i.e. using the weight factor mentioned in Section 1.2)?

Note: "MinBias" has been changed to "Event dataset", and "BeaconLaunch" dataset has been changed to "Forced-triggered (FT) dataset" from the recommendation of Referee 2.

The MinBias and BeaconLaunch histograms were normalized to the same livetime using what we called a "weight". Fig. 3 has since been updated to show the rates of both datasets rather than total number in order to eliminate the discussion on the term "weight."

Why the lowering of the WaveDeform (modified WaveDeform) threshold changes the BeaconLaunch distribution at large charge values?

The lowering of WaveDeform is actually a modification on the termination condition when fitting SPE pulse templates to the digitized waveforms. The termination condition controls when the algorithm should terminate trying to improve the fit. It is determined through the amplitude of the gradient of the least-squares residual in the fit. When the improvement per PE, of allowing one more non-zero pulse, drops below the value of the termination condition, the algorithm will terminate. Larger numbers will produce a faster, coarser fit using fewer pulses, and smaller numbers will provide a slower, better fit with more. The nominal WaveDeform uses a Stopping tolerance of 9, the modified WaveDeform uses a Stopping tolerance of 6.

The modification of WaveDeform at low-PE can also now be seen in Fig. 3. We've updated the description of the termination condition in the text:
"The falloff in charge below 0.13 PE is due to the software threshold mentioned in Sec. 2.1. The threshold is a result of the settings employed within WaveDeform. These settings include a termination condition on the SPE pulse template fitting algorithm that monitors the gradient of the least-squares residual in the fit to the digitized waveforms. When the improvement to the fit, per PE, of allowing one more non-zero pulse drops below the value of the termination condition, the algorithm will terminate. By reducing the value of the termination condition setting, WaveDeform is slower but will produce a better fit to the data and fit and include smaller pulses. The extension to the charge distribution is shown as the green data points in Fig. 3 (right). Reducing this value, however, will also increase the electronic noise contamination. An assessment of this follows."

How the normalization factor M is evaluated?

As described in the above answers, the distributions are now shown in terms of rates.

Page 9, last paragraph: The statement that "The modified WaveDeform datasets show a minimal increase in the contribution of noise to the low- charge region" is not supported by Fig 3.

The argument was meant to say that while there is an increase in the noise, it is still vastly sub-dominant to the non-noise distribution. This is probably more obvious when looking at the linear scale shown below. Since we background subtract the noise, the sentence that is referred to above is not needed, and has since been removed.

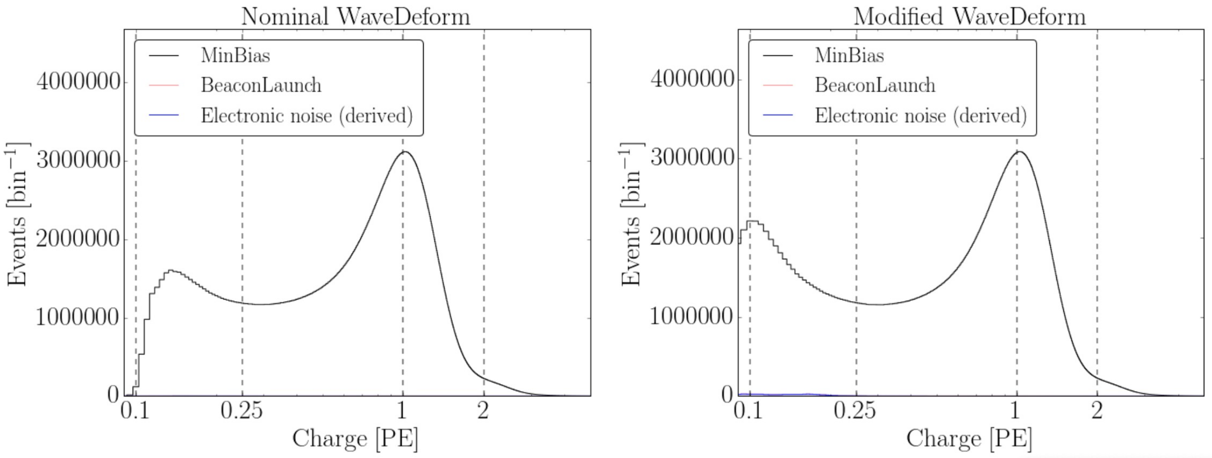

Figure 1: Linear version of Fig. 3 in the SPE paper.

Page 10, 4th paragraph: The assumption that "This analysis assumes the same shape of the steeply falling exponential component (Exp1) for all DOMs in the detector to avoid large fluctuations in the DOM-to-DOM efficiencies." needs confirmation (e.g. by fitting several average charge distributions, corresponding to samples of randomly selected DOMs, estimate the parameters of exp1 and then examine their spread).

You bet. We've tested this by fitting the subset of Std. QE DOMs and the HQE DOMs. They were within error of each other, however the uncertainty in the fit parameters for Exp1 are large, especially for the HQE DOMs. The following response describes why we cannot fit subsets of DOMs for this component.

How is possible to avoid large DOM-to-DOM efficiency fluctuations by the above data-analysis assumptions?

Since Exp1 lies in the low-charge region, the shape and normalization of Exp1 have a significant impact on the DOM efficiency. That is, if you have a significant portion of the photoelectrons producing charge below 0.13PE, they do not make it into the pulse series used for physics analyses. This therefore appears as an efficiency term.

The values we find for Exp1 using the aggregate data (5,160 DOMs in total) are: P_{e1} = 0.186 $\pm$ 0.041 and w_1 = 0.027 $\pm$ 0.002 PE.

Variations of these parameters within the quoted uncertainty, change the percentage of photoelectrons being reconstructed below the 0.13PE threshold by 11%. This corresponds to an equivalent 11% change in the efficiency. This is at a level larger than the expected DOM-to-DOM efficiency differences, therefore it is not possible to fit for this component either on a DOM-to-DOM level (even if we had 5,160 times as much data), or for different subsets of DOMs.

A global efficiency systematic is included in all IceCube analysis. This systematic accounts for a global DOM efficiency change. That is why it is safe to use a single shape and normalization for this component. We need to make some assumption in the region where we have no data (below ≈0.13PE) since the simulation requires us to describe the distribution down to 0.0PE.

Page 10, 4th paragraph: "we background-subtract the BeaconLaunch distribution from the MinBias data". Obviously the BeaconLaunch distribution is dominated by noise. Nevertheless, to be consistent with the previous Section, the noise distribution evaluated in Section 2.2 (i.e. blue histo in Fig 3) should be used for background subtraction. Anyway you should comment on your choice.

Good point, the text has been updated to reflect your comment:
"By subtracting the modified WaveDeform electronic noise spectrum (shown in Fig. 4) from the Event dataset, ..."

Page 10, 5th paragraph: In principle, a first, biased fit will influence the second fit as well (by attributing more weight to the "wrong" data points). Further explanation is needed to support that the applied fit-technique produces unbiased estimations.

While this is true, the bias is extremely small (two orders of magnitude smaller than the uncertainty in the fit).

The second fit amounts to approximately a 1% shift in the Gaussian mean location, or equivalently 2 bins widths. The concern in the above question is whether or not selecting the window around the Gaussian mean can create a bias. We can test this by offsetting the window, say by 5 bins, and seeing if the result changes.

Comparing the result for the Gaussian mean with the window described in the paper to the window shifted 5 bins over, we find that the bias in the Gaussian mean is on average 0.026%. This is negligible when compared to the correction that the second fit performs ($\sim$ 1%) as well as the uncertainty in the location of the Gaussian mean (also $\sim$ 1%). Therefore this is not an issue.

Update to text: "The re-fitting amounts to a percent level change in the Gaussian mean location and the chosen region around the original estimate of the Gaussian mean was found not to produce a significant bias."

Page 10, 6th paragraph: Define better the term“average residual“. Do you mean a correction function as shown in Fig 4?. The residual correction cannot be used as a scaling factor. Do you mean an additive correction (which is proportional to the SPE template, with the proportionality factor given by the function of Fig 4)?

That's correct. The average residual (Fig 4) shows the sum of all the residuals divided by the number of DOMs. The text has been updated:

"Upon fitting the standard WaveDeform Event dataset with the determined values for Exp_1, the residual of each fit is calculated by measuring the percentage difference between the fit and the data. By averaging the residuals over all the DOMs (shown in Fig. 5), a correction function is generated and will be used as a global additive factor for all SPE charge templates to account for the difference between the chosen model (Eq. **??**) and the actual data. All SPE charge templates include the residual correction function in Fig. 5."

Page 11, 1st paragraph: The estimated values of Pe1 and w1 are in principle correlated. Please give the correlation factor.

That's correct and they are highly correlated. This is another reason to use a single value value for Exp1 rather than fitting different sets of DOMs. The correlation between the two parameters is -0.895.

The text now states: "The result of the fit yielded P_{e1} = 0.186 $\pm$ 0.041 and w_1 = 0.027 $\pm$ 0.002 PE, with a Pearson's correlation coefficient of -0.895."

Page 11, 2nd paragraph and last paragraph: Please avoid the use of the term "cumulative".

Removed.

Page 11, Table 1: When fitting the SPE charge distribution of an individual DOM, the estimated parameter values are mutually correlated. A comment on the size of the correlations is needed. In principle, because of these correlations, it is not correct to treat the estimated parameters (Pe1,2 w2, μ and σ) independently, i.e. to form independently their averages, as in this Table. Please comment.

That's a good point. The correlation between the parameters is shown below.

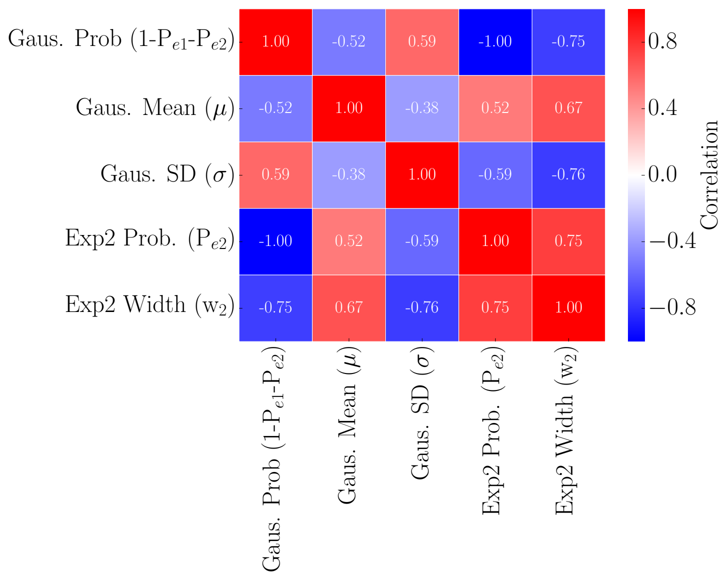

Figure 2: Linear version of Fig. 3 in the SPE paper.

Page 12, Fig. 5: The blue curve, representing the SPE template, differs from the red curve (the result of the fit) because it does not include the contribution of multiple pes. However, the Gaussian component of the SPE template is systematically lower than the data even at charges where the 2PE contribution is negligible. An explanation is needed.

That's correct. The difference between the red and the blue curves in Fig. 5 is the MPE contribution. The 2PE distribution is the convolution of the SPE distribution with itself. For example, since the SPE distribution (blue) is finite at low charges, that means that the 2PE distribution will also have a contribution at low charges. That is why the 2PE contribution is not simply a Gaussian centered at 2PE. In the central limit theorem, it will approach a Gaussian with high levels of MPE contamination.

We've updated the text to make this more explicit. Fig. 5 caption now states:

"The result of the convolutional fit (red) for DOM(1,1) using the Event dataset (black data points with statistical uncertainties), including all data from seasons IC86.2011 to IC86.2016. The extracted SPE charge template from the fit result is shown in blue. For both the convolution fit and the SPE charge template, the curves include the correction from the average residual shown in Fig. **??**. As discussed in Sec. **??**, P_{e1} = 0.186 and w_1 = 0.027 PE are fixed for the individual DOM fits. The difference between the fit result and the SPE charge template is the 2PE contamination. The 2PE contribution is defined as convolution of the SPE charge template with itself, and the SPE charge template is non-zero at low charges, the 2PE contribution is non-zero at low charges."

Page 12, 1st paragraph: The numbers given in this paragraph to quantify the differences between technologies, are not supported by the numbers given in Table 1. Please clarify.

Good catch. The code that printed out the differences had a typo and was printing out the incorrect numbers. Everything is now updated. All errors reported in this section are also now reported as the standard error in the measured quantity or the standard error in the measured difference.

Page 13, 1st Paragraph: The numbers given in this paragraph are not supported by the numbers given in Table 1.

Same issue as above.

Page 14, Fig 8: Please explain how the confidence intervals have been evaluated.

The term confidence interval is not a good choice of language here. Instead, it is now described as the spread in calculated SPE charge templates for a given charge. The 68% interval shows that 68% of the SPE charge templates pass through that region, likewise for the 90% interval.

The caption in Fig. 8 now reads:
"The inner (outer) dark blue region shows the 68% (90%) spread in the SPE charge templates for a given charge. Superimposed are the average SPE charge templates for the variety of hardware configurations shown in white. The TA0003 distribution, for comparison, is shown in orange. All curves have been normalized such that the area above 0.25 PE is the same."

Page 14, 1st Paragraph: Rephrase the sentence "The largest deviation ... in Ref. [9]."

Updated text:
"...The measured distributions were found to be consistent with randomly scrambling the order of each seasonal measurement. Greater than 96% of DOMs are found to have less than 1.0% change per year in the position of the Gaussian mean, in agreement with Ref. [**?**]. This observation holds for the individual subsets of DOMs with different hardware configurations as well. "

The whole Section 3.3 (including 3.3.1) should be rewritten.

We attempted to address the statement that 'The whole Section 3.3 (including 3.3.1) should be rewritten' but without specific criticisms it was difficult for us to ascertain what the reviewer found problematic (typos, language, or the structure of the section). Therefore, we corrected several grammatical errors and we added several sentences at the start of Section 3.3 to better motivate the subsequent discussion:

"In this section we investigate the extent to which changes to the SPE charge templates impact quantities such as the trigger fraction above our discriminator threshold or the measured charge over threshold, since these quantities are used in the analysis of IceCube physics events. If the choice of template has a significant impact on high-level physics results reported by IceCube, it would be a serious concern. Thus we explore the effect of changing the assumed gain response in simulation. The change has different implications depending on the typical illumination level present in different physics analyses. Therefore we explore the effect of the templates in three illumination regimes: dim, semi-bright, and bright illumination."

Section 3.4, Page 17: The significance of this Section to this work is not apparent. Please explain.

We agree that this section is very technical and has little significance to the presented work. As it does add some context to other work in IceCube and will likely be referenced in this way, we would still like to keep it.

Page 17, 2nd paragraph: The sentence "Two simulation sets consisting of the same events were processed through the IceCube Monte Carlo simulation chain to the final analysis level of an update to the IC86.2011 sterile neutrino analysis [24]" is very confusing. Do you mean that two different simulation sets of events are employed, which differ only on the SPE templates used to describe the detector?

That's correct. Two simulation sets of the exact same events were processed to the final level of an analysis, with the only difference being the description of the SPE charge templates versus the TA0003 distribution.

We've clarified the wording:
"Two Monte Carlo simulation datasets were produced, one using the SPE charge templates to describe the charge distribution for every DOM in the detector and the other using the TA0003 distribution. These two simulation sets were produced using the same events and pass through the same IceCube processing chain to the final analysis level of the IC86.2011 sterile neutrino analysis [**?**]. Here, the events at the analysis level are $>$99.9% upward-going (a trajectory oriented upwards relative to the horizon) secondary muons produced by charged current muon neutrino/antineutrino interactions. The muon reconstructed energy range of this event selection is between approximately 500 GeV and 10 TeV.."

Page 17, 3rd paragraph: Please explain further what are the data points (trigger condition etc.).

The data points represent real data after passing through an IceCube event selection. We've added a reference to the event selection criteria (how neutrino events are extracted):

"Here, the events at the final analysis level are $>$99.9% upward-going (a trajectory oriented upwards relative to the horizon) secondary muons produced by charged current muon neutrino/antineutrino interactions. The event selection criteria is described in Ref.26."

The text was also updated to improve the description of the data points:

"Fig. 12 (left) shows a histogram of the total measured charge per DOM, during each neutrino interaction, after noise removal. The black data points represent the measured IC86.2012 data but are statistically equivalent to the other seasons."

As well as the caption in Fig. 12: "A comparison between IC86.2012 (black data points with statistical-only uncertainties) to two sets of simulation at an IceCube analysis level. Both simulation sets were produced with identical procedures, except for one describing the SPE charge distribution using the SPE charge templates (blue) and the other using the original TA0003 (orange) model. Left: The total measured charge per DOM, per neutrino event. Right: The distribution of the total measured charge (Qtot) per DOM, divided by the number of DOM channels (NChan) that participated in the event."

Page 18, 1st paragraph: The statement that "The result of this measurement was shown to be useful for improving the overall data/MC agreement" applies to a particular analysis only. Please rephrase.

Updated text: " The result of this measurement was shown to be useful for improving the overall data/MC agreement for an IceCube analysis as well as calibration of the individual PMTs..."