Response to: Report of Referee 2

IceCube Collaboration 03/2020

We reviewed this interesting paper that describes an improved in-situ calibration of the single-photoelectron charge distributions for each of the in-ice photomultiplier tubes (PMTs) of the IceCube Neutrino Observatory. The characterization of the individual charge distributions is important for PMT calibration and for the understanding of the detector performance. In the paper, the single photoelectron identification procedure and the method to extract the single-photoelectron charge distribution are discussed in detail. The method to extract the single-photoelectron charge distribution uses a deconvolution of the multiple-photoelectron distributions. The paper is of interest and contains sound information and valid scientific methods, thus proving the validity of the results obtained. Therefore, we support the publication of the scientific content. However, we find that the discussion can be substantially improved by complementing the text with clarifications and a systematic organization of the presentation and discussion of the content. A major revision of the text and discussion of the results is required. A few suggestions are listed below.

Thank you. Your comments were very useful and we think that the product of the latest preprint is much improved.

This section contains detailed information and becomes very technical. It contains the technical details of the detector structure and components that are needed later. It would be better to start with an overview of the main goals discussed in this paper, i.e. further expanding on the PMT calibration and why this is relevant to the experiment. It would certainly benefit if a sentence explains or justifies what are the resolution values that are needed to reach a specific measurement. Later, and in a different section called "Experimental setup" or "Detector", the detailed description of the detector can be summarized.

Good suggestion. The text has been updated so that we start with an overview of the SPE charge distribution in the "Introduction" section, then move onto the description of the detector in the "Experimental setup." The introduction text has also been updated.

A sketch of the detector array where the names of the different parts are presented would be a useful and helpful addition for the reader.

We've added a sketch of the detector to Fig. 1.

"There are two versions of AC coupling". If AC coupling is explicitly mentioned, it would be useful to explain what/why they are and/or how those affect the detector response.

The description has been improved with specific detail to how it affects the detector response, as well as more detail relating to their differences. Two references are also included that provide further detail.

"The PMTs are operated with the anodes at high voltage, so in order to remove the DC voltage from the signal, the PMT is AC coupled to the amplifiers (front-end amplifiers). There are two versions of AC coupling employed in the PMTs, referred to as the *new* and *old toroids*, both of which use custom-designed toroidal transformers. The new toroids were designed with a larger ferrite core and more turns resulting in a droop time constant of 15μs, ten times larger than the time constant of the old toroids. The DOMs instrumented with the old toroids were also designed with an impedance of 43 Ω, while the new toroids are 50 Ω. The toroidal transformer effectively acts as a high-pass filter with an approximately flat frequency response up to 100MHz. Single photoelectrons, measured at the output of the secondary winding, however, show percent-level shape differences between new and old toroids. The toroidal coupling method was chosen partial since it offers a higher level of reliability than capacitive coupling and has lower stored energy for the same frequency response [**?**]. Conventional AC-coupling high-voltage ceramic capacitors can also produce undesirable noise from leakage currents and are impractical given the signal droop and undershoot requirements [**?**]. The locations of DOMs with the different versions of AC-coupling are shown on the right side of Fig. **??**. All HQE DOMs are instrumented with the new toroids."

Later in the paper, we specify how it affects the detector response (when discussing WaveDeform):

"The digital waveforms are then deconvolved into a pulse series using software called "WaveDeform" (waveform unfolding process) [**?**]. The pulse series for a particular DOM corresponds to a list of single photoelectron charges and timestamps. WaveDeform is a fitting algorithm that determines the pulse series by fitting a superposition of single photoelectron basis functions to the the digitized waveforms. The basis functions, or SPE pulse templates, depend on the version of the AC-coupling instrumented in the particular DOM. The pulse series, extracted from various datasets, represents the fundamental unit of data used in this analysis. "

Please explain what is "...bifilar wound 1:1 ...". It is very technical as it is.

We've removed this particular description, since we agree that it is too technical. It is left to the reader to look at the provided reference if they would like a more technical description of the toroidal transformer.

Page 3, line 3: what is a "scaled single PE pulse"? is it a "superimposition of single PE pulses"? Please clarify.

That's correct. This is a better way of describing it. The text now reads:

"This is accomplished by deconvolving the digitized waveforms into a superimposition of single photoelectron pulses (so-called pulse series), and the integral of the individual pulses divided by the load resistance defines the observed charge. "

page 3, 1st paragraph: 1.2mV is equivalent to 0.23PE. How is the equivalence set? please explain. Also, later in the text, a different equivalence of 2V and 560PEs is mentioned (beginning of page 6). It would be useful to clarify the source of the apparent inconsistency.

The description should have said "560 SPEs" rather than "560PEs". The text has been updated. This is probably the origin of the inconsistency that you mentioned.

The equivalence is set by: "... the integral of the individual pulses divided by the load resistance defines the observed charge."

The peak voltage of a properly amplified SPE in the R7081-02 PMT with a gain of 10^7 will be $\approx$5.2 mV (Sec. 2). The discriminator is set to 1.2mV via a DAC. It's a linear conversion between peak voltage and charge, therefore 1.2 mV is approximately 0.23PE. This can also be verified by fitting an error function to the discriminator threshold, which also yields 0.23PE.

In order to produce a 2V signal, we require a superposition of many SPE pulses. The average SPE pulse does not produce 1.0PE since there is the low-PE charge component in the SPE charge distribution. Instead, the average SPE pulse produces on average approximately 0.7PE of charge (the mean of the charge distribution). Since a 1.0PE pulse has a peak amplitude of 5.2 mV, we would require approximately 560$\times$0.7PE pulses to reach 2V. This is why we now state 560 SPEs rather than 560PEs.

P.3, 2nd paragraph: high-gain and low-gain. It is not clear how many and where these different stages are.

Since we only use the highest-gain channel for single photoelectron detection, we've removed the description of the other channels to simplify the description. This update should address the next three questions as well. The updated text is:

"When one or more photoelectrons produce a voltage at the anode sufficient to trigger the onboard discriminator, the signal acquisition process is triggered. The discriminator threshold is set to approximately 1.2 mV, or equivalently to $\sim$0.23 PE, via a digital-to-analog converter (DAC). The signal is presented to four parallel channels for analog amplification and digitization. The highest-gain channel has a nominal amplification of 16 and is most suitable for single photon detection, and is thus used in this analysis. This channel is digitized at 300 million samples per second (MSPS) for 128 samples using a 10-bit Analog Transient Waveform Digitizer (ATWD). Two ATWD chips are present on the DOM Mainboard (MB) and alternate digitization between waveforms to remove dead time associated with the readout.

Are the three channels separately amplified with different gains? and why? perhaps the sentence should read "are each subject to different levels of separate amplification"?

Updated text in the previous answer addresses this.

Answer: The three channels are separately amplified with different gains. The low-gain channels are used for very bright events (e.g. a cosmic ray muon passing near the DOM) that saturate the high-gain channel. The high-gain channel is used for single photoelectron counting.

Three channels are mentioned but then there is a reference to only one high-gain channel. Please clarify and structure the sentence such that a complete description is given.

The text above addresses this.

It is incomplete and confusing as it is. Also, it does not serve any purpose unless there

is a clear indication of the reason why this detailed information is relevant. Rather, it would be useful to clarify the reason why the signal is split into four channels.

The new text above hopefully relieves the confusion as well.

P.3, 3rd paragraph: It may be intuitive, but please define "SPE", as it is the first time it appears.

Now, "SPE" is defined in the first sentence of the paper:
"This article discusses a method for determining the in-situ single-photoelectron (SPE) charge distributions of individual photomultiplier tubes (PMT) used in the IceCube neutrino detector."

Line 2: it would be better to replace "corresponds" with " is calibrated to correspond"

Good idea.

Line 3: "that create a spurious structure"

Thanks.

Of the items listed there are a few that are intrinsic properties/imperfections of the PMTs and can therefore be testes separately in the lab. Have those effects been tested before? are similar results obtained to those presented here?

There exist (mostly internal) lab measurements of several of the listed properties. These include: afterpulse, late pulse, pre-pulse, and dark noise measurements. There was also an internal test performed many years ago that illustrated that the SPE charge distribution changes depending on the location that the photon struck the PMT.
A description can be found in https://arxiv.org/pdf/1002.2442.pdf. The in-situ measurement naturally takes into account the arrival position and direction distribution as encountered in the physics data.

The statistical fluctuation due to cascade multiplication and dark noise come from the description by Hamamatsu.

With that said, the addition of Exp_1 was empirical. It was added to improve the fit quality to the data – since this is primarily what we are after. Attempting to measure the charge distribution of each of these properties (say, measuring the charge distribution solely from pre-pulses) would be interesting though, and potentially a good expansion to the measurement in the future.

We've added a statement about the empirical validity of the updated functional form:

"The steeply falling component above the WaveDeform threshold and below the discriminator (0.13 PE to 0.23 PE) is in agreement with the laser measurements mentioned in Sec. **??**. Since small pulses below the DOM's discriminator threshold will be recorded in events with multiple photoelectrons, it is important that the SPE charge distribution accounts for this shape. Exp_1 was specifically introduced into Eq. **??** to account for this component. It was an empirical choice to use an exponential to represent this region of the charge distribution."

Eq.1.1: It would help the clarity of the paper to clearly define the different parts of Eq.1.1, provide a justification for the choice, and state how each component is estimated. For example, do exp1 and exp2 provide a description of the background? if so, why not stating this clearly? Please clarify.

The exponentials do not describe a background. The electric background contribution (usually called "pedestal" in SPE charge distributions), is negligible to this study as discussed in Section 3.2. They rather describe the under-amplified portion of photoelectrons as found in any PMT. The exponential plus Gaussian is a very common (yet empiric) choice in the field. As mentioned two questions ago, the second exponential is an empirically motivated extension of this model to better describe the very low charge contribution. Lab measurements confirm that these very low charges are in-time with a triggered light source and as such a real effect of the PMTs.

Eq.1.1: Please provide the explicit definition of ”Erfc” used. This can be done in the text (preferably) or in a footnote.

Thanks, updated to have it defined inline.

Sec.1.2, 3rd paragraph: The sentence including "droop and undeshoot" and "expected reaction" is technical/jargon and it should be clarified. What is the "expected reaction"?

Droop and undershoot are common terms used for signal processing so we’ve added a good reference to a text book that describes them (Techniques for Nuclear and Particle Experiments), as well as an improved description in the text:

"After waveform digitization, there is a correction applied to remove the DC baseline offset. Distortions to the waveform, such as from droop and undershoot [**?**] introduced by the AC coupling, are compensated for in software during waveform calibration. The area encompassed by the distortion of the waveform due to these effects is proportionally related to the charge that produced them, and is therefore commonly a problem for bright events or high frequency trains of photons. The expected distortion can be calculated if the input charge is known, and is compensated for by adding the determined reaction voltage of the distortion to the calibrated waveform. If the undershoot voltage drops below 0 ADC counts (outside of the minimum ADC range), the ADC values are zeroed and then compensated for once the waveform is above the minimum ADC input."

Sec.1.2, ”MinBias dataset” subsection: it is not clear what ”1:1000 events” is. Does it refer to a dataset that collects 1:1000 of the randomly-selected events? Please clarify.

It corresponds to collecting on average 1 in every 1000 random events. The text was modified to:

"This dataset preserves the full digital waveform readout of randomly-selected HLC events, collecting on average 1 in every 1000 HLC events."

”16 triggered DOMs”: are there 16 DOMs that trigger the events? are they adjacent?

please clarify.

We've rephrased "16 triggered DOMs" to "16 DOMs that observed a discriminator crossing," which should clarify the wording. The DOMs do not have to be adjacent, the trigger condition (see the description of a HLC in Sec. 2.2) requires only that a DOM and its nearest or next-to-nearest neighbor observe a discriminator threshold crossing.

Updated text: "The average event consists of approximately 26 PE distributed over an average of 16 DOMs that observed a discriminator crossing. The full digital waveform of these events allows us to extract the raw information about the individual pulses. This dataset will be used to measure the individual PMT charge distributions."

Sec.1.2: This is an example that occurs elsewhere and in similar acronyms in the text. Usage of acronyms like "WaveDeform" or "BeaconLaunch" that are used several times throughout the text could be avoided. Those acronyms are jargon, they are not intuitive, and do not correspond to the names given in parentheses, thus making hard to connect to the actual meanings. It is certainly a straightforward correspondence for those in the IceCube collaboration, but it is hard for the average reader that has to constantly refer to the prior and non-intuitive explanation. Please consider using alternative names, or provide correct acronyms.

Good comment. We've changed the term "MinBias dataset" to "Event dataset", and "BeaconLaunch dataset" to "Forced-triggered dataset (FT)". The term WaveDeform itself is an abbreviation that describes the algorithm rather well, so we'd prefer to keep it.

Fig.1, caption. "OTS": same as the comment above.

OTS is an acronym commonly used for "OuT of Service", however, we agree that it's not very intuitive. Instead, we've changed it to RFS for "Removed From Service", which is also a better description of the condition.

Sec. 2.1, 3rd paragraph: "ATWD time window"? does it correspond to the "time integration window"?

The ATWD time window corresponds to the total length of the ATWD readout "300 million samples per second (MSPS) for 128 samples," which corresponds to 427ns.

We've update the text: "300 million samples per second (MSPS) for 128 samples (a readout time window of 427 ns)"

Sec.2.1: 3rd paragraph: "light grey region". It is hard to identify the different areas. Sometimes it is referred to the color, other times to the time span. Letters for each area would simplify the identification, remove ambiguities, and make the text uniform.

That's a good idea. We've added lettering for the windows, and updated the description. The time values are important, therefore they were left in the text. See Fig. 3 for the update.

Sec. 2.2, 1st paragraph: "threshold at ... 0.13PE" is hard to see/justify in Fig.2 (right) is

referred as due to a termination, and it is referred to in the Fig.2 caption as "fall off in charge ... 0.13PE". Please give consistent explanations.

The vertical line in Fig.2 (now Fig. 3 right) is now drawn at 0.13PE in order to make the area that we are describing more explicit. Below 0.13PE, the rapid fall off in charge is due to the termination condition in WaveDeform.

Fig. 3 (right)caption now states: "The falloff in charge below 0.13 PE (red dashed line) is due to the termination condition in WaveDeform."

Sec.2.3: 2nd paragraph: "...contamination from two-photon events...". Why is this more important than other effects? is this due to what? why is this effect highlighted or important? Please clarify in the text. The next paragraph mentions "three photon contribution". Same as above.

It has to do with the MPE contamination being a much larger unwanted contribution to the charge distribution than the other effects, and has a distribution much different than that expected from a pure sample of single photoelectrons. For context, we state that we find an estimate of the MPE contamination is ≈6.5%.

As stated in the text, the contamination from two-photons is due to:

1. MPE contamination
2. The summation window in the pulse selection. It is necessary to perform a sum over some period since occasionally split an SPE pulse into multiple smaller pulses. We must reassemble pulses that may have accidentally been split.

Dark noise from the photocathode, late pulses, and after pulses have a distribution similar to an SPE distribution. Late pulses and after pulses were already at the percent level before the pulse selection, and further reduced by the pulse selection.

Prepulses are rare (sub 1%) and result in a 10-20 fold decrease in the measured charge therefore landing below the 0.13PE threshold. Thermionic emission from the first (or later) dynode also falls in this region. Therefore these are also not of concern.

The photocathode thermionic emission rate is measured to be 350Hz. The accidental coincidence with the ATWD readout time window (in the summation window from 100ns to 375ns) is approximately 0.1%. So this too is subdominate to the MPE contamination.

We show in the paper that the electronic noise has a SNR of 1.98x10^5 and is therefore negligible.

The three-PE is also discussed in the text: "We do not account for the three-photon contribution, which is justified by the lack of statistics in the 3 PE region as well as the significant rate difference between the 1 PE and 2 PE region, as shown in Fig. **??** (right)."

Sec.2.3, 4th paragraph: "...assumes the same shapes ... for all DOMs...". The assumption is based on what? This assumes that the dependence is purely due to the PMT intrinsic

characteristics and not to local in-situ differences? Also, the PMTs are of different nature/characteristics. Does the assumption remain valid? Please clarify in the text.

It was a choice that was needed to be able to make this measurement. The text in the Fitting procedure section has been updated:

"Although the pulse selection and the modification to WaveDeform allow us to probe the charge distribution below the discriminator, there is insufficient data to accurately fit Exp_1 to each DOM independently. Rather, we determine the value of Exp_1 after summing all the data together, and use that result for Exp_1 in all subsequent fits.

The distribution is purely given by PMT and electronics properties, with no dependence on local in-situ differences. The motivation to not fit for Exp_1 independently for all DOMs is as follows. Pulses that fall below the WaveDeform threshold do not contribute to the pulse series. This is equivalent to a efficiency term for DOM in question. Having large variations in the Exp1_1 component will in turn cause large variation in the individual DOM efficiencies. Rather, setting all DOMs to have the same Exp1_1 component means that each DOM has the same efficiency penalty. In IceCube analyses, we include a global efficiency systematic uncertainty which accounts for this.

Using the aggregate of the entire ensemble of DOMs with the modified WaveDeform datasets, we background-subtract the FT distribution from the Event dataset, fit the resulting distribution to determine the components of Eq. **??**. The determined shape and normalization of Exp_1 is then used in all individual DOM fits using the standard WaveDeform."

Sec.2.3, last paragraph: "predetermined values" from what/where? what does it refer to? Is the value fixed within certain predetermined boundaries? or does it refer to values determined earlier from tests in the laboratory? Please clarify.

Apologies, this paragraph was out of order. It was referring to the measured value of Exp1. This was largely re-written (Sec. 3.3).

Fig.3: It would help the reader to extend the range of the horizontal axis to 4. As it is, it is not clear what is the current boundary. Please adjust.

Good idea. We've added the 5PE tick to the axis.

Page 11, 3rd paragraph: This paragraph is confusing. A few points: "failed fits". Are they due inefficiency or to what? what is the "number of valid pulses"? what is a "valid pulse"? "goodness" by large chi-square? Please clarify.

The text has been modified to:
"An SPE charge template for a particular DOM is classified as having a "failed fit" if it is flagged by one of the validity checks. The validity checks ensure that there is sufficient data extracted from the pulse selection and that the fit was successful, as well as goodness of fit check on the reduced chi^2. The majority of the failed fits (approximately 95%) are due to DOMs which have been removed from service (see Fig. **??**). The remaining failed fits were found to be associated with DOMs that have known issues associated with their charge collection. All the DOMs with failed fits are not included in this analysis. In the

simulation, the average SPE charge template is assigned to these DOMs. "

To answer your question directly:

1. Failed Fits: fits to the collected charge distributions that did not meet a validity check. 95% of these failed fits are DOMs that have been removed from service and therefore do not collect any data to fit to. The remaining 5% of the failed fits are DOMs that are partially operational (for example, one of our DOMs has the discriminator set to 0.75PE, so the fit doesn't perform well).
2. Number of valid pulses: Number of pulses extracted from the pulse selection for a particular DOM.
3. Valid pulse: The pulses that pass the pulse selection.
4. Goodness: We're more explicit now. It is a reduced χ^2 test that flags DOMs as well as a check to ensure the fit was successful. Essentially, IceCube has 6 DOMs out of 5,160 that have some issue with them but are still taking data.

Page 3, 3rd paragraph: "between 107 and 111 DOMs": In caption of Fig.1, it says that the DOMs shown in white have been removed from service. However, here it says that DOMs have been removed depending on the season. The two statements seem to be contradicting. Please clarify.

The number of DOMs that have been removed from service (RFS) varies over time due to DOM failure. From IC86.2011 to IC86.2016 the number of RFS DOMs was found to range between 107 and 111.

Updated text: We've moved the number of failed DOMs description into the caption of Fig. 2. It now states: "...DOMs that were removed from service (RFS) for the IC86.2016 season are shown in white. These are primarily DOMs that resulted in failures during deployment and failures during subsequent operation. Due to subsequent DOM failures, the number of RFS DOMs from IC86.2011 to IC86.2016 varies between 107 to 111."

Why only a few and not all parameters of the fit are indicated in the Table? - What is NQE?

The only parameters not explicitly included are that of Exp1. We hold these fix at P_{e1} = 0.186 $\pm$ 0.041 and w_1 = 0.027 $\pm$ 0.002 PE. It is a good idea to include the value of Exp1 in the caption though. The caption has been updated: "The average values and standard error of each fit parameter for the subset of hardware configurations listed in the first column. The fitted parameters for Exp_1 for all configurations are P_{e1} = 0.186 $\pm$ 0.041 and w_1 = 0.027 $\pm$ 0.002 PE, as described in the text."

NQE was previously "Normal QE", however we recently changed the naming convention to Standard QE – this was a typo and should have been "Std. QE".

"Width": does it refer to "sigma" of the distribution? Please clarify.

In Eq. 1, the Gaussian width was defined as σ, corresponding the the standard deviation of the Gaussian. To address the concern here, we've change the term "Gaussian width" to "Gaussian standard deviation", or "SD".

Column 2: does it refer to Exp1 or Exp2?

The column header states that it is for Exp2. Exp1 is held fixed for all subsets of DOMs. This is now stated in the caption.

Caption: what is the ”standard error”? please clarify.

The standard error is the statistical accuracy of a measured quantity. Here, we are calculating the mean values of Pe1, w1, 1-Pe1-Pe2, sigma, and mu. The uncertainty in the value of the mean is what is called the standard error. The standard deviation is the width of the distributions of Pe1, w1, 1-Pe1-Pe2, sigma, and mu.

Standard error of the mean is calculated as the standard deviation of the distribution (Pe1, w1, 1-Pe1-Pe2, sigma, and mu), divided by the square root of the number of samples (number of DOMs).

To make the description a bit more rigorous, we've replace the term "standard error", with "standard error of the mean".

We could also think about reporting the standard deviation.

What is ”validity”, ”PVR”, etc? please clarify in the caption.

Thanks. Validity corresponds to whether or not there the DOM was a failed fit or not. PVR is the peak-to-valley ratio. We've removed the "validity" text and wrote out "Peak-to-valley ratio" explicitly in the image.

Legend: add something that indicates the data points.

The MinBias histogram was the data. Rather than a histogram for the data, there are now data points with statistical errorbars. We think this addresses the confusion.

The fit results correspond to ”convolution fit” in the legend? Please clarify in the caption.

That's correct. The legend now states "Fit result" rather than "convolution fit", and in the caption is says "..the result of the convolutional fit..." to tie the two together.

caption: ”black histogram” is rather ”black points”

The data was represented as a histogram. The new figure is now data points with uncertainties.

caption: ”Pe1 and w1 are fixed as discussed at the end of Sec.2.2.” Where exactly is this discussed? where are these values indicated? the discussion in Sec.2.2 does not indicate the values. If so, please clearly discuss and rephrase the Section accordingly.

Thanks, it looks like reordering of some paragraphs changed the location of these values. Pe1 and w1 are now discussed in the updated Sec. 3.3. The values for the fit are presented in Sec. 3.4.

Sec.3.1, 1st paragraph: "it is evident that... depends on the DOM hardware". Results in Table 1 are all within one sigma from each other. They are all consistent with each other. Please clarify.

Could it be that you are looking at Table 2? Table 1 shows the uncertainty between the different DOM hardware configurations as many sigma from each other. This is part of the reason why we report the standard error of the mean rather than the standard deviation.

Sec.3.1, 1st paragraph: how is the "peak-to-valley ratio" defined? Please clarify in the text.

Updated text: "...These differences can also be seen in the measured peak-to-valley ratios and average charge of the SPE charge template (see Fig. **??**). The peak-to-valley ratio is defined as the ratio between the maximum in the SPE charge template near the SPE peak divided by the minimum found near the discriminator. "

page 13, 2nd paragraph: " a ... change of the fit parameters... according to their PMT serial number...". Given the difference between PMTs, is it reasonable to use the assumption in Sec.2.3 ("...this analysis uses the same shape...")? Please clarify in the text.

We've attempted to perform a separate fit for the HQE DOMs versus the Std. QE DOMs to determine Exp1. They were found to be within error of each other. However, this is strictly due to having larger uncertainties in the fitted value of Exp1. If we were able to accumulate several orders of magnitude more data, then perhaps we would be more comfortable to differentiate the Exp1 between the different Hardware differences.

Sec.3.2, 1st paragraph: "excluding Exp1". Why is Exp1 excluded? Please explain.

We have a single value for Exp1 fit using all the data.

Sec.3.2, 1st paragraph: "randomly": Seasons are not random- but time/temperature- dependent. Please justify the choice.

Apologies, here, the word "Season" is referring to the "IceCube season" defined in Sec. 1.2, which corresponds to near a full year of data (May to May). The Caption in Fig. 10 now reads: "The change in the individual DOM fitted parameters from IceCube season IC86.2011 to IC86.2016. The histogram represents the rate of change, per IceCube season, of the fitted parameter labelled in the legend."

The text was also updated:
"The SPE charge templates were also extracted for each IceCube season independently (IC86.2011 to IC86.2016), in order to investigate the time dependence of the fit parameters. The data was broken up into IceCube seasons rather than years, since calibration to remove gain drift is performed at the beginning of each IceCube season.

For every DOM in the detector, the change over time of each fit parameter (excluding Exp_1, since it was held fixed for this analysis) was calculated. Fig. **??** shows the change in a given fit parameter, relative to the mean value, per Ice-

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

Sec.3.3, 4th paragraph: ”Q 0.25”. what does it mean? Is 0.25 the threshold level? Please clarify.

This is approximately the discriminator threshold (0.23PE). To avoid confusion, the numbers were re-calculated using 0.23PE.

Sec. 3.5, 1st paragraph: ”we must implement”. Does it mean that it has been implemented already? if so, please correct. Also, how is this implemented in simulation?

It has been implemented already. The sentence following this one describes how it is implemented. The text has been updated:

"To model the IceCube instrument, the PMT response is described in the simulation. The IceCube simulation chain assigns a charge to every photoelectron generated at the surface of the photocathode. The charge is determined by sampling from a normalized SPE charge distribution probability density function (PDF). A comparison to data between describing the SPE charge distribution PDF using the SPE charge templates and the TA0003 distribution follows."

Sec.3.5: perhaps, you may want to replace: ”between describing” ->”with”. - Eq.3.4: is it 0.10 or 0.13? What is ”Q 0.10”? please clarify.

We were approximating the threshold as 0.10PE. Since we discuss in detail the threshold at 0.13PE, we have now changed the description to 0.13PE.

Sec.3.3.1, 1st paragraph: What is a ”bright” or a ”dim” event? please clarify.

We describe the definitions of bright and dim sources in the subsections: "Dim source measurements" and "Bright source measurements".

To improve the description, we now give an example of a bright event: "For light levels that are large (e.g. a cosmic ray muon passing near the photocathode), ..."

Sec.3.3.1, 1st paragraph: what does the ”bright-to-dim ratio” represent? please clarify.

Measuring the shape of the charge distribution below 0.13PE is difficult, however, it is expected that these "runt" pulse exist. These pulses are small, and we can think of them as "extra", in addition to the above-threshold ones that we can easily measure in the lab and count. In the case where only few photons arrive at a given PMT, so no more than one would be detected, any runt pulses will fail to trigger the DOM and therefore cannot have any effect. Only when there are enough photons for runt pulses to “ride along” with others that send the PMT response over threshold, then the runt pulses can shift the total charge. However, the shift is small (since by definition they carry little charge):

for example, if we add in 10% more pulses, each with typical charge around 0.05 spe, then the total charge collected is only going to be affected by 0.5% (a very small number). That is, bright events would then collect 0.5%.
The bright-to-dim ratios show this explicitly. They show that even though the shape, particularly at low charges, is changed rather dramatically with respect to the TA0003 distribution, we do not expect a bias to be introduced into bright events (compared to dim events).

Table 2: the errors seem to be extremely small. How are they calculated? do they include systematical as well as statistical uncertainties? For example, by looking at Fig.8, it seems the uncertainties are rather large and not as tiny as those indicated in Table 2. If those are the errors on the mean, it is not clear from the presentation.

The errors in Table 2 represent the standard deviation of the calculated quantities. We've attached the distribution for the subset of DOMs in the figure below.
The caption was updated to: "The distribution in bright-to-dim ratios for the previous charge distribution (TA0003) and the individual DOM SPE charge templates for the IceCube and DeepCore detectors. The uncertainty reported represents the standard deviation of the distribution of the calculated quantity."

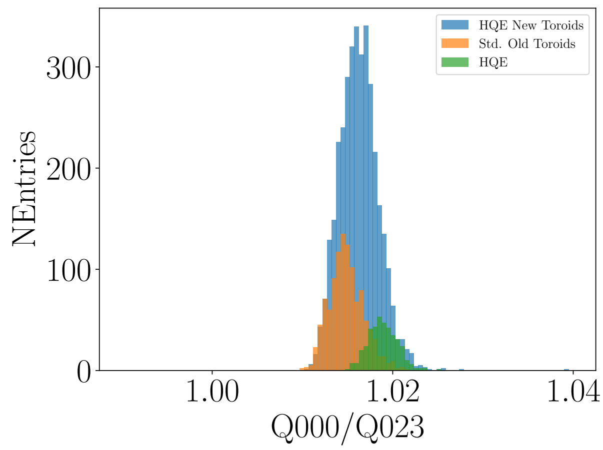

Figure 1: The distribution of one of the columns in Table 2 of the paper.

Fig.10: The original "TA0003 model" differs from the "SPE template"

The caption has been updated to:
"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."

caption: is it simulation or from the SPE template fit?

It shows data/mc agreement at an analysis level when we use the SPE charge templates to describe the SPE charge distribution in simulation, compared to when we use the TA0003 distribution.

Fig.10, caption: a few suggestion to clarify/correct the text: - ->"...to the 2012 IceCube season (data points)".
->"...data are shown in black."
"...number of DOM channels (NChan)

Thanks, the changes have been added.

Page 18, 2nd paragraph: why HQE DOMs have a smaller peak-to-valley ratio with respect to standard QE DOMs? From where is this information taken? Is it clear why? Naively, one would expect the opposite. Please clarify.

We expect this behaviour if the HQE DOMs are more efficient at collecting photoelectrons emitted from near the edge of the photocathode. The peak-to-valley ratio for a photoelectron emitted near central axis from the photocathode is much higher than that emitted near the perimeter. We find that the peak-to-valley ratio approaches unity near the edge, this is described in Sec. 8 of https://arxiv.org/pdf/1002.2442.pdf.

a few lines below: "...corresponding in an average increased peak-to-valley ratio...". What does it mean? Please clarify.

It should say that the old toroids have a larger peak-to-valley ratio than those instrumented with the new toroids.

Updated text: "...had narrower Gaussian component and an increased peak-to-valley ratio of $14.36 \pm 0.01\%$. "