The manuscript trains four hybrid deep learning models with naïve Bayes labels to map GLOF susceptibility in the HKKH region for 2010–2020. The topic is relevant, and the effort to account for label uncertainty over such a large region is appreciated. However, I have some concerns about the overlap between the labelling indicators and the model inputs, the training/testing design, the consistency of several numbers, and the comparison with the 2025 events. In my view these issues need substantial work, so I would suggest major revision and resubmission.

Major comments:

1. It seems that the labelling indicators (Lines 98–103, Table 2) largely overlap with the model inputs (Section 2.3), and that the non-GLOF lakes are selected by low values of the same indicators (Lines 113–117), whereas GLOF-prone lakes are defined by high values (Table 2). If Table 6 is computed against the Bayesian labels (Lines 213–214), it may mainly reflect how well the networks reproduce the labelling rule; if it is computed on the GLOF inventory (Lines 108–112), the negative samples are still lakes selected for low indicator values. It would also help to clarify whether this inventory was used to set the priors (Lines 172–175). I would suggest listing the input features, evaluating against observed GLOF occurrence with representative non-GLOF lakes, and comparing with simple baselines (e.g., naïve Bayes, logistic regression, random forest).

   Response : We thank the reviewer for this valuable suggestion. In the revised manuscript, we have clarified the predictor variables used by the four hybrid deep-learning models and their relationship to the Bayesian labelling framework. We have also clarified the composition of the independent test dataset, comprising 900 documented historical GLOF lakes and 900 stratified pseudo-absence lakes, which were reserved for model evaluation referred line 152 to line 165. We acknowledge that the overlap between the labelling indicators and model inputs may influence the interpretation of model performance, and we have addressed this limitation in the revised discussion. Although comparisons with conventional classifiers, including naïve Bayes, logistic regression, and random forest, would provide additional insight, the present study focuses on evaluating the performance of the four selected hybrid architectures. We have acknowledged the absence of these baseline comparisons as a limitation and identified them as a priority for future research.

2. I found the design of the training and test sets in Lines 104–112 difficult to follow: 675 + 225 lakes are used for training, and "900 lakes with confirmed past GLOFs" for testing. If the test set contains only positive samples, AUC, PR-AUC and calibration could not be computed on it. Could the authors give the size and class composition of each set and fold, the data used for Table 6 and Figures 3–4, and the source of the 900 GLOF lakes?

3. Response : to reduce the length of paper some of important details had been missing we have added the necessary details again it's 625 + 625 lakes are used for training it's a typo error for training and second it secondly 900 lakes with confirmed past GLOFs Plus 900 pseudo-absence samples were selected from eligible glacier lakes without documented historical GLOF occurrences using stratified pseudo-absence sampling strategy base on which AUC and other performance matrix has been evaluate we have mentioned details from line – line . the 900 GLOF identified from Glacier Lake points ICIMOD is available at ICIMOD and Global Glacial Lake Inventory: 2015. Available at Zenodo.

4. More detail on Section 2.4 would help readers reproduce the method, in particular how the class-conditional parameters were estimated, how the Beta prior (Line 189) and the weights in Table 5 enter the posterior of each lake, and whether the networks used soft labels or labels binarized at 0.7 (Lines 106–107, 192). Definitions of lake volume, breach volume, excess volume and IPP, and the sensitivity analysis mentioned in Lines 335–337, would also be useful.

   Response: We thank the reviewer for this constructive comment. We agree that additional methodological detail is necessary to improve the reproducibility and transparency of the Bayesian classification framework. Accordingly, we have expanded Section 2.4 to clarify the estimation of class-conditional likelihoods, the roles of the parameter-specific prior weights and the class-level Beta prior, and the procedure used to derive the posterior probabilities for individual lakes.

   We have also clarified the distinction between continuous posterior probabilities and binary classifications based on the 0.7 threshold, specifying how the resulting labels were used to train the hybrid deep-learning models. In addition, we have expanded the definitions of lake volume, breach volume, excess volume, and IPP, including their calculation procedures, units, and supporting references where applicable. Finally, the description of the sensitivity analysis has been revised to identify the parameters and thresholds examined and explain how changes in these settings affected the resulting susceptibility classifications. These revisions improve

the methodological clarity and reproducibility of the proposed framework. Line 194 to 250

5. Some numbers appear inconsistent and may need checking: Figure 2 and the text differ in the Sentinel-2 period (2016–2022 vs. 2016–2020) and the minimum lake area (0.1 vs. 0.002 km^2); the breach volume threshold in Line 115 (0.5 million m^3) is ten times the lake volume threshold in Lines 116–117; the colour bar of Figure 7 ends at about 5.5 million m^3, below the 20 million m^3 threshold in Table 2; Table 4 implies mean depths of only 0.5–5 m; Lines 210–212 do not match Table 6; and Figure 8 totals about 250,000, versus 23,000 lakes (perhaps lake-years?).

   Response: thanks it's 2016–2020 we had corrected the typo error in image. 0.1 km^2 is used as threshold value with Bayesian framework to select the glof and non glof samples 625+625 and 0.002 km2 for screening of water bodies extracted using NDWI. The two thresholds refer to different parameters: the lake-volume threshold (0.05 million m^3) represents the estimated volume of water stored in a lake, whereas the breach-volume threshold (0.5 million m^3) refers to the breach-related volume criterion used in our framework. the colour bar of Figure 7 ends at about 5.5 million m^3 we crop the image because of it size now the full range of volume is visible and it's also justify the use of 20 million m^3 threshold for volume which not a universal threshold value but many studies in HKKH used this thread hold we had also provided literature references (Table 4 implies mean depths of only 0.5–5 m ?) I couldn't exactly understand what you ask but table 4 contain almost 250,000 lakes over the years 2010 to 2020 and we randomly presented some of them but if required we can provide the whole table as supplementary files. (Lines 210–212 do not match Table 6) corrected. Figure 8 totals about 250,000, versus 23,000 lakes (perhaps lake-years?) exactly the figure trace all 23000 lake over the year of 2010 – 2020.

6. It would be helpful to describe how the hazard escalation and bivariate scores (Figures 10–11) were calculated and how Landsat and Sentinel-2 were harmonized. Sections 3.2 (Lines 243–247) and 3.3 (Lines 280–289) seem to describe different regional patterns, which could be reconciled. Section 3.4 is qualitative, and some events (Uttarkashi; Gilgit-Baltistan melt floods; the ice-dammed Shishper lake) may not be representative of the processes modelled; some cited references also predate the events. Reporting the predicted susceptibility of the actual source lakes would make this comparison more convincing.

   Response: the description about hazard escalation and bivariate scores in (Figures 10–11) were calculated and how Landsat and Sentinel-2 were harmonized are

mentioned accordingly reference line 335 – line 350. We thank the reviewer for identifying the apparent inconsistency between Sections 3.2 and 3.3. We have revised both sections to distinguish the spatial distribution of lake area and volume, model-derived susceptibility, the bivariate hazard score, and temporal hazard escalation. Section 3.2 now clarifies that the concentration of larger lakes in parts of the northern and central HKKH does not necessarily imply the highest susceptibility across all models. Section 3.3 identifies the south-eastern Himalaya as the most consistent hotspot of very high susceptibility across the four deep-learning models and acknowledges the greater model-to-model variation in the Hindu Kush and Karakoram. We have also clarified that the bivariate hazard score combines lake volume and susceptibility, whereas the escalation score represents the change in that indicator between 2010 and 2020. These revisions provide a consistent interpretation of the regional patterns while preserving the differences represented by the individual figures. We thank the reviewer for highlighting the limitations of the original Section 3.4. We have rewritten the section to clarify that the post-2020 event review is a qualitative comparison of subsequent event locations with the spatial susceptibility patterns identified during 2010–2020, rather than a representative modelling exercise or formal statistical validation. The revised discussion examines events in the eastern and central Himalayan arc alongside localized events in the northern HKKH, while distinguishing conventional GLOFs from other glacier-related or high-mountain flood processes. The revised interpretation identifies broad geographical consistency with the study findings while explicitly acknowledging that individual source-lake susceptibility classes and changes in GLOF frequency cannot be established from this qualitative comparison alone. Citations also provided and cross checked.

Minor comments:

1. The manuscript would benefit from language editing; Lines 287–299 are somewhat repetitive. Typos: "tiggers" (Line 39), "take use of" (Line 62), "according on" (Line 301).

   Response: typos corrected

2. Susceptibility, hazard and risk are used interchangeably; consistent use would help.

   Response: We thank the reviewer for highlighting the inconsistent use of susceptibility, hazard, and risk. We have standardized these terms throughout the

manuscript according to the quantities assessed in our framework. Susceptibility now refers to the relative propensity of individual lakes to experience a GLOF, as represented by the four deep-learning models. The combined indicator is consistently termed the bivariate hazard score, calculated from normalized lake volume and model-derived susceptibility, while hazard escalation refers to its change between 2010 and 2020. We have also clarified that these indicators are not calibrated probabilities of GLOF occurrence and do not constitute a comprehensive risk assessment because exposure and vulnerability are not explicitly quantified. References to "risk" have been revised where they previously described susceptibility scores or the bivariate hazard indicator.

3. Some references seem unrelated to the statements. Please check.

   Response : references rechecked

4. Table 3: please consider giving the final hyperparameters, the CNN input, and the definition of the HNN.

   Response: Hybrid Neural Network (HNN), hyperparameters table updated accordingly.

5. Line 118: "Table 3" should probably be Table 4. Equations are not numbered; superscripts are missing in Lines 95 and 114–117.

   Response : yes it's table 4 thanks reference corrected. Equation numbering and missing superscripts checked .

6. Most figures have low resolution and small fonts (Figures 1, 2, 5–15). A basemap would help in Figures 5–11; "HKHH" should be "HKKH".\

   Response: resolution ,fonts size improved and HKHH corrected

7. Figure 4: please give the number of bins and samples per bin.

   Response: suggestion noted for figure 4.

8. Section 4: separating the Discussion from the Conclusions and comparing with previous studies would be useful.

   Response: discussion and conclusion section improved accordingly thanks.