Masaki Ishitsuka (https://t2k.org/comm/pubboard/review/SK-T2K2023/TN460%20Review/1st%20PB%20Review%20Comments/Comments Ishitsuka)

- It is mentioned in L692 that the impact of ND fit strategy (simultaneous fit by MaCh3 and use input from BANFF by PTheta). What is the time scale of this?

Response: ~~The time scale is after the announcement at NNN, but well before the publication of the results. A conservative estimate would be in a month's time personally.~~ This analysis has been completed by ptheta over the weekend and will be in the technical note on the week starting September 25.

- As the difference of MaCh4 and PTheta was smaller before in OA2020, one possible cause is due to influence of systematic variation from SK atmospheric neutrino data. Is it possible to check the variation of the systematic errors after the fit (or in post burn-in steps) in MaCh3 and PTheta, especially for some parameters with large variations?

Response: We added a dedicated section to describe the comparison of the systematic parameters' posterior distributions from both fitters (Section 5.4). In conclusion, we have observed some differences in the systematic posterior distributions, but these differences are on a understandable level considering the different implementations of each systematic between MaCh3 and P-theta. The study of the effect of the ND fit which you mentioned above (ptheta using MaCh3 steps for T2K systematics) also studies this.

- I assume both results in Fig. 113 and Fig.117 are after Delta m^2 smearing but can you confirm? I would also like to confirm the order to apply Delta m^2 smearing in both MaCh3 and PTheta, i.e. whether the smearing is applied only Delta m^2 in the final step for plotting and set the range of allowed region, or smear Delta m^2 in fit procedure. In the later case, Delta m^2 smearing affects other oscillation parameters and optimization of systematic variations.

Response: In MaCh3 the smearing is always applied *after* the MCMC analysis. It is applied by taking a Gaussian variation of the additional uncertainty from fake-data, centered on zero with a width equal to the fake-data inflation value, for every step in the chain. This is valid since the MCMC has already sampled the regions of the smeared result. So the smearing does have a knock-on effect on other variables too, although it's quite small.

- L.15: "The Bayes factor B ... is 50.0, which indicates evidence for CP violation" is misleading. This tells that if CP is violated, sin(delta_CP) < 0 is strongly favored, while this ratio does not tells the significance of CP violation. Following sentence with "above 2 sigma" is correct as a test of CP violation.

Response: Agreed, this has been corrected and was an oversight from us. We also updated the abstract to include a statement on 2 sigma exclusion when a prior flat in sindCP is used (it is not excluded).

- Fig. 83-90: It will be easier to recognize the data if we plot them with thicker lines and in front of the posterior predictions, if possible.

Response: Will be updated asap. Sorry the time is a bit limited.

- Table 10 and L487-497: Definition of the numbers (%) should be explained in caption or in text (I assume they are relative normalization).

Response: Yes, correct. We have added this detail to both the caption and the main text, because the table comes long after the text.

- Table 10: I see Total are smaller than Total sys. for some samples (e.g. SK SubGeV-elike-0dcy). I assume this is because the variation from the oscillation parameters are small but it's not clear. I would suggest to explain which "oscillation parameters" are varied for "Total" (in caption of Table 10 or in L486).

Response: All oscillation parameters are varied here. The reason that “Total” may be less than “Total syst" is due to correlations between oscillation parameters and systematic parameters which are not accounted for when only calculating the “Total syst". Such correlations can lessen the overall uncertainty on the number of events at SK. We have added this explanation to the text.

- Table 10: "SK corr. det." in caption means "SK det."?

Response: Yes indeed, this has been corrected, thank you

- Table 10: Are the Adler angle dials included in "Cross sections"? (low momentum PID is mentioned in caption)

Response: Yes they are, and we’ve added it to the caption for clarity

- L504-506: We can read from the subscripts of $\chi^2_{data, draw}$ but better to explicitly write "χ^2 between the daw and the data ($\chi^2_{data, draw}$) " or something like that.

Response: Yes indeed, we agree and have changed this. With your suggestion it also agrees with figure 91 and the other posterior predictive p-value figures.

- L531: This is not a suggestion to update the analysis but comment on the interpretation of p-value. (1) Track length of 10 GeV muon is >40m and the acceptance to such long track is limited with ~35m inner detector size. (2) Uncertainty on nu_e flux above 10 GeV is large due to contribution from Kaon. (3) PID for >10GeV multi-ring events is much more complicated than that for ~1GeV single-ring events as hadron production and scattering are relevant for those events (actually, we see less data in e-like and more events in mu-like, which could be due to PID). I wonder if the estimated systematic uncertainties are still valid for such events. Anyway, I assume these events have no impact or have very small impact to the extraction of oscillation parameters in joint-fit.

Response: I agree with your concerns, but these events use the recommendation by the SK collaboration in the joint fit group. Our assumptions is that SK’s treatment of systematics for these is satisfactory, as outlined in the MoU between the two collaborations. Of course there is always a possibility for improvement in the systematics, and this first analysis can shed lights on where to first look. The second MoU between T2K and SK that is under construction will try to look into these sort of details more, especially the detector systematics.

- L531: Is the data statistics accounted in chi2 calculation? I assume so but would like to confirm as I see the deviation at the highest energy bin in Fig. 87 center-right and bottom two plots are up to 2 sigma levels.

Response: Yes the data statistics is included in the chi2 calculation via a Poisson likelihood in all samples: T2K ND, T2K SK, and SK atmospheric. Since the technical note does not change the likelihood calculation from the sensitivity note, we chose to not write it out in this note. If you want the full formula, please consult section 8.2 (“Likelihood Calculation”) of T2K-TN 426.

- Fig. 92 (and Fig. 93): I would suggest to change the range of horizontal and vertical axes to see the structures.

Response: We tried that when first making this plot, but it was very difficult to compare them to each other with a varying x and y-axis between samples. So we intentionally kept the same range on the plots. By having this range, we can also see which samples contribute the most to the chi2. For example, we can easily compare the shape from T2K vs SK (figure 92 vs 93). The shape is largely visible for the important samples. Therefore, we have left this plot as is. We are happy to provide you the plot in private correspondence. We also added some discussion on comparing SK and T2K samples contribution to the chi2.

- Fig.91-93: It is helpful for readers to write some explanation (interpretation) of the shapes of the scatter plots. For example, center-top plot in Fig. 93 is wider in horizontal and narrower in vertical. I understand the range of vertical variation represents the variation due to systematic uncertainty, while horizontal variation includes both systematic and statistical fluctuation of the draw. Deviation from red line in left-hand direction indicates the systematic bias between the data and nominal MC.

Response: The interpretation of each axis is not so straight forward. Your guess is pretty much correct, except the y-axis uses the data and MC, and the x-axis uses MC and MC. Basically, the method compares each draw’s expected chi2 if there was just a statistical variation of the draw, to the realised chi2 from the data. If you have a bad model, you expect the chi2 from the data vs the draw to be larger than the fluctuation (the y-axis is greater than x-axis). If you have a good model (or possibly a little too much freedom), you expect the chi2 from the fluctuation of the draw to be similar in size to the chi2 between the data and the draw. So having a p-value very close to 1 is not necessarily a good sign either. We added this discussion to the beginning of the section on posterior predictive p-values. If you are interested, the references in the text have interesting discussions on the subject.

- L625-626: Comparing Fig. 105 in TN460 and Fig. 61 in TN393, the direction is the same but the magnitude is different. It is misleading to write as "A similar result is observed".

Response: The magnitude is different because this analysis has a stronger constraint on dCP. The features are still the same. The point we wanted to make is that this behaviour has been seen before and is not new for this analysis, so is not a cause for concern. The largest difference is in a region with deltaChi2 is >10. For the region we’re trying to exclude (dcp=0, pi), the difference are truly very similar in the analyses. We have tried to highlight this in the new text.

Joe’s questions

(https://t2k.org/comm/pubboard/review/SK-T2K2023/TN460%20Review/1st%20PB%20Review%20Comments/CommentsJoeWalsh)

Physics questions

- For the impact on reactor constraint th23 octant preference, it may be nice to have side by side th13-th23 2D contour plots showing
  T2K Only, T2K+RC,
  SK Only, SK+RC,
  T2K+SK, T2K+SK+RC
  to see how this evolves with the constraint from each set of data
  Something similar for th13 and dCP may also be interesting

Response: These are on our to-do list, but we do not currently have the computational means to do this. For similar indicative results we recommend looking at p-theta’s technical note for now, and the comparisons between ptheta and Osc3++ for the SK only analysis. In MaCh3 we are improving the ND code to run faster CPU-only fits, so the SK+ND fit can be run on CPU clusters and not require GPUs (the oscillation code we have for SK atmospheric is much faster than the T2K beam oscillation code). This will be a priority moving ahead, as we too are interested in these results. The relevant content will be updated to the TN afterwards.

- L489: Is this also the sample for which the new detector matrix method has a much lower uncertainty?

Response: You mean in our analysis or in OA2023? Our analysis has similar sized uncertainties due to detector systematics to OA2020. OA2023 gets a much smaller uncertainty on this sample, yes. Can you clarify? [Figure 30 of TN456 has this comparison which shows that our new detector matrix also has a similar size of uncertainty with OA2023]

- L496/Table 10: Do we have a good idea of why the T2K uncertainties are larger in similar samples? The FHC 1Re1de sample has a 2% larger xsec uncertainty than SK SubGeV e-like-1dcy, and the detector uncertainty is roughly double. I would naively expect these to be similar. Does it come down to reaction mode composition? Or are there beam related uncertainties such as the masking of time windows in the beam structure for Michel tagging which drive this?

Response: Clarence/Dan/Junjie can double check. I think the T2K FHC 1Re1de sample statistics is very small compared to SK SubGeV e-like-1dcy sample, despite they are similar energy, the constraint from the beam sample is not great.

- How does the SK p-value compare to SK-only fit p-values?

Response: SK-only fit is on the to-do list, will have this interesting comparison when we have the fit results.

Wording

- L180: This is a bit unclear, are these Markov chains the test chains? And when you say "follows from", do you mean that the end step in one chain is used to start the next? "Shorter" than what.

Response: A bit of clarification added here. The test chains were not included in the final result. The short chains are not test chains. A short chain will pick up where the previous short chain stops because of limited wall time set by the computing clusters. The queueing time for the nodes that have longer wall time (eg. 48hrs, 72hrs) is usually much longer than nodes with less wall time (eg. 24hrs).

- L300: "comfortably excluded between 2-3 sigma", I think typically we just say "Excluded at greater than 2sigma", the 2-3 maybe frames this as "we exclude some at 3sigma" which is only true of pi in I.O., but not 0 and not in N.O. or when marginalised.

Response: Addressed

- L306 and L375: I personally am not keen on the "by-eye" estimate of the Bayes factor when we have the values computed, at least in the first case where to me it looks closer to 90/12 ~7.5 judging from the HPD, so saying 9/1 feels like it could be overselling the preference somewhat.

Response: Addressed, it was only meant to be indicative to guide the reader how these are calculated. The statement has been removed and now refers to the relevant section.

- Fig 79-82: I would change the legend to say "With R.C." rather than "With th13". "With" or "Without th13" sounds like you are doing fits with or without varying th13

Response: Addressed. Originally we thought having “RC” in the legend would risk it being jargon, but we also see your point of th13. Either way, a legend title is easy to change so it’s been done.

Minor speling/corrections

- L2: I think this should maybe say "and fit to the real data"

Response: Addressed

- L128: "the no correlation is coded" -> "no correlation is coded"/"the lack of correlations is coded"

Response: Addressed

- L575: "Publication" is misspelt

Response: Addressed

Comments from Magda:

Magda’s comments to the TN460 v.2.0 titled: ”A Markov Chain Monte Carlo oscillation data analysis using SK atmospheric and T2K beam neutrinos” -09.2023

General comment:

- This note describes a combined analysis of the five beam T2K samples and 18 atmospheric neutrino samples selected from SK-IV data using MACH3 framework.

This is a huge amount of work. Congratulations !

Response: Thank you, it’s been very rewarding!

Major comments:

- My general comment is that this TN has very limited text with the conclusion coming from the presented results. Many plots are presented but with very limited conclusions coming from them and the reader is referred to review few other TNs. In my opinion all the plots with the comparison of MACH3 and p-theta framework should have conclusions in both TNs, while TN459 was fine for reading, in this TN460 I had to come back to TN459 several times. I will give few examples below.

Response: This follows what has been requested in the past: to keep technical notes concise and focus on the results, and refer to other technical notes for work that has already been documented.

- Figure 105 - wonder why we see for the Normal Ordering (NO) almost the same values of with the highest posterior density while this is not the case for Inverted ordering?

Here there is a shift between p-theta and Mach3 results observed

Response: This is not understood yet. We tried the joint fit with several configurations on the P-theta side to understand which difference is causing this as shown in Section 5.3 Figure 115. However, none of them reproduced the shift of best-fit point in dCP for IO. A similar (but smaller) shift is observed in the T2K official analysis as well (see Figure 111), so this could be triggered by the intrinsic P-theta/MaCh3 implementation difference. We have also checked near-detector constraints, and compared the parameter values, and they could help explain the difference here.

- Figure 107 - wonder why we see the shift of the best fit values between the MACH3 and p-theta results for ? This does not depend on the neutrino mass ordering assumed.

Response: Similar response to question regarding Figure 113.

- Figure 110 - wonder why MACH3 has smaller values of posterior density in the peak for J with respect to p-theta framework? This is even more visible for the IO.

Response: Jarlskog is proportional to sindcp. MaCh3 and PTheta also show their posterior distribution of dcp with flat prior on sindcp. PTheta has slightly better constraint than MaCh3 in this dcp distribution. Thus MaCh3 is a bit flatter in Jarlskog distribution.

- Figure 111 -114 : general comments: these are very important plots because they should show our improvement in the oscillation analysis while we add 18 SK atmospheric samples to the game, and many more additional updates. The plots are too small, the problem to compare the results is clearly visible especially for the Fig.111 figures and for IO where for the highest values of the we see the jigsaw results for MACH3 framework. BTW why do we see it? Can we improve on this ?

Response:

The plotting style will be updated soon after due to the limited time. The jigsaw behaviour in the positive dcp region is due to a relatively small amount of steps. Strictly speaking, MaCh3 produces posterior probability distribution for model parameters, not delta-chi2 distributions. In order to compare to OA2020 results, we transferred the posterior probability distributions to a delta-chi2. But the jigsaw region is away from the interested region and does not influence the conclusion.

- figure 111: I don’t understand why we see the improvements with p-theta framework for joint fit while for MACH3 these results are almost the same as MACH3 with T2K only samples for ?

Response:

Two possible reasons:

1. Atmospheric samples do not contribute a lot to the sensitivity of dcp, but rather mass ordering
2. Two fitters frameworks have some differences. The implementation of joint-fit in MaCh3 is based on MaCh3 OA2020 framework and you can see good agreement between MaCh3 joint-fit sensitivity and MaCh3 OA2020 sensitivity (AsimovA) in TN426 v1.2 Fig39~41. Though there are some model updates, they are

investigated to not have significant influence on sensitivities. On the other hand, PTheta is developed in a hybrid way, part of its implementation follows OA2021, not OA2020.

- Figure 112; results for for NO - we observe huge improvements for the lower octant between joint fit and T2K only - this is seen for both fitters - and this is explained in TN459 and also here the reader is referred to review TN459, but it would be good to add here also comment that SK only data prefer lower octant while T2K only upper so when we make joint fit we see what we see here.

Response: Updated.

- Figure 113; results for for both orderings - we observe that join fit for MACH3 and previous with T2K only results are very similar in case of finding the best fit value while the p-theta results are shifted - why ? It is hard to understand. I see in the text : “Detailed discussion of this shift of Δm2 can be found in TN459” but I have checked this appendix and I don’t see why p-theta is shifted vs mach3 results.

Response:

1. In TN459 v1.1 L687~694 & Fig 33 lowest plot, PTheta tried to isolate the influence of each major framework change (in different colors) and they do explain partly the shift observed here.
2. In TN459 v1.1 Fig 34 lowest plot, if PTheta uses OA2020-like framework, it can almost reproduce OA2020 results. Hence, the shift does not mainly come from additional atm samples, but rather the fitter framework changes. As we mentioned before, MaCh3 joint-fit framework follows most of the implementation of MaCh3 OA2020 and thus the joint-fit results are closer to OA2020 than PTheta.

- In Section 5.3 Potential explanations of the results differences, there is attempt to find the source/sources of the differences we see but presented studies are not sufficient to fully explain it. We see for example that there is visible impact on these results coming from Erec binning, which pulls PTheta results towards MaCh3. (Fig. 115 and Fig 116.)

Thank you for performing these studies.

Response: We have not fully understood the reasons and will keep investigating. It is challenging to trace back which part of this complicated mechanism is responsible for this shift since it’s impossible to fully isolate each change in the framework.