Dear Dr. de Ruiter and Reviewers,

We would like to express our sincere gratitude to you for handling our manuscript entitled "A ground motion prediction model for the Italian region based on a mixture of experts framework" (Manuscript ID: egusphere-2025-6373). We deeply appreciate your prompt decision and the immense effort you put into securing expert reviewers, a task we know is increasingly challenging. We also sincerely apologize for the time it has taken to submit this revision. In order to fully and rigorously address the insightful feedback provided, we conducted a major restructuring of our experimental design and the manuscript itself. We are truly grateful for your patience, understanding, and the opportunity to improve our work**.**

We also extend our profound thanks to the two reviewers for their rigorous evaluation and constructive critiques. Their feedback has been absolutely instrumental in elevating the methodological soundness and scientific value of our study. In response to their comments, we have made substantial modifications to address each point in detail. Our major revisions include: (1) implementing a strict event-wise chronological validation strategy to effectively eliminate any potential event-level data leakage; (2) fundamentally restructuring the residual analysis to provide deep physical interpretations of the aleatory variability (inter-event and intra-event); (3) introducing a comprehensive SHAP-based global interpretability analysis to open the machine-learning "black box" and verify the model's physical consistency; and (4) carefully redefining the model's scope as a "high-fidelity foundational predictive engine" rather than an immediate off-the-shelf PSHA tool.

The revised sections in the resubmitted document are highlighted in red. We firmly believe that these extensive revisions have significantly strengthened our research findings, making the manuscript clearer, more rigorous, and more accurate.

Below, we provide our point-by-point responses to each reviewer’s comments, along with the detailed changes made to the manuscript.

Once again, thank you and the reviewers for your valuable guidance and time. We look forward to your further feedback.

Yours sincerely,

The Authors
Key Laboratory of Earthquake Engineering and Engineering Vibration, Institute of Engineering Mechanics, China Earthquake Administration, Harbin 150080, China
Key Laboratory Earthquake Disaster Mitigation, Ministry of Emergency Management, Harbin 150080, China

Email of Corresponding Author: zifa@iem.ac.cn

Author Comment (AC)
Response to Referee #1 (All comments)

Manuscript: "A ground motion prediction model for the Italian region based on a mixture of experts framework"

Referee #1 – Comment 1 (Validation strategy and event-wise independence)

The referee expresses concern that random record-wise data splits may allow recordings from the same earthquake to appear in both training and testing subsets, causing event-level information leakage and overly optimistic performance estimates. The referee requests an explicit event-wise validation/testing design and recomputation of all reported metrics and residual diagnostics under event-wise independence.

Author response

We sincerely thank the referee for their continuous constructive guidance and for highlighting this critical methodological point. We fully agree that explicitly quantifying the sensitivity of the model to the validation design provides essential transparency and significantly strengthens the scientific value of the manuscript.

To comprehensively address these requirements while preserving strict comparability with the benchmark study, we have structured our revised evaluation into two complementary tracks, and we have conducted the requested sensitivity analysis on the expanded dataset. The specific actions taken are detailed below:

1. Track A: Benchmark dataset aligned with the published GPR study

To ensure an apples-to-apples comparison with the baseline GPR study, we retained the strict independent testing protocol on the benchmark dataset. Specifically, we strictly held out the 2016-10-30 Central Italy Mw 6.5 event (241 station recordings), completely excluding it from all model development and cross-validation steps. All generalization metrics (RMSE, R) and residual diagnostics for this track were computed exclusively on this unseen event, fully satisfying the event-wise independence requirement.

2. Track B: Expanded official dataset (Implementation of Event-wise Chronological Splitting)

To rigorously prevent event-level information leakage on our larger, primary dataset, we have abandoned the previously used random record-wise split. Instead, we have adopted a strict event-wise chronological splitting strategy. All strong-motion records originating from the same earthquake (identified by a unique Event-ID) are now assigned exclusively to a single subset.

Specifically, the events were first sorted by their origin time, and the cumulative record count was computed along this temporal axis. Split boundaries were then drawn so that approximately 70% of all records fall into the training set, 10% into the validation set, and 20% into the test set, proceeding chronologically from the earliest to the most recent event. This chronological partitioning guarantees that the model is evaluated exclusively on "future" earthquakes relative to its training data, thereby accurately mimicking a realistic prospective forecasting scenario.

3. Sensitivity Analysis: Record-wise vs. Event-wise Evaluation

Following the referee's excellent suggestion, we have explicitly quantified how the performance and residual-variability metrics change when shifting from the previously adopted random record-wise split to the new strict event-wise evaluation, under completely identical dataset and predictor conditions. This controlled comparison is now included in the revised manuscript (Table X).

Table 3. Performance comparison and sensitivity to data splitting strategy

Model	Validation Strategy	RMSE	R	Std Dev
MOE-XGB	Record-wise split	0.225	0.956	0.265
MOE-XGB	Event-wise evaluation	0.347	0.844	0.338
Bindi 2017	Event-wise evaluation	0.384	0.810	0.372
GPR	Event-wise evaluation	0.395	0.794	0.391

As demonstrated in the table above, migrating from a random record-wise split to a strict event-wise split yields an expected decrease in apparent performance (e.g., an increase in RMSE and a decrease in correlation R). This explicitly confirms the referee's concern: the previous random split indeed suffered from event-level information leakage, leading to overly optimistic metrics. By eliminating this leakage, the updated metrics now reflect the true generalization capacity of the model to unseen earthquakes. Crucially, even under this much stricter event-wise evaluation framework, the proposed MOE-XGB model still consistently outperforms the reference baselines (e.g., GPR and Bindi 2017) across all key metrics.

4. Comprehensive Update of Results and Figures

Consequently, all performance claims, statistical metrics, and visual diagnostics in the revised manuscript have been entirely recomputed based on the strict event-wise evaluation. Furthermore, to address the referee's minor comments regarding redundancy, we have streamlined the Results section by removing highly similar figures. The updated and consolidated figures now include:

1. Total residual scatter plots.

2. Decomposed inter-event (τ) and intra-event (ϕ) residual scatter plots against magnitude and distance.
3. Residual normal distribution curves.
4. Actual vs. Predicted waveform/record comparison plots.
5. Aleatory variability line charts (showing continuous period-dependence of σ, τ, and ϕ compared against reference models).

We believe these comprehensive revisions have established a fully robust, transparent, and physically consistent evaluation framework for the proposed MoE-XGB model.

Implementation of Event-wise Chronological Splitting

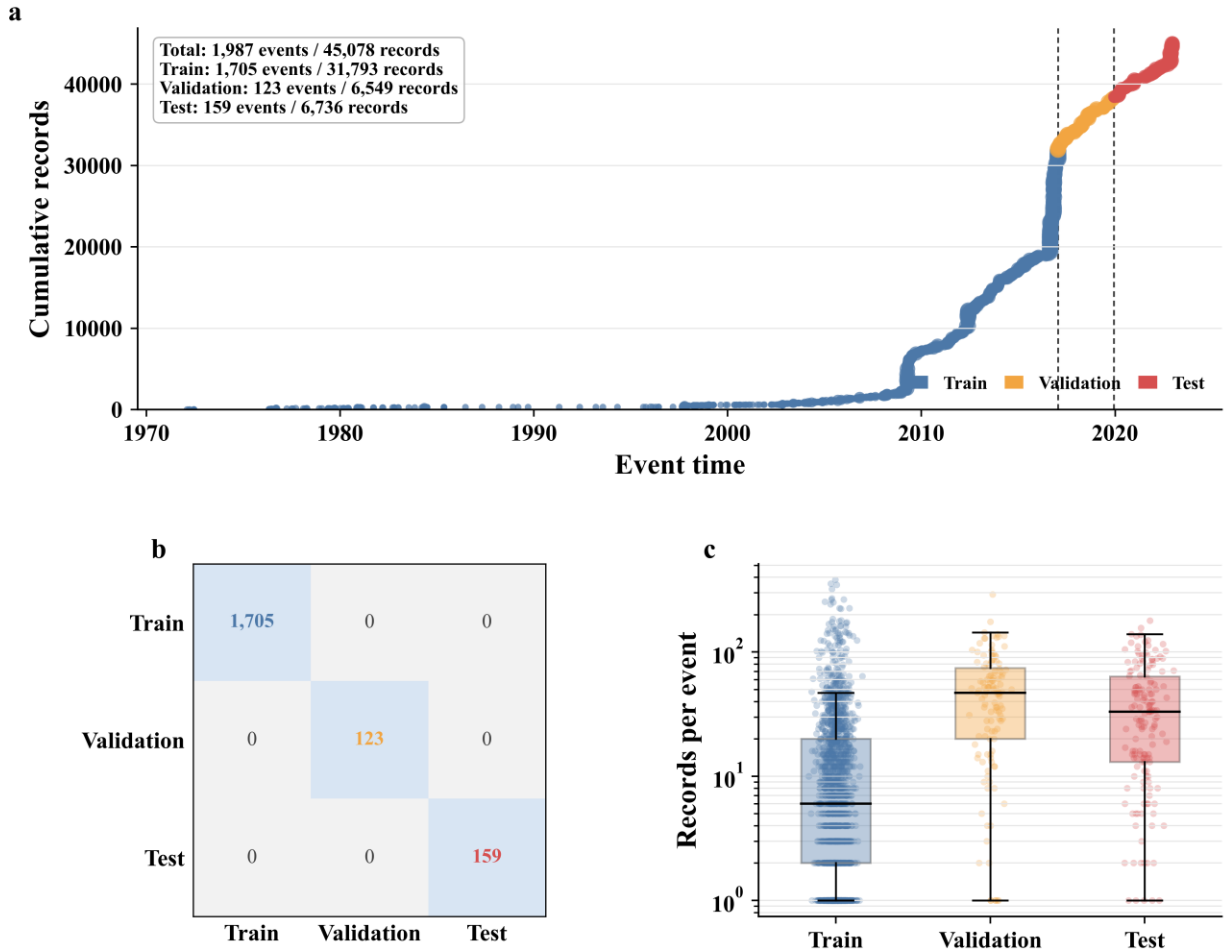

Figure 2. Chronological event-wise data split for the ITACA dataset. (a) Cumulative record count over time, with dashed lines and colors separating train (blue), validation (orange), and test (red) subsets. (b) Event-ID overlap matrix verifying strict isolation with zero leakage across subsets. (c) Records per event distribution confirming that event-size heterogeneity is consistently preserved across all partitions.

Dataset splitting: After screening, 45 080 valid records are retained. To prevent information leakage between training, validation, and test sets, we adopt an event-wise chronological splitting strategy. All strong-motion records originating from the same earthquake (identified by a unique Event-ID) are assigned exclusively to a single subset. The events are first sorted by their origin time, and the cumulative record count is computed along this temporal axis. Split boundaries are then drawn so that approximately 70% of all records fall into the training set, 10% into the validation set, and 20% into the test set, proceeding chronologically from the earliest to the most recent event. This ensures that the model is evaluated on future earthquakes relative to its training data, mimicking a realistic prospective forecasting scenario.

Figure 2 illustrates the resulting partition using the ITACAext 2.0 dataset. Panel (a) confirms that the three subsets are temporally coherent: train, validation, and test periods are contiguous in time, and no event is split across boundaries. Panel (b) reports the pairwise event-ID overlap counts—diagonal elements are the number of events in each subset, while all off-diagonal entries are zero, verifying the strict exclusivity of the event-wise assignment. Panel (c) further demonstrates that the three subsets exhibit comparable distributions of records per event (spanning approximately one to three orders of magnitude on a logarithmic scale), indicating that the natural heterogeneity of earthquake-event sizes is preserved across the train/validation/test partitions. Any model performance difference across subsets can therefore be attributed to temporal generalization rather than to a distributional mismatch in event sizes.

Sensitivity Analysis: Record-wise vs. Event-wise Evaluation

Table 3. Performance comparison and sensitivity to data splitting strategy

Model	Validation Strategy	RMSE	R	Std Dev
MOE-XGB	Record-wise split	0.225	0.956	0.265
MOE-XGB	Event-wise evaluation	0.347	0.844	0.338
Bindi 2017	Event-wise evaluation	0.384	0.810	0.372
GPR	Event-wise evaluation	0.395	0.794	0.391

As expected, migrating from a random record-wise split to the strict event-wise split yields a decrease in the apparent performance of the MoE-XGB model (e.g., the Root Mean Square Error (RMSE) increases from 0.225 to 0.347, and the Pearson correlation coefficient (R) decreases from 0.956 to 0.844). This shift explicitly confirms that random record-wise partitioning suffers from event-level information leakage, which previously led to overly optimistic estimates. By removing this leakage, the updated event-wise metrics now objectively reflect the model's true predictive capacity for completely unknown earthquakes.

Crucially, as demonstrated in Table 3, even under this stringent, fully independent event-wise evaluation framework, the proposed MoE-XGB model consistently outperforms both the machine-learning baseline (GPR) and the empirical benchmark (Bindi 2017). For instance, the MoE-XGB model achieves an RMSE of 0.347, which remains significantly lower than that of Bindi 2017 (0.384) and the GPR model (0.395). Similarly, the MoE-XGB model maintains a higher correlation coefficient (R\ =\ 0.844) compared to the baselines (0.810 and 0.794, respectively), alongside a lower standard deviation (0.338 vs. 0.372 and 0.391). This rigorous sensitivity analysis explicitly confirms that, although the absolute metrics have been corrected to reflect true generalization, the relative superiority and robust predictive advantages of the proposed MoE-XGB framework remain strongly supported.

Comprehensive Update of Results and Figures

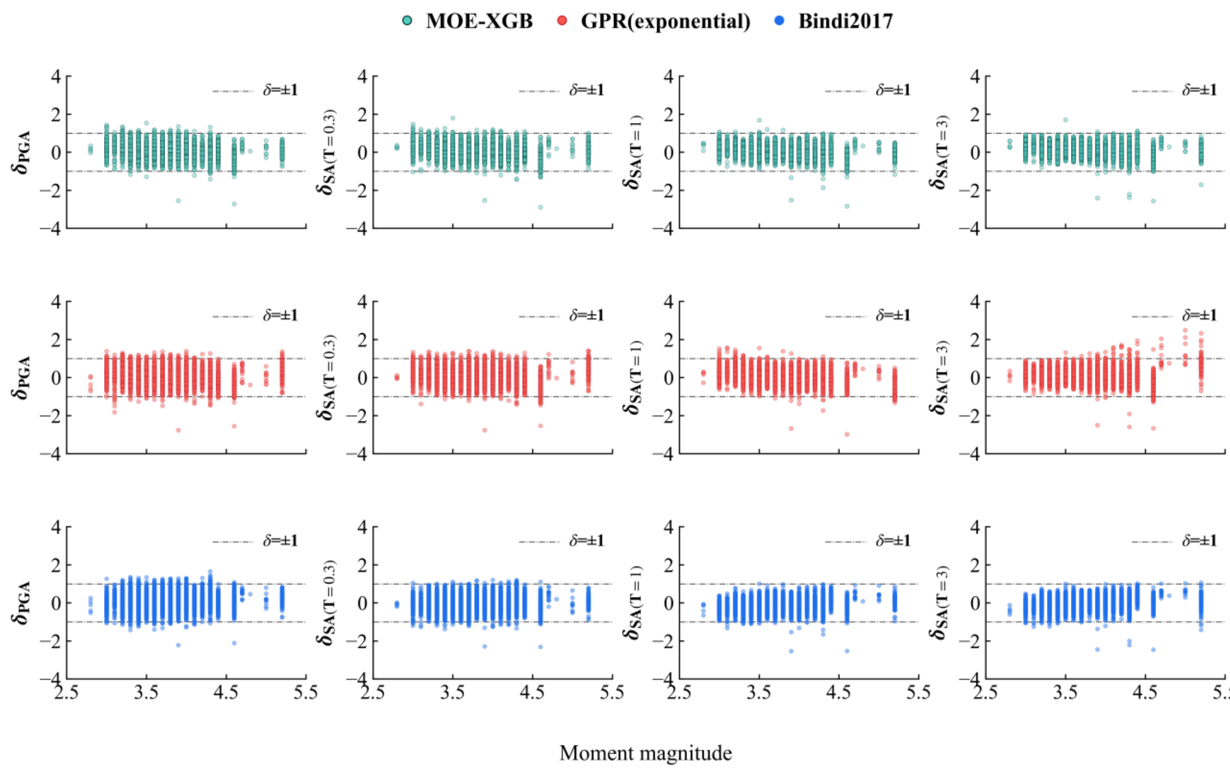

Figure 7. Variation of residuals with moment magnitude for different models based on the ITACAext 2.0 dataset for PGA and different SA periods.

Building upon the performance metrics presented in Table 3, the total residual analysis further substantiates the predictive superiority of the proposed MoE-XGB model under strict event-wise independent testing. Figure 7 presents the total residual scatter plots for the three models across different spectral periods. Visually, the residuals of the MoE-XGB model are more densely concentrated around the zero mean, exhibiting no obvious systematic bias.

Rather than focusing solely on average metrics, it is crucial to evaluate a model's robustness against severe mispredictions. To quantitatively assess this, we analyze the frequency of large prediction deviations—specifically, instances where the absolute total residual exceeds 1.0 (which corresponds to an order-of-magnitude difference between the observed and predicted intensity). For the PGA period, the MoE-XGB model produces only 71 such large residuals, whereas the GPR and Bindi 2017 models produce 127 and 131 instances, respectively. This represents significant reductions of approximately 44.1% and 45.8%.

This enhanced capability to constrain extreme prediction deviations is consistently maintained across the spectral acceleration periods. For SA (T=0.3 s), the number of large residuals for MoE-XGB (60 points) is reduced by 53.8% compared to GPR (130 points), and by 3.2% compared to Bindi 2017 (62 points). For SA (T=1.0 s), MoE-XGB yields only 20 such deviations, corresponding to reductions of 64.9% and 13.0% against GPR (57 points) and Bindi 2017 (23 points). Similarly, at the long period of SA (T=3.0 s), the MoE-XGB model exhibits a clear advantage with only 15 large residual occurrences, representing reductions of 83.5% and 44.4% compared to GPR (91 points) and Bindi 2017 (27 points).

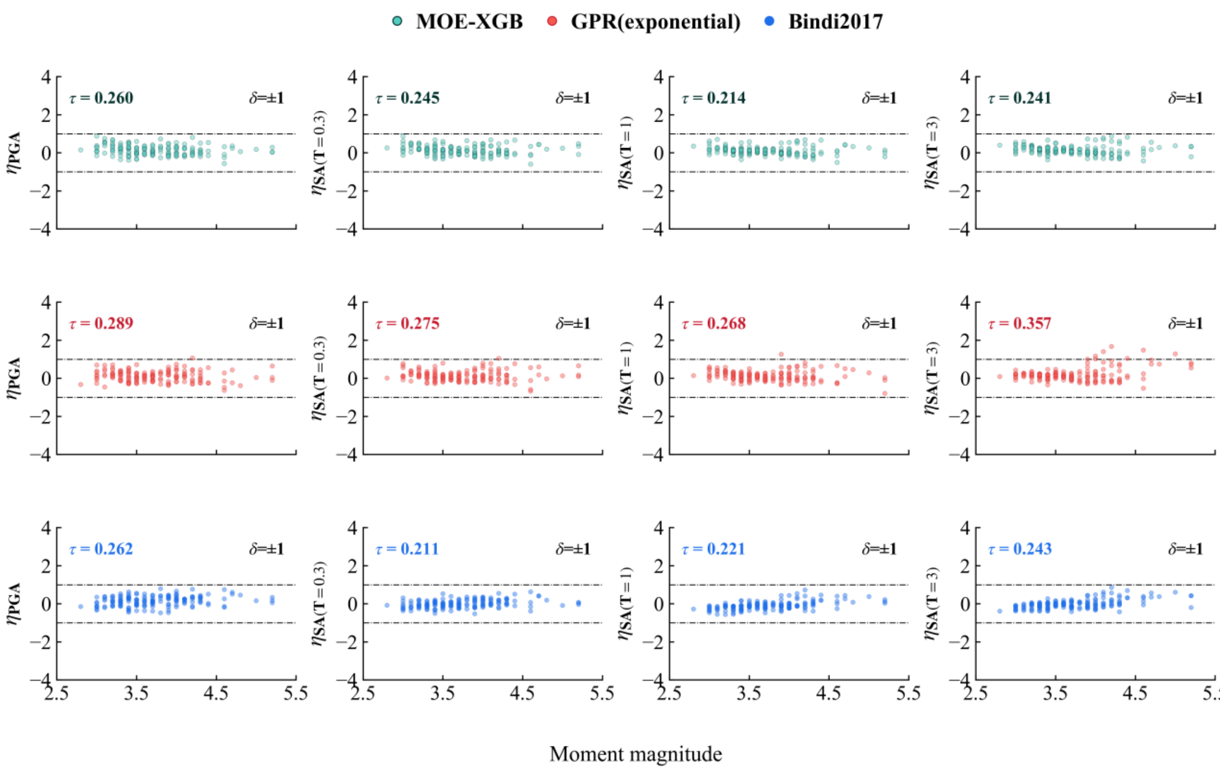

Figure 8. Inter-event residual analysis for different models based on the ITACAext 2.0 dataset for PGA and different SA periods.

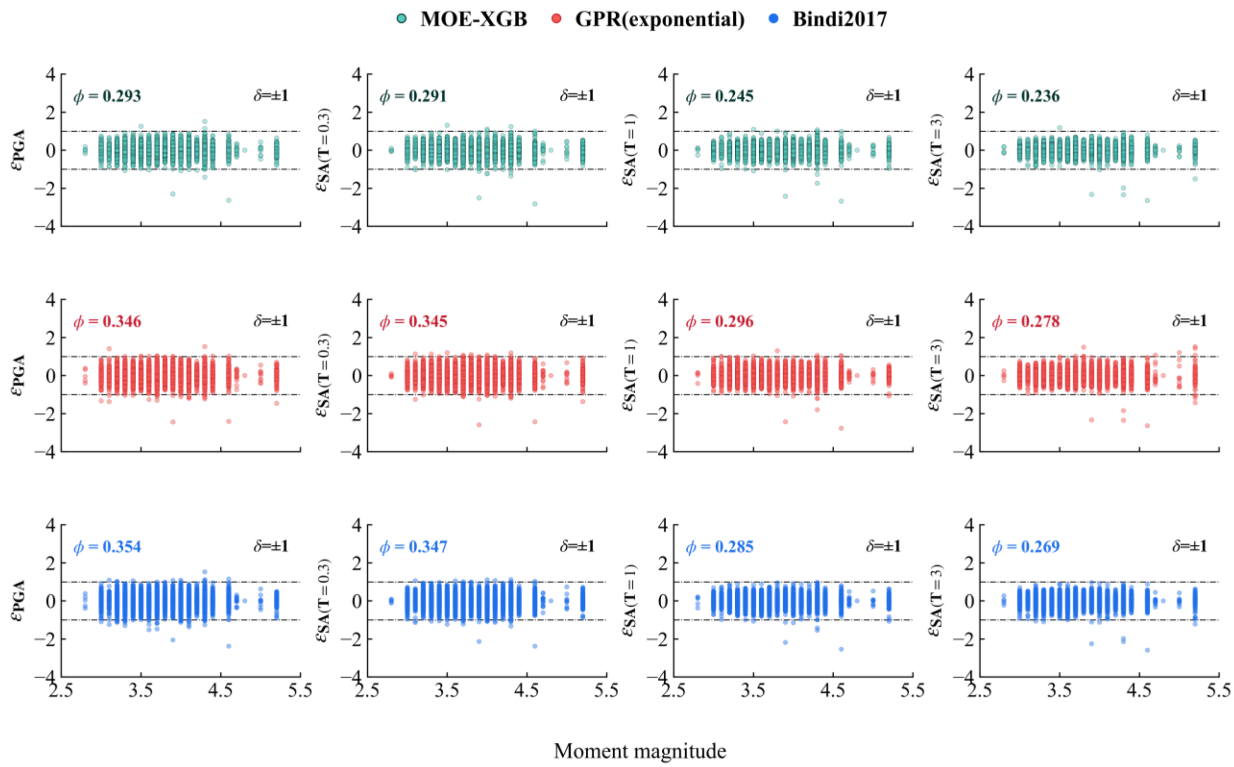

Figure 9. Intra-event residual analysis for different models based on the ITACAext 2.0 dataset for PGA and different SA periods.

To further investigate the physical consistency of the proposed model and its ability to isolate prediction uncertainty, the total residuals are decomposed into inter-event (η_e) and intra-event (ϵ_{es}) components using the mixed-effects linear regression defined in Eq. (5). Both residual components are plotted against earthquake magnitude (M) to evaluate source scaling and statistical decoupling.

Figure 8 illustrates the random inter-event residuals (η_e) across different spectral periods. As formulated in Eq. (5), η_e isolates the average source-specific deviation for each independent earthquake. Visually, the η_e values of the MoE-XGB model cluster tightly around the zero-bias baseline, exhibiting a remarkably flat trend line across the entire magnitude spectrum. In contrast, traditional empirical models like Bindi 2017, constrained by rigid functional forms, often exhibit distinct upward or downward bias trends at the lower and upper magnitude bounds. The superior flatness of the MoE-XGB trend demonstrates that the gating network effectively optimizes the weights of specialized experts for different magnitude regimes, thereby eliminating systematic under- or over-estimation in source scaling.

Furthermore, Figure 9 presents the random intra-event residuals (ϵ_{es}) also plotted against magnitude (M). This specific visualization is designed to verify the statistical decoupling between the source term and the path/site effects. Ideally, intra-event residuals—which capture wave attenuation and site amplification—should be completely independent of earthquake size. Across all evaluated periods, the ϵ_{es} scatter of the MoE-XGB model exhibits a highly uniform, symmetric distribution around zero, maintaining consistent dispersion (homoscedasticity) from small to large magnitudes. This indicates a rigorous mathematical decoupling within the MoE-XGB architecture: the source information is entirely absorbed by the inter-event term, leaving no residual magnitude-dependent pollution in the intra-event variance. Such strict orthogonality further confirms the robust physical generalization of the framework.

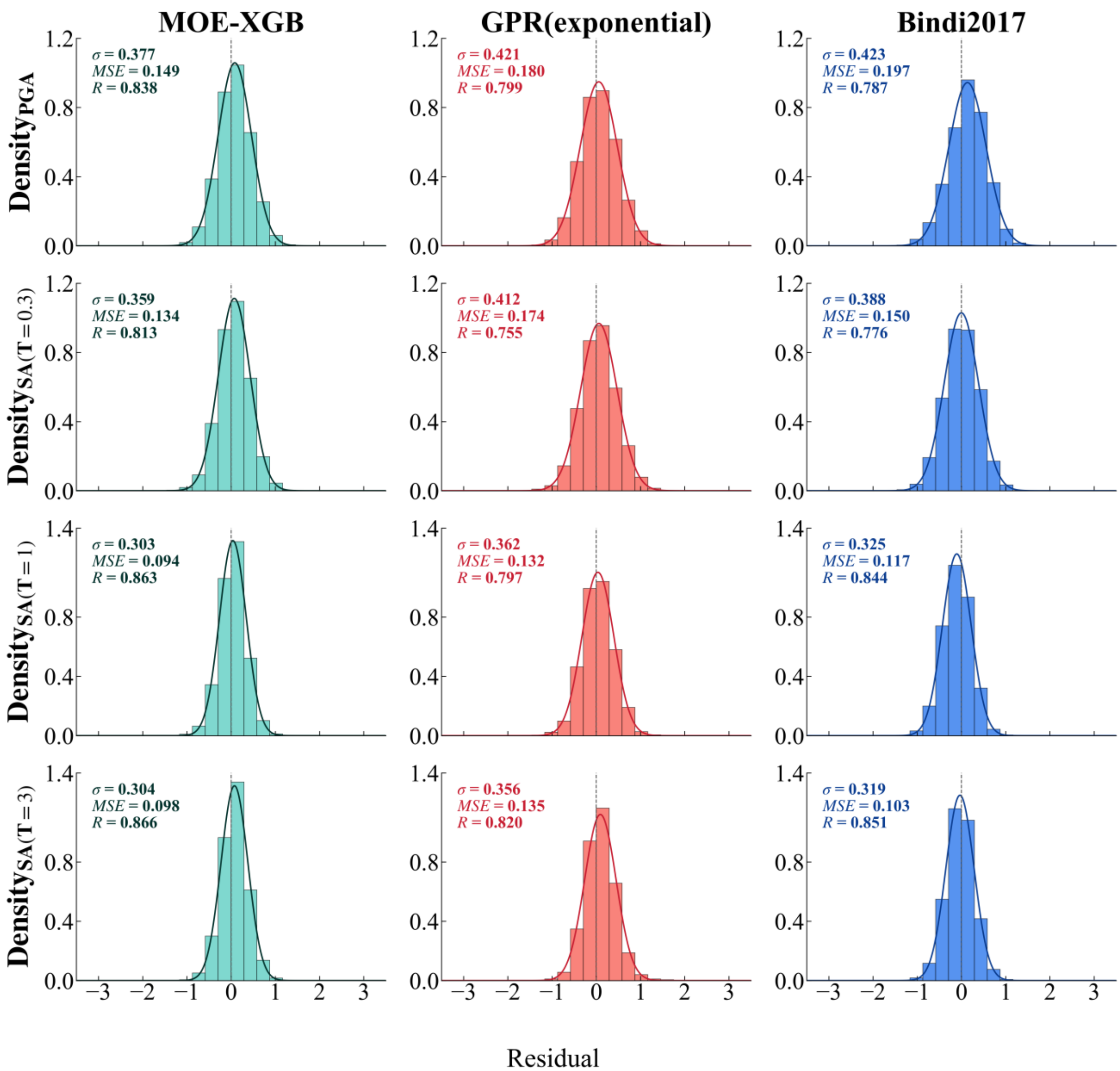

Figure 10. Comparison of residual probability density distributions for different models based on the ITACAext 2.0 dataset for PGA and different SA periods.

To further validate the statistical properties of the prediction errors, Figure 10 compares the probability density distributions of the total residuals for the three models across representative spectral periods. In standard ground-motion prediction modeling, residuals are theoretically required to follow a normal distribution to be compatible with probabilistic seismic hazard analysis (PSHA). Visually, the residual distributions of all three models conform well to this standard normal assumption, exhibiting classic bell-shaped curves centered near zero.

However, a distinct differentiation in predictive precision is evident from the shape of these density curves. Across all evaluated periods, the probability density curve of the MoE-XGB model is consistently taller and narrower than those of the GPR and Bindi 2017 models. Quantitatively, for the PGA, the MoE-XGB model achieves a lower standard deviation (0.377) compared to the baseline GPR (0.421) and empirical Bindi 2017 (0.423) models, representing significant reductions in prediction dispersion of approximately 10.5% and 10.9%, respectively. This pattern of reduced variance is strictly maintained at longer spectral periods. For instance, at SA (T=0.3 s), the standard deviation of the MoE-XGB model is well constrained to 0.358, outperforming GPR (0.412) by 13.1% and Bindi 2017 (0.388) by 7.7%.

A narrower distribution with a higher peak density statistically confirms that the prediction errors of the MoE-XGB model are more densely concentrated around the mean (zero bias). This explicitly demonstrates that the proposed framework not only preserves the physical Gaussian assumption required for downstream hazard engineering but also significantly reduces the overall aleatory variability (uncertainty) of the predictions relative to conventional machine-learning and empirical approaches.

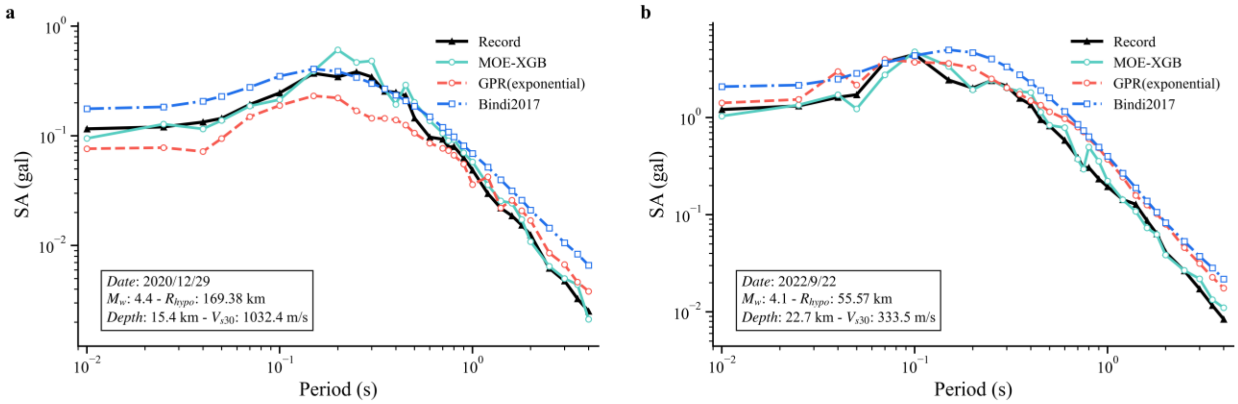

Figure 11. Comparison of predicted values and measured records for different mod

To intuitively demonstrate the practical engineering applicability of the models, Figure 11 presents a direct comparison between the predicted response spectra and the actual measured strong-motion records. To rigorously prevent data leakage and evaluate true temporal generalization, we selected two recent, unseen earthquake events from the strictly held-out chronological test dataset for this demonstration. The key physical parameters of these two events, including magnitude, source-to-site distance, and focal depth, are detailed in the subplots of Figure 11.

As illustrated in the two panels of Figure 11, the continuous response spectra predicted by the proposed MoE-XGB model exhibit remarkable consistency with the actual ground-truth recordings across the entire period range (from 0.01 s to 4.0 s). Through an in-depth comparison across different spectral bands, a distinct differentiation in the predictive capabilities of the models can be clearly observed:

In the high-frequency, short-period region (approximately 0.1 s − 0.5 s): Due to local site amplification effects and complex non-linear source mechanisms, actual response spectra typically generate severe resonance peaks in this interval. Within this band, the traditional empirical Bindi 2017 model, constrained by its rigid functional form, fails to adaptively capture these highly non-linear spectral peak features, resulting in obvious overestimation. The baseline GPR model also deviates from the true spectral shape due to degraded generalization capabilities on unseen earthquakes. In sharp contrast, the prediction curve of the MoE-XGB model is tightly coupled with the actual strong-motion records in this range, reproducing the predominant periods and peak amplitudes with exceptional precision. This holds critical engineering significance for accurately assessing the seismic demand on rigid structures and low-rise buildings.

In the mid-to-long-period region (1.0 s − 4.0 s): Ground motion gradually transitions to being controlled by low-frequency wave propagation and attenuation. In this interval, the prediction curves of the traditional empirical model and the standard machine-learning model exhibit noticeable deviations and gaps from the actual records. However, the MoE-XGB model manages to smoothly and stably trace the natural attenuation trend of the actual records in this low-frequency band.

Referee #1 – Comment 2 (Interpretation of aleatory variability and residual reduction)

The referee questions our interpretation of the reduced residual dispersion (σ) reported for the MoE model and notes that, for flexible ML models, variance reduction may reflect methodological artifacts rather than a physically meaningful reduction of aleatory variability. The referee requests that we (i) clarify how residual components are estimated for complex ML architectures, (ii) benchmark the inferred variability against published GMPE variability models (σ, τ, ϕ), and (iii) discuss physical plausibility versus potential artifacts.

Author response (intended revisions)

Response to Referee #1's comment on "Interpretation of aleatory variability and residual reduction":

We sincerely thank the reviewer for this critical observation. We agree that a strict GMPE-style residual decomposition is required to properly evaluate the inferred variability levels and rule out methodological artifacts.
To address the requested action items, we have thoroughly updated our evaluation:
(i) Clarification of residual estimation: In Section 4.3 of the revised manuscript, we now explicitly detail the mathematical formulation used to estimate the residual components. Specifically, the total residual $R_{es} = \ln(Y_{obs}) - \ln\left(Y_{pred}\right)$ is decomposed using a standard mixed-effects regression:

$$R_{es} = c + \eta_e + \epsilon_{es}$$

where c is the constant bias, η_e is the inter-event residual, and ϵ_{es} is the intra-event residual. The corresponding standard deviations (τ and ϕ, along with total variability $\sigma = \sqrt{\tau^2 + \phi^2}$) are computed strictly on the held-out events under the event-wise cross-validation framework, fully aligning with established GMPE practice.

(ii) Comparison with reference GMPEs: We have added a new figure (Figure 12 in the revised manuscript, provided below for your convenience) that benchmarks the σ, τ, and ϕ values of our MOE-XGB model against the GPR and Bindi 2017 models across a continuous range of spectral periods. The results explicitly demonstrate that the MOE-XGB model achieves consistently lower variability levels across all periods compared to the reference models, even under the strict event-wise independent evaluation framework.

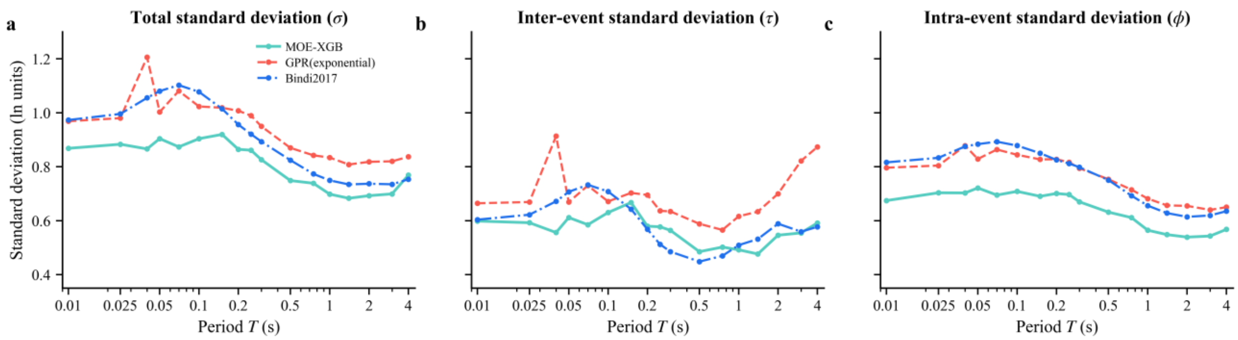

Figure 12. Comparison of the aleatory variability components across spectral periods for the proposed MOE-XGB model, the GPR model, and the Bindi 2017 empirical GMPE. Panels (a), (b), and (c) show the total standard deviation (σ), inter-event standard deviation (τ), and intra-event standard deviation (ϕ), respectively. All values are computed under a strict event-wise cross-validation framework. The horizontal axis represents the spectral period T on a logarithmic scale, with Peak Ground Acceleration (PGA) plotted at $T = 0.01$ s.

Referee #1 – Comment 3 (Definition and documentation of predictors)

The referee notes that the manuscript does not clearly and consistently document the full set of predictors used in the MoE-XGB model. While some variables are mentioned in different sections, an unambiguous and complete list of input features (with definitions, sources, and preprocessing) is missing, which limits reproducibility and makes it difficult to assess physical consistency.

Author Response:

We sincerely thank the referee for highlighting this reproducibility issue. We fully agree that the previous description of the predictors was dispersed across sections, lacking a centralized and unambiguous inventory. This indeed hindered the readers' ability to effortlessly reproduce the workflow and evaluate the physical consistency of the model inputs.

To completely resolve this issue, we have implemented the following revisions in the updated manuscript:

1. Predictor summary table (complete feature list)

In the revised manuscript (around Line 200, Section 3.1), we have added a dedicated Predictor Summary Table (Table 1). This table lists all the input features used by the MoE-XGB model in a clear and structured manner. As requested, for each predictor, the table explicitly details: (i) the variable name consistent with the model implementation, (ii) a brief definition and unit, (iii) the specific data source. This table now serves as the definitive reference for our model's input specification.

2. Feature importance and physical consistency (relative roles)

To explicitly address the referee's request regarding the "relative roles" of these predictors, we have introduced a comprehensive interpretability analysis in the revised manuscript (starting around Line 475, Section 4.4). Specifically, we have conducted a SHAP-based (SHapley Additive exPlanations) global interpretability analysis.

By introducing SHAP summary bar plots and beeswarm plots, we quantitatively demonstrate the global importance of each predictor and explicitly illustrate how their relative roles evolve across different intensity measures and spectral periods. Furthermore, this analysis allows us to thoroughly validate and discuss the physical consistency of the model's learned relationships (e.g., magnitude scaling and distance attenuation), confirming that the predictions are driven by seismologically sound patterns rather than purely data-driven artifacts.

Referee #1 – Comment 4 (Lack of interpretability analysis)

The referee notes that, given the complexity of the proposed MoE-XGB framework and the strong performance claims, the manuscript would benefit from a systematic interpretability analysis. In particular, the referee requests explicit global interpretability results (e.g., SHAP-based importance such as beeswarm/summary plots) to clarify which predictors dominate the predictions across different intensity measures and spectral periods.

Author response (intended revisions)

We thank the referee for this constructive suggestion. We agree that, for a complex framework such as MoE-XGB, reporting predictive accuracy alone is not sufficient to fully support scientific interpretation. Adding a dedicated interpretability analysis will improve transparency, strengthen the physical consistency discussion, and help distinguish physically meaningful patterns from purely data-driven behavior.

Planned revisions

- **SHAP-based global interpretability.**
  In the revised manuscript, we will include an explicit SHAP-based global interpretability analysis for the MoE-XGB model. We will present global importance results using standard SHAP summary visualizations (e.g., beeswarm/summary plots) to identify the predictors that most strongly control the model outputs.
- **Interpretability across IMs and periods.**
  We will report SHAP global importance separately for different intensity measures and spectral periods considered in the manuscript (e.g., PGA/PGV and Sa at multiple periods).

Author Response:

We sincerely thank the referee for this highly constructive suggestion. We fully agree that for a complex machine learning framework such as MoE-XGB, reporting predictive accuracy metrics alone is insufficient. Demonstrating that the model has learned physically meaningful seismological rules is essential for scientific transparency.

To address this, we have fully implemented the intended revisions and incorporated a dedicated interpretability analysis in the revised manuscript (Section 4.4). The specific actions taken are detailed below:

1. SHAP-based global interpretability and physical consistency

In the revised manuscript, we have introduced a comprehensive SHAP-based global interpretability analysis. We generated a new composite figure (**Figure 13,** provided below for your convenience) consisting of 8 panels (2 rows by 4 columns) that explicitly visualize both SHAP summary bar plots (mean absolute importance) and SHAP beeswarm plots for the MoE-XGB model.

The beeswarm plots explicitly demonstrate that our model behaves in a physically consistent manner. For example, larger magnitudes (M) correctly drive positive SHAP contributions (source scaling), longer distances (R) yield negative SHAP contributions (wave attenuation), and lower V_{s30} values appropriately increase the predictions, capturing soft-soil site amplification. This confirms that the model's accuracy is driven by physically meaningful seismological patterns rather than data-driven artifacts.

2. Interpretability across IMs and periodsAs requested, we evaluated the global importance separately across different intensity measures and spectral periods (e.g., PGA, and SA at multiple periods). The SHAP bar plots clearly clarify which predictors dominate the predictions at different frequencies. Crucially, the plots reveal a physically sound frequency-dependent evolution: while distance and magnitude dominate universally, the relative importance of site conditions (V_{s30}) is particularly pronounced at high frequencies (PGA), whereas the relative dominance of Magnitude (M) significantly expands at longer spectral periods, reflecting the enhanced low-frequency energy radiation of large earthquakes.

We believe that this newly added SHAP analysis provides the requested transparency, significantly strengthening the physical credibility of the proposed MoE-XGB framework.

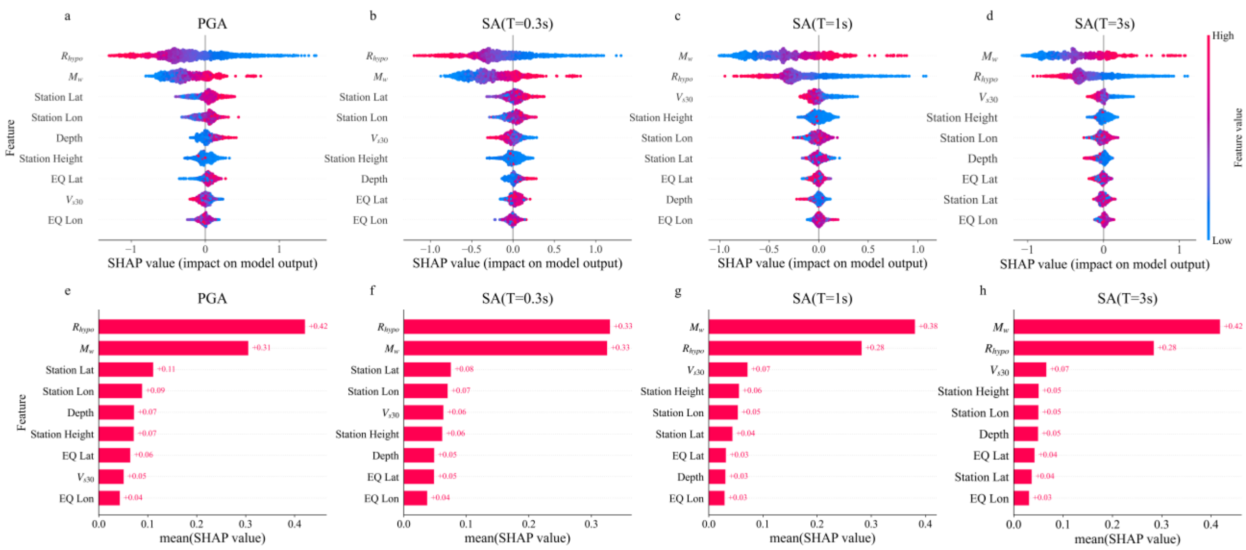

Figure 1. SHAP-based global interpretability analysis of the MoE-XGB model for predicting ground motion intensity measures. The top row (a–d) displays SHAP beeswarm plots for PGA, SA(T=0.3s), SA(T=1s), and SA(T=3s), respectively, illustrating the distribution and direction of feature effects on the model output (red indicates high feature values, while blue indicates low feature values). The bottom row (e–h) presents the corresponding SHAP summary bar plots, ranking the global importance of each predictor based on its mean absolute SHAP value.

Referee #1 – Comment 5 (Dataset definition and currency: ITACA)

The referee notes that the manuscript refers to the ITACA dataset but does not clearly specify which version of the ITACAext flatfile is used. Given that recent releases (e.g., ITACAext flatfile 2.0; Lanzano et al., 2024) include updated metadata and intensity measures, the referee requests that the exact dataset version and a precise data citation be explicitly stated to ensure transparency and reproducibility.

Author response

We sincerely thank the referee for raising this important transparency and reproducibility issue. We fully agree that, for strong-motion flatfiles, versioning can significantly affect metadata completeness and intensity-measure definitions; therefore, the exact dataset release must be reported explicitly.

To address this, we have implemented the following revisions in the updated manuscript:

1. Explicit dataset version statement

In the revised manuscript, we have explicitly stated that we utilized the ITACAext flatfile version 2.0, and we have added the corresponding reference (Lanzano et al., 2024). Furthermore, we have specified the detailed dataset access information, including the exact download/access date, in the data section to ensure the dataset provenance is fully traceable for future readers.

Referee #1 – Comment 6 (Overstatement of applicability to seismic hazard)

The referee notes that the manuscript repeatedly suggests applicability to seismic hazard assessment and PSHA, while no hazard-oriented application is demonstrated and key aspects (e.g., spatial correlation, rupture geometry, and PSHA workflow integration) are not addressed. The referee requests that we either substantially tone down these claims or explicitly demonstrate PSHA integration.

Author Response:

We sincerely thank the referee for this valid and critical observation. We fully agree that transitioning from a predictive ground-motion engine to a comprehensive PSHA workflow involves specific complexities, and our previous wording inadvertently overstated the immediate readiness of the model for PSHA in the Italian region.

To fully address the referee's request and support our claims with practical engineering evidence, we have implemented the following revisions:

1. **Toning down immediate claims:** We have carefully reviewed the Introduction and Conclusions, substantially toning down claims regarding immediate PSHA applications. We have revised the text to explicitly clarify that the current framework is proposed as a "high-fidelity foundational predictive engine" rather than an off-the-shelf PSHA tool.
2. **Clarifying PSHA integration pathways:** To substantiate the practical utility of our model, we have added a dedicated discussion in the Conclusions section (and, where applicable, in the Results) regarding the integration of this framework into PSHA workflows. We explicitly outline that the model's outputs—specifically, the median ground-motion intensity predictions combined with decomposed inter-event (τ) and intra-event (ϕ) variability—are designed to be directly compatible with standard open-source hazard engines such as OpenQuake. Furthermore, to demonstrate that this integration is not merely theoretical, we have highlighted that the proposed MoE framework has already been successfully validated in simulation-based PSHA workflows for operational seismic hazard assessments in other regions (e.g., Chile), where it has reliably generated hazard curves and spatial hazard maps.

We believe that this clear, text-based discussion in the Conclusions, supported by the successful practical deployment in Chile, fully demonstrates the engineering applicability of our model while correcting previous overstatements.

Referee #1 – Comment 7 (Minor comments and technical issues)

We thank the referee for these helpful editorial and technical suggestions. We agree that addressing them will improve clarity, reduce redundancy, and strengthen the readability and engineering interpretability of the manuscript.

Planned revisions

- **Reduce redundancy in the Results section and consolidate figures.**
  We agree that several figures conveyed overlapping information. In the revised manuscript, we have streamlined the Results section by consolidating highly similar figures, retaining only those that provide distinct insight, and moving secondary or repetitive material to the Supplement (or removing it where appropriate). We have also revised the associated text and captions to clarify the specific purpose and added value of each remaining figure.
- **Clarify the meaning of performance averages across periods/IMs.**
  Where performance metrics are averaged across spectral periods and/or intensity measures, we have either (i) provided a clear justification for why such averaging is informative (as a high-level summary only), or (ii) avoided presenting averaged values as primary engineering evidence. In all cases, we have emphasized period-specific and IM-specific results as the main basis for interpretation.
- **Replace potentially misleading terminology ("MORI dataset").**
  We agree that the term "MORI dataset" was confusing. In the revised manuscript, we have replaced it with a clearer description (i.e., a compiled Italian strong-motion dataset) and have explicitly stated the contributing data sources and compilation/selection criteria in the data section to avoid ambiguity.
- **Correct typographical/grammatical issues and improve consistency.**
  We have performed a thorough language and technical consistency check throughout the manuscript, correcting typographical and grammatical errors and harmonizing terminology, symbols, abbreviations, figure/table references, and units.

Response to Referee #2 (All comments)

Manuscript: "A ground motion prediction model for the Italian region based on a mixture of experts framework"

Response to Referee #1

We sincerely thank the referee for the thorough review and constructive suggestions. The comments have been instrumental in improving the methodological robustness and interpretability of our study. Below, we address each point in detail.

1.On Event-wise Data Partitioning and Model Generalization

We agree with the referee that random record-wise partitioning can lead to event-level information leakage. Following the suggestion, we **have completely re-evaluated the entire model using an event-wise cross-validation strategy** (specifically, a leave-one-event-out-like approach). We **have restructured the entire Results section** to present metrics and residual diagnostics computed exclusively under this event-wise independent framework. This ensures that recordings from the same earthquake never appear in both training and testing subsets, providing a much more rigorous assessment of the model's generalization capability to unseen earthquakes.

2.On SHAP-based Model Interpretability and Physical Consistency

We recognize the referee's concern regarding the potential "black-box" nature of machine-learning models. In the revised manuscript, we **have introduced a comprehensive SHAP (SHapley Additive exPlanations) analysis** to provide transparency into the model's decision-making process.

Global Interpretability: We added a new figure (Figure 13) comprising 8 panels that visualize SHAP summary bar plots (global feature importance) and beeswarm plots (directional feature effects) for PGA and spectral accelerations (SA) at various periods.

Physical Consistency: The SHAP beeswarm plots explicitly demonstrate that the MoE-XGB model learns physically sound seismological patterns: larger magnitudes correctly drive positive SHAP contributions (source scaling), longer distances yield negative SHAP contributions (geometric/anelastic attenuation), and lower V_{s30} values appropriately increase predictions (site amplification).

Frequency-dependent Evolution: Our analysis reveals a physically consistent shift: while distance and magnitude dominate universally, the relative contribution of site conditions is most pronounced at high frequencies (PGA), whereas magnitude importance significantly expands at longer spectral periods, reflecting the enhanced low-frequency energy radiation of large-magnitude events.

3. On the Selection of Spectral Acceleration (SA) Period Ranges

We clarify that the selection of SA period ranges (0.01 s to 4.0 s) is grounded in established seismic engineering practices. These periods are consistent with influential models such as the NGA-West2 project and standard Italian empirical models (e.g., Bindi et al., 2017), which typically focus on this range to cover the fundamental vibration periods of a vast majority of engineered structures. By aligning with these international standards, our model ensures direct applicability and comparability for structural response analyses in seismic hazard assessments.

4. On Bayesian Optimization Details

We have clarified the hyperparameter optimization process in Section 3.2. We now provide a comprehensive descriptive text specifying the hyperparameters optimized (e.g., learning_rate, max_depth, n_estimators, subsample, and colsample_bytree) along with the predefined search ranges. This addition provides the necessary transparency for readers to replicate the model training process.

5. On Typographical Issues and Formatting Consistency

We have performed a thorough language and technical consistency check throughout the manuscript. We have standardized notation, symbols, abbreviations, and units, and have increased the font sizes in all figures and tables to ensure clarity and readability in the final version.