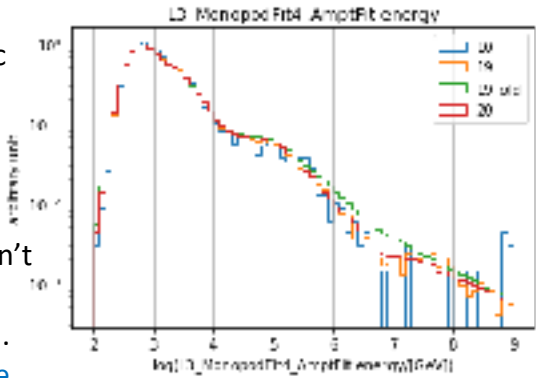
The answers in red will be done soon and the answers in black are finalized.

BAC: I added more comments/questions in blue.

Feedback:

1. Please clarify what “baseline analysis based on pass 2 MC” means
    - Sorry I mean I did some analysis without considering systematic uncertainties. These include data mc agreement check, study on the change/update of following components: 1. baseline atmospheric neutrino flux model, 2. ice model for monopod reconstruction, 3. calculation of self-veto effect. They are described in section 6 more detailedly.
2. For the “self veto”: please give a few line explanation of the main changes.
    - The main changes includes: 1. the new self veto allows including systematic uncertainties and detector characteristics. 2. the new self veto considers some cases which is approximated by the old one. For example, the new self veto takes into account fluctuations from stochastic losses of muons propagating through the Earth toward the detector and includes muons from other branches of the shower in which the neutrino was produced.
    - BAC: Can you please add this information to the wiki. E.g. the line about fluctuations from stochastic losses.
    - Sure. I have added it.
3. On the “Level 3” section: what are the backgrounds that the “Cascade L3 processing” is trying to reduce?
    - Cascade L3 processing reduce both muon and atmospheric neutrino background. For example containment cut reduces the muon background because muon is track like signal which is less likely to be contained in the detector region. In addition, coincident events (coincident muons and coincident muon+neutrino) are reduced.
    - BAC: Can you please add this information to the wiki.
    - I have added it.
4. What are “single events”? Do you mean events w/o coincidence atmospheric muon?
    - yes, single events means events w/o coincidence.
    - BAC: Can you please add this information to the wiki.
    - I have added it.
5. How is the 2.55/3 decided on? It is not the ratio of 79/86 (which I also do not think is the right thing to do).
    - 2.55/3 is the ratio of event rate of 2010 and other years. I just rescaled ic79 data to compare the shape of other variables with other years.
    - BAC: So the 2.55/3 was decided totally empirically? You just scaled it by eye?
    - Yes it is.

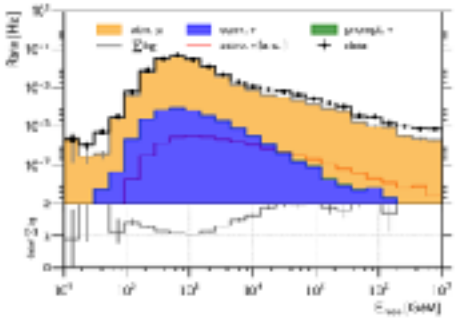
6. Was this followed up on? This is a very clear difference for 2010 and 2020, separated by almost 10 years?
    - This dues to the change of software from icerec to combo. 2010 and 2020 are processed by combo while other years are processed by icerec. I reprocessed 2019 with combo (yellow line) and it moves close to 2010 and 2020 from original position (green line). We don't think this is critical to our analysis because the data is very background dominated at this level.



    - BAC: Can you please add this information to the wiki? I would buy that this doesn't affect the analysis at this level, but this makes me uncomfortable. Have you performed year-to-year consistency checks at a higher level? Does the software group know about this?
    - I will check it later

7. Sorry, this does not make sense to me. Can you link to further explanation about why the cascade passing rate changed? Does the new filter just have a different passing rate?
    - We looked up related document. And here is the conclusion:
    - This is the proposal for CascadeFilter_13 (https://user-web.icecube.wisc.edu/~mlesiak-bzdak/Documents/Proposals/cscd-filter-l2-ic86-2013-v07.pdf). It listed the cuts and changes comparing with CascadeFilter_12. The main changes are that a new cut on NString > 1 is proposed and the cutting parameters are changed. For pass1 CascadeFilter_13 is used starting from 2013 while for pass2 it is used for all the years. According to the proposal (Table 1 on page 6), the difference between the level2 event rate for 2012 filter and 2013 filter does not explain the difference shown in our figure. The table shows that 2013 filter has a higher rate (30.0) than 2012 filter (28.5), while our figure shows a higher rate for 2012. Then we looked up Hans' thesis (https://www.stonybrook.edu/commcms/grad-physics-astronomy/_theses/niederhausen-hans-may-2018.pdf). The table 3.1 on page 50 agrees with our plots. Year 2012 has a higher level3 single contained rate (193.9 mHZ) than other years (~155 mHZ). The reason is unknown yet **but the data agrees with corsika well for pass1 (2012 corsika also has a higher rate than 2011 corsika).** Besides, for pass2 analysis, we don't necessarily require the agreement with pass1, especially for the very early years when the filters and software are not stable yet. The uniform across years and MC/data agreement is more critical.
    - BAC: So, there are two issues here, right? The first is: 2012 has a higher passing rate than 2013-2015. This is unexplained, but consistent with what Hans saw, and the data/MC looks okay apparently. The next issue is that pass 1 disagrees with pass 2 in 2012 in a way that it doesn't disagree in 2013-2015. So pass1->pass2 is doing something different in 2012 than in other years. This is also unexplained, correct? But again you are suggesting the data/MC is fine, and so this is alright? In any case, can you please add all this information to the wiki. Also, I'm not sure I see the data/MC plots to support this? Sorry if I'm just missing them!
    - answer it later

8. Sorry, this is also not clear to me. Can you add more documentation about why this is true? This otherwise seems a clear change in the high-energy cascade rate.
    - The plot on the right hand side is from Hans' thesis. It shows each component of level3 single contained events. So at this level, the sample is still muon background dominated. And we can see in the later part that in the final cascade sample, the high energy events looks identical. Then this difference should



not influence our fitting result.

- BAC: Can you please add this explanation to the wiki (in a spoiler would be fine.)
- Sure. I have added it.

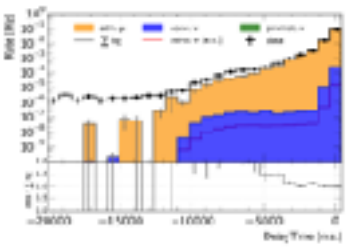
9.

	2010	2011-2020
active strings requirement	79	86

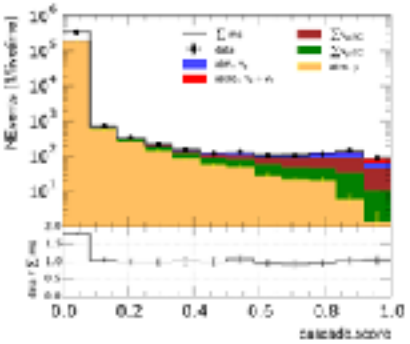
I am confused. In previous sections, you gave comparisons for 2010—the 79 string version of IceCube. Why are you now saying that you only select IC86 configurations? Also, does this means that runs with strings dropped (does this ever happen?) are not included in the analysis? Also, can you please clarify if deepcore is included in the analysis? Or is this only based on non-DeepCore DOMs?

- Sorry, I did not make it clear here and the wiki is updated. We request the runs with all strings are good. Which means for years 2011-2020, we request the runs with 86 strings are active and for year 2010 we request the runs with all 79 strings are active. What we dropped is the runs with bad strings, for example the active strings of run 122299 in 2013 is 85. (https://convey.icecube.wisc.edu/data/exp/IceCube/2013/filtered/level2/IC86_2013_GoodRunInfo.txt).
- And deepcore is included in the analysis. We did monopod reconstruction both with and without deepcore. The cuts depend on both of them and the final energy and direction is from the reconstruction with deepcore.
- BAC: Thank you for the explanation and for cleaning up that part of the wiki!
- No problem!

10. How are the outliers defined quantitatively? Can you also add some small comment about why the rate various throughout the year?
    - The outliers are defined by 20% deviation from rolling average of the neighborhood. And the rate variation throughout out the year due to the change of atmospheric density throughout the year, lead by the change of atmospheric temperature.
    - BAC: Okay, thank you. Do the final good run lists have the outliers already removed?
    - No problem. No, they do not have the outliers.
11. Please add more narrative. E.g. “After we select events passing the L3 filter, we apply a further selection. The selection is different for high…”. Also, please try to make it clearer the sets of cuts being pursued. I might suggest that you re-order the wiki page so that it goes “level 3 cascade selections” and then immediately “higher level cascade selection.” And then you can put all the pass1/pass2 checks in to the “appendix” so to speak. Otherwise the cuts are interrupted by this big section on pass1/pass2 comparison, and on the year-to-year comparison.
    - I re-ordered my wiki, thank you.
    - BAC: Thank you! I think this reads much better.
    - Thanks!

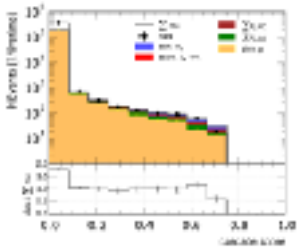


12. Please explain the purpose of this cut.
    - The background is dominating at this level. This cut is inherited from the previous cascade analysis. Previously it is used for getting a better data/mc agreement. As the figure shows there is a large discrepancy below -5000.
    - BAC: Please add this explanation to the wiki. (In a spoiler would work.)
    - I have added it.
13. Like For point 11, please add more narrative. Also, explain why the L5 cut is needed. I guess it's because the muon background in the cascade L3 sample is still too high?
    - Yes. You are right. From the plot above, after level4D, atmospheric muon still dominants.
14. Can you please add more explanation of why applying the pass 1 BDT to pass 2 data is okay? BDTs, or ML techniques in general, can be extremely sensitive to e.g. data/MC mismatches. So I think justification needs to be added here.
    - We compared the bdt variable distributions on level5a. Pass1 corsika, which was used for training bdt, matches with pass2 data. And we re-trained the bdt for pass2 and burn sample below 10 TeV. The new model does not performs better. This is added to wiki, see my wiki for detailed information.
    - BAC: As you point out, there is better data/MC agreement in zenith when you use the pass 2 BDT on pass 2 data instead of the pass 1 BDT. But your second set of plots demonstrate that the data/MC agreement between applying pass 1 trained to pass 2 is similar to applying pass 2 trained to pass 2, so I guess it's alright on the whole because the BDT affects both the data and MC the same. From our offline discussions, I think the basic conclusion is that you don't want to take the time and burn the MC statistics to retrain the BDT over the full energy range (0-60 TeV), and you aren't seeing a quantitative concern or argument that it's needed. Is this correct? If so, can you add this explanation to the wiki? (And then I'm satisfied on this point.)
    - Yes you are right. I have added this to my wiki.
15. What is the intuitive meaning of the hybrid score? If the hybrid score is >0.75, does that mean it is 75% chance (or better) of being hybrid event? Or is it really some arbitrary discriminant variable?
    - Three bdt scores ranges from 0 to 1 and the sum of three bdt scores is 1 in our case. Larger score means that it is more likely to be in this group.
    - BAC: Okay, thank you, this is helpful. But it is still not a percentage, right? So 0.75 means it is more likely to be in this group, but it doesn't quantify how much more likely? Please add this explanation, and the answer to this question, to the wiki.
    - BDT score is not percentage. For example, this is the bdt cascade score distribution from pass1 low energy cascade selection (https://wiki.icecube.wisc.edu/index.php/Multi_Year_Cascades/Low_Energy#Level_4C_.28BDT.cascade.29). An



event with cascade score 0.8 does not mean it is more like to be cascade sample than 80% of all events. I have added this explanation to wiki. And page 5 from https://cds.cern.ch/record/2684657/files/report_Chenji_Han.pdf has a good discussion about the relationship between BDT scores and probabilities.

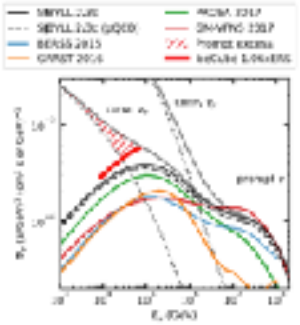
16. Why is the sample split up this way? And which sample is the signal sample which will be used in the fit? I guess it's the "cascade sample"?
    - The reason we split the neutrino sample into cascade sample and hybrid sample is that we want to split and constrain the conventional atmospheric neutrino background. The shape of hybrid sample is cascade + track. Numu charge current decay is the event with that kind of shape. Numu charge current decay produces hadrons (looks like cascade to detector) and muon (looks like track). So this sample constrains the flux numu events then constrains the atmospheric neutrino flux. And cascade sample is the signal sample because most of astrophysical neutrinos fall into this sample. But we used all three samples (cascade, hybrid, muon) in the fit.
    - BAC: This makes sense. Can you please add this explanation to the wiki?
    - I have added it.
17. Again, I don't know what this cut/coordinate means? For muons, why would you want a large cascade.score? Perhaps some plots need to be added to this section to better explain the selections.



    - According to pass1 wiki (https://wiki.icecube.wisc.edu/index.php/Multi_Year_Cascades/Low_Energy), this cut is applied because of the discrepancy in the first bin.
    - BAC: Please add this explanation/plot (in a spoiler might work).
    - I have added it.
18. It would be nice if something like this was added to the low-energy section. In fact, I suspect this answers some of the questions I had about e.g. "Delay Time". See if you can rework these sections so the explanations come in the right order, etc.
    - Sure. I have added that.
19. Link to the software which makes this cut.
    - I have added it to wiki.
20. What is the motivation for wanting to recover this event? As stated, this is non-scientific, and must be explained.
    - This was proposed at the beginning of the review process but finally we decided to not change it in the analysis. I updated the wiki to make it clear.
    - BAC: Okay, please just add one line to the wiki which says something like "we could have recovered it by changing the safety margin, but didn't" (and you could even link to the more thorough description later).
    - I have added it.
21. Which one did you go with at the end? And now that you say that, please specify which version of xgboost you used, and why you chose that version.

- Xgboost-0.47 is used for pass1 analysis while I was using xgboost-0.90 for those plots. (I updated my wiki to make it clear). The main point is that we should keep bdt used for training and classification the same version. If we use previous bdt model, we should keep the same software version. If we need to retrain the bdt, we are free to use any version we want but still need to keep the version same for training and running.
- BAC: Please add this explanation to the wiki.
- I have added it.

22. This is not quite my job, but this is an unacceptable location for permanent storage of analysis code (/data/user isn't even backed up!!). This should be moved to GitHub/SVN, and reproducibility checks ensured, before the analysis can proceed. Please discuss with WG leads and WG/technical reviewer.
    - I will upload it to the github.
    - BAC: Thank you!
    - No problem!
23. See question 20. Why should this event be rescued? This seems like the definition of cherry picking, and I don't understand why you do it. Also, can you please provide the energy and zenith of this event? A steamshovel view might also be nice.
    - See answer to question 20.
    - BAC: Thank you!
    - No problem!
24. The distribution for cscdSBU_MonopodFit4energy is clearly *very* different. This must be explained.
    - Sorry, the labels are incorrect and I updated the wiki. There are three plots. From left to right, they are astrophysical+conventional+prompt neutrino flux, conventional components only and prompt components only. In this analysis we propose to switch from berss(blue line) to prompt from sibyll2.3c. See figure 6.4 in https://journals.aps.org/prd/abstract/10.1103/PhysRevD.100.103018 (attached here). The prompt flux difference is up to one order of magnitude for different model but we have not seen prompt flux in any analysis.



    - BAC: Ohh, I see. Can you make sure the plots have proper labels? It's not clear that the three are plotting sum, conventional, and prompt in the three different panels. But I see your point. In the third panel, the unmeasured prompt component just has many different potential normalizations, all of which are compatible with data since the prompt component is yet undetected. Please add this explanation to the wiki.
    - The labels have been corrected and the explanation has been added.
25. What is meant by "lower the muon energy veto"? Is this a parameter of the software? This needs to be explained better. It seems we also really need data/MC comparison to know if the new self-veto calculation is working?

- "muon energy threshold" is a parameter in self veto software that we tune to match down going atmospheric neutrinos in Monte Carlo with our data. Yes you are right, we performed a data/mc comparison and updated the self veto study with the muon energy threshold selection result to my wiki.
- BAC: Please add the explanation of the threshold parameter to the wiki. X and Y axes need labels. Looking at the data/MC plot, I'm actually concerned… None of the muon step values appear to affect the data/MC agreement at all?
- answer later

26. Much more details are needed (imo) about how the fit is done. Can you please write down the likelihood? I would also like to see the 2D histograms for data for all three samples.
    - I updated my wiki page.
    - BAC: I don't see all three samples, or all the plots? Maybe something is wrong with the wiki display? Just to re-iterate, I think we need to see the data, the MC, and then the ratio. You can even write the data/MC values over the relevant bins (python has a way to do this, the osc group does it). Also, I don't understand the binning. Why is the entire northern sky + horizon in a single bin? And why is the final_hybrid and final_muon in a single all sky bin? I think this needs to be explained in the wiki. Just to note, Chris Weaver showed (see Fig 2.10 in his thesis) that the shape of the conventional flux is actually more peaked for nu-e. So having that finer binning might help you? I would also like to see data/MC agreement in various projections, e.g. along the zenith axis, along the energy axis, etc.
    - answer later
27. Can you specify clearly which are to be solved for, and which are nuisance? I think everything that is not the 2 astrophysical neutrino flux parameters are nuisance, right?
    - We have 5 parameters to be fitted. They are astro_norm, astro_index, conv_norm, prompt_norm and muon_norm. And the rest are nuisance parameters. See table 2 from (https://arxiv.org/pdf/2001.09520.pdf).
    - BAC: Thanks!
    - No problem!
28. Please add a Q&A section to your wiki.
    - I have added it.
29. I'm afraid I don't see any discussion of data/MC agreement. I think this needs to be added.
    - I have added it.
    - BAC: I would like to see this be done in more detail; see question 26.
    - Ok, thanks!
30. I think post-unblinding checks and plans needed to be added as well.
    - I have added it.
    - BAC: I am not thrilled about doing the beyond SPL fits post unblinding. This runs the risk of being biased, and susceptible to the analyst cherry picking certain models based on how the data looks like. I think it would be better to specify the

full range of model tests, and to run them all as blind fits. Then, before unblinding the model parameters, check that all p-values/goodness of fits make sense, and check that the systematic parameters are all within normal bounds (no significant pulls) before unblinding the model parameters. I would encourage you to bring this up with the WG and your technical reviewer.

- answer this later

31. From a high-level perspective: Have you considered testing beyond the single power law? Cascades should, in principle, give you the ability to probe e.g. broken power laws, piece-wise unfoldings, log-parabolas, etc. This would be very interesting. I think the previous analysis made such tests?
    - Yes, these will be parts of the test after the unblinding.
    - BAC: See my reply to 30.
    - Thanks!
32. I expected some plots of e.g. the expected sensitivity, and examples of performance on burn sample only. For example, what result does your fit give if run only the burn sample of the 10 years? What is the expected sensitivity on the 10 year sample, and what is the impact of the systematics on the sensitivity? Because this analysis is (essentially) an extension of a prior selection, I think it would be good for you to quantify in what ways the sensitivity improves and have those projections in the wiki before moving to unblinding request. I assume Shigeru as your WG reviewer is also thinking about this.
    - For sensitivity part, this can and will be done. I will update the sensitivity part to my wiki. For burn sample fit, we can do it to test background because currently we are only allowed to look events below 10 TeV.
    - BAC: Great, thanks! I will keep my eye out.
    - Thanks!