Response to the referee report (DE14040)

We thank the referee for the thorough and thoughtful review of our manuscript. We appreciate the constructive feedback and helpful suggestions, which have significantly improved the clarity and completeness of the paper. Below we provide detailed responses to each comment and describe the corresponding revisions made in the manuscript.

Blue: referee comments
Orange: replies
Red: additional revision

Report of the Referee -- DE14040/Han

The article reports measurements of the neutron production from atmospheric neutrino interactions in the water or Gd-doped water SuperKamiokande detector. In particular the article focuses on comparing the result with various model's prediction, identifying those that better match the data. For most of the models it is identified the potential source of discrepancy with respect to data as well as what are the factors that results in a better agreement. The presented results are especially important for neutrino physics and responds to a need of its community. Neutrons, such as those produced by atmospheric neutrinos, are, in fact, common to nearly all neutrino physics experiments, being either a signal or a background. Having a model that correctly predicts their distribution and number is therefore crucial.

The paper is well structured and written. It starts with a description of the detector, the signal selection, and a complete list of the model tested. Results and discussion are well supported by the figures.

The abstract is well written, although I think it is missing the important information the signal selection analysis is performed using a neural network algorithm.
↪ We have revised the abstract to include this information:
"Neutron signal selection is based on a neural network trained on simulation, with its performance validated using an Am/Be neutron point source."

The conclusions well represent the findings in the paper.

I believe the paper meets the standard of PRD and should be accepted for publication provided minor changes listed below.

- I think it would be really beneficial for the reader to have a summary table of the findings of the paper. Highlighting the key strengths and weaknesses of the model tested.
↪ We have added Tables V and VI to the manuscript, summarizing key aspects of the primary neutrino interaction models and the secondary hadron-nucleus interaction models, respectively. These tables highlight their agreement with data, and include footnotes indicating specific model features that influence neutron signal multiplicity predictions.

- Page 1, introduction: [...] Accurate prediction OF detectable neutrons is essentiall [...] OF is missing
↪ This typographical error has been corrected.

- Page 7: I find the paragraph describing [Fast neutrons ..... energy] quite confusing and I would appreciate a rephrasing of it, specifically the part of the "assuming it top be near another known vertex". Near by which amount? Does this introduces a bias?
↪ Approximately 50 cm is considered “near,” corresponding to the 1σ displacement between the Cherenkov photon vertex and the neutron capture vertex. Using a fixed vertex as a proxy can introduce bias and additional uncertainty.

We have rephrased the paragraph as follows:
Since fast neutrons, γ-rays, and scattered electrons travel only a few tens of centimeters, the associated Cherenkov photons can be reliably treated as originating from a single vertex. This enables time-of-flight (ToF) corrections to suppress random PMT hits from dark noise. The vertex can either be assumed *a priori*—e.g., using the reconstructed lepton vertex—for better resolution and signal efficiency at the cost of potential bias, or reconstructed independently from the lower energy neutron signal (Appendix A), yielding lower resolution and efficiency but greater robustness to bias and kinematic uncertainties. Residual Michel electrons from muon decays are reduced using timing and energy cuts.

- Why the range is selected to be 18 to 534 micro-s in SK-IV and V and from 3 in SK-VI? How were these numbers selected? I understand that the full description is in the cited paper but a short repetition of the choice would make the current paper more clear.

↪ The upper bound of 534 μs corresponds to the end of the 535-μs event window (Sec. III C). The 18-μs lower bound in SK-IV/V was chosen in earlier studies to avoid PMT afterpulses and Michel electrons. In SK-VI, this was extended to 3 μs, thanks to improved rejection of these backgrounds through dedicated cuts and a neural network. This extension was applied only to SK-VI, as earlier MC simulations in SK-IV/V assumed the 18-μs threshold.

We have revised the manuscript accordingly:
In the candidate search stage, PMT hit times are corrected for photon ToF from an \textit{a priori} vertex given by the reconstructed lepton vertex. A 14-ns sliding time window is applied, with thresholds of 5 hits for SK-IV/V (pure water) and 7 for SK-VI (Gd-loaded), optimized for signal efficiency \cite{han_mthesis}. Overlapping candidates within 50 ns are resolved by selecting the one with the most PMT hits. In SK-IV/V, the search window was set to [18, 534] μs from the event trigger, following earlier analyses [6, 7, 47] designed to avoid PMT (<15 μs) and Michel electrons. For SK-VI, it was extended to [3, 534] μs to capture more of the faster neutron captures with Gd, aided by improved candidate classification.

- Equation 3: it is not specified the value of tau_thermal.

↪ We assume the referee is referring to the measured value of $\tau_{thermal}$.

We have rephrased the text below Equation 3:
…where the normalization constant A, the background constant B, the neutron thermalization time scale $\tau_{thermal}$ (set to 0 for SK-IV/V assuming a velocity-independent neutron capture rate), and the neutron capture time constant $\tau_{capture}$ are free parameters. The measured $\tau_{thermal}$ in SK-VI was 4.71±0.04 μs.

- Page 10 and 11: why sometimes is used NEUT 5.4.0 and others 5.1.4? Are tested neutrino events produced with 5.4.0 or 5.1.4? Please explain.

↪ The neutrino events used in the results section were generated with NEUT 5.4.0, which also served as the training dataset for the generalized additive models (GAMs) used to predict signal detection efficiency. Assuming that this efficiency is largely independent of the underlying neutrino interaction model, we evaluated the robustness of the GAMs by applying them to simulations produced with NEUT 5.1.4, which predicts approximately 10% lower true neutron multiplicity. As shown in Figure 11, the GAMs generalized well to this alternative dataset. The discrepancy between the predicted and true multiplicities was taken as a systematic uncertainty associated with the GAM performance.

We have revised the text:
To evaluate uncertainty in its performance, trained GAMs were deliberately tested on unseen simulation data produced with NEUT 5.1.4, which predicts roughly 10% lower (n,γ) multiplicity overall.

- Figure 6: different types of signal style would make it more clear.
- Figure 7 top: same comment as for figure 6.
- Figure 9: it is extremely difficult to separate the different contributions

considering the choice of colours. The various contributions should have more distinctive colours.
- Figure 12, 13, 15, 16, 18, 23: as for the preceding figures, a different line style would help.

↪ We have improved the clarity of the figures by introducing varied line styles and applying hatching to filled histograms, ensuring better visual distinction between different contributions.

- Figure 24: there is no top/bottom but left/right.

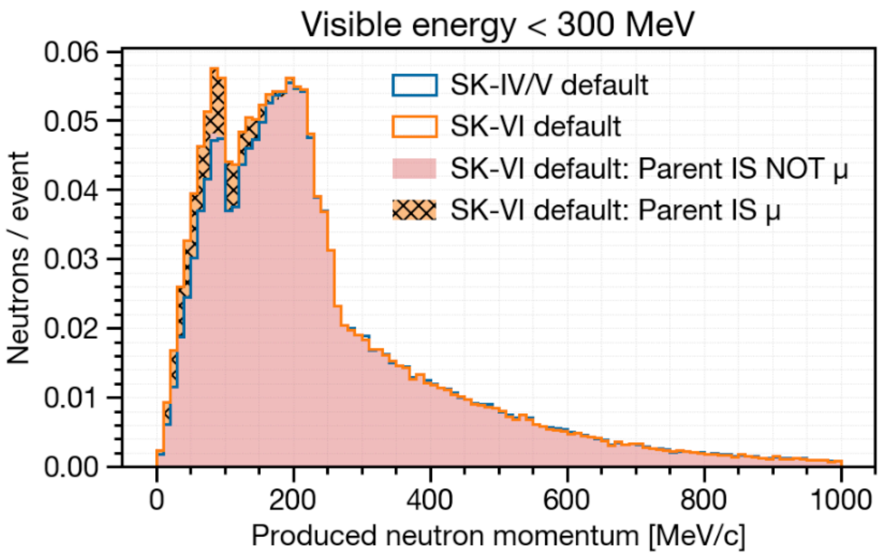
↪ This has been corrected.

- I think there is some inconsistency in the way the references and listed (full journal name vs short version, for some the title is present for others not). A cross check would be necessary.

↪ We have cross-checked and revised the references to ensure consistency. In line with Physical Review guidelines, titles were removed for journal articles and theses, and retained for technical reports.

Additional revisions

In addition to addressing the referee's comments, we identified that the 5–10% difference in predicted neutron multiplicity between the `SK-IV/V default` and `SK-VI default` simulations primarily arises from the inclusion of neutrons emitted via μ- captures on oxygen in the SK-VI model. Initially, we misattributed this discrepancy to differences in the low-energy neutron cross section libraries—ENDF/B-VI for SK-IV/V and ENDF/B-VII.1 for SK-VI—under the assumption that this was the only difference between the two setups. Upon closer examination, we found that the observed discrepancy is largely accounted for by neutrons produced through μ- captures, which are simulated in the SK-VI default but not in the SK-IV/V default configuration. This was verified by explicitly checking the μ-capture-induced neutron yield in the `SK-VI default` simulation, which closely matched the difference in the model predictions (figure attached).



To clarify this point, we have added the following clarification in Sec. VI:
The `SK-VI` setup differs from `SK-IV/V` in its use of the Geant4 NeutronHP and Bertini cascade models for neutron tracking below 20 MeV and $\mu-$ captures, respectively.

We have also revised the text in Sec. VII.C to read:
Neutrons produced from μ- captures contribute 5--10\% of the average (n,γ) multiplicity for visible energies below 300 MeV, as shown in Figure 16 (e.g., `SK-VI default`, which includes μ- capture simulation, vs. `SK-IV/V default`, which does not).