Reviewer Comments

Monitoring-Informed Scenario Forecasting of Spatial Seismic Demand for Proactive Resilience in the İzmir Basin

General assessment

The manuscript presents a relevant and potentially useful monitoring-informed framework for spatial seismic-demand forecasting in the Izmir Basin. The integration of strong-motion observations, residual-based spatial updating, active-fault scenarios, uncertainty propagation, and resilience-oriented indicators is coherent, and the long-term monitoring database provides a valuable basis for the proposed application. The manuscript is generally well organized and the digital-twin-oriented interpretation is appropriately qualified. However, several methodological aspects require further clarification and, in some cases, additional validation before the main conclusions can be fully supported. In particular, the interpretation of station-level residual corrections, predictive validation across stations/events/time, consistency of source-distance definitions, treatment of uncertainty, and reproducibility of the reference and spatial models should be strengthened. The comments below are intended to improve the methodological robustness, transparency, and reproducibility of the study while preserving its overall framework and practical relevance.

Author's response: We sincerely thank the Reviewer for the careful and constructive assessment of the manuscript. The comments have helped us strengthen the methodological validation, reproducibility, terminology, and interpretation of the proposed monitoring-informed forecasting framework. We have addressed all comments point by point. In particular, we added event-centered residual sensitivity analysis, simple-baseline benchmarking, grouped-event validation, source-distance sensitivity, uncertainty-component clarification, magnitude-range sensitivity, spatial-block validation, record-count sensitivity, and a fixed-network model-version analysis. We have also expanded the methodological details and standardized the terminology throughout the manuscript. Our detailed responses are provided below.

1.Separation of event effects from station-level residual tendencies: The current updating layer is constructed by averaging event–station residuals at each station. However, the residuals may contain event-specific effects in addition to persistent site/path effects. Please clarify how event effects are prevented from being absorbed into the station-level correction field, and provide a sensitivity analysis using event-centered residuals or an equivalent mixed-effects formulation.

Author's response: Thank you for the comment. We agree that the raw event–station residuals may contain event-specific effects in addition to persistent station-related spatial tendencies. To evaluate whether such event effects materially influence the monitoring-informed correction layer, we performed an additional sensitivity analysis using event-centered residuals.

$$\bar{r}_{\text{event},j} = \frac{1}{N_j}\sum_i r_{ij}$$

$$r_{ij}^{*} = r_{ij} - \bar{r}_{\text{event},j}$$

The resulting event-centered residuals were subsequently aggregated at the station level and interpolated using the same IDW procedure adopted in the baseline analysis. The sensitivity analysis retained 36,254 of the 37,077 modelling records (97.8%), representing 2,428 multi-station earthquake

events. A total of 105 stations satisfied the minimum-record requirement for the event-centered spatial update, compared with 108 stations in the baseline analysis.

The event-centered station residual tendencies remained strongly consistent with the original station residuals (Pearson $r = 0.960$; Spearman $\rho = 0.955$). The resulting spatial residual correction fields also showed very high agreement (Pearson $r = 0.976$; Spearman $\rho = 0.964$), with an RMSE of only 0.040 log units. Spatial correlations between the original and event-centered scenario PGA fields exceeded 0.997 for all five scenarios, and the scenario hierarchy remained unchanged.

The downstream indicators were similarly stable: MPI fields showed Pearson and Spearman correlations of 0.996 and 0.995, respectively, with an 84.7% overlap among the top 10% priority grid cells. Persistence patterns remained highly consistent (Pearson $r = 0.968$), with identical persistence values at approximately 96% of prediction-grid locations. These results demonstrate that the principal spatial correction structure and scenario-priority patterns are not materially controlled by event-specific mean residual effects. The corresponding sensitivity analysis has been added to the revised manuscript.

2.Benchmark against a no-correction / simple-correction baseline: The leave-one-station-out results show only modest rank association for both IDW and Gaussian-process residual prediction. Please demonstrate whether the spatial residual layer adds predictive value relative to simpler alternatives such as no correction, a global mean residual, or a nearest-station correction.

Author's response: Thank you for the comment. We expanded the leave-one-station-out validation to include three simple reference cases: no residual correction, a leave-one-out global-mean residual correction, and a nearest-station correction.

The no-correction and global-mean baselines produced RMSE values of 0.396 and 0.386, respectively, while nearest-station prediction yielded an RMSE of 0.474. The IDW surface produced an RMSE of 0.405, whereas the Gaussian-process benchmark achieved the lowest RMSE and MAE, at 0.379 and 0.297, respectively.

These results show that the simple baselines are competitive and that the spatial residual surfaces provide only modest pointwise predictive improvement. We have therefore clarified the interpretation of the residual layer in the revised manuscript. It is not presented as a high-precision local residual predictor; rather, it provides a spatially varying regional correction layer for scenario-based forecasting. The expanded benchmark results have been incorporated into Table 4 and the corresponding discussion has been revised accordingly.

3.Event-blocked and/or temporal out-of-time validation: The stated objective is forward scenario forecasting, yet the current validation focuses mainly on spatial interpolation at omitted stations. Please add a validation that more directly tests generalization to earthquakes not represented in the calibration data, preferably through grouped-event validation and/or an out-of-time experiment.

Author's response: Thank you for the comment. We added a five-fold grouped-event validation in which all records belonging to the same earthquake were assigned exclusively to either the training or test set. The reference model and spatial residual correction field were reconstructed independently within each training fold, ensuring that the test earthquakes were not represented during model fitting.

The validation included 36,262 records from 2,974 identifiable earthquake events. When evaluated against station-level residual tendencies derived solely from the held-out events, the IDW correction reduced the mean RMSE from 0.377 to 0.337 and the mean MAE from 0.301 to 0.261, corresponding to improvements of approximately 10.4% and 13.4%, respectively. The mean Pearson and Spearman correlations between predicted and observed held-out station residual tendencies were 0.442 and 0.416.

These results show that the monitoring-derived spatial correction contains transferable information on persistent station-level residual tendencies across earthquakes not used in calibration. We have also clarified that the correction layer is intended as a regional updating representation rather than as a high-precision predictor of individual event–station residuals.

4.Consistency of the source-to-site distance definition between calibration and scenario forecasting: The reference model is calibrated using an epicentral-distance-based effective distance, whereas the scenario calculations use the shortest horizontal distance to the mapped fault trace combined with an assumed depth. Please discuss the transferability of the fitted distance term between these two distance definitions and, if possible, provide a sensitivity check or clearer limitation statement.

Author's response: Thank you for the comment. We agree that the calibration and scenario stages use related but not identical source-distance measures. The reference model is calibrated using a hypocentral-style effective distance, whereas the scenario stage uses a depth-adjusted shortest distance to the mapped fault trace:

$$R_{eff} = \sqrt{{R_{epi}}^2 + \mathrm{Depth}^2}$$

$$R_{proxy} = \sqrt{{R_{trace}}^2 + h^2}$$

We have clarified in Section 3.2 that scenario source-to-site distance is represented using the shortest horizontal distance to the mapped fault trace combined with the representative source depth, providing a consistent distance measure for comparison among the selected scenarios.

As an additional sensitivity check, the scenario calculations were repeated using representative depths of 5, 10, and 15 km. Although the absolute PGA levels varied, the spatial demand patterns remained highly stable, with Spearman correlations exceeding 0.997 relative to the baseline 10-km case for all scenarios. The Tuzla Fault also remained the dominant scenario under all three depth assumptions. We have added this sensitivity result and clarified that absolute demand levels remain conditional on the adopted source-distance representation.

5.Distinction between aleatory residual variability and uncertainty in the correction field: The Monte Carlo analysis uses the interpolated station residual standard deviation as the local dispersion term. Please clarify whether this quantity is intended to represent event-to-event aleatory variability, epistemic uncertainty in the estimated station correction, or both. These concepts should be separated conceptually, and the terminology should be revised accordingly.

Author's response: Thank you for the comment. We agree that event-to-event residual variability and uncertainty in the estimated correction field should not be treated as equivalent quantities.

The station residual standard deviation is now explicitly interpreted as event-to-event aleatory variability. The sampling uncertainty of the station-mean residual correction was additionally characterized using its standard error:

$$\mathrm{SE}(\bar{r}_i) = \frac{s_{r,i}}{\sqrt{n_i}}$$

Across the 108 spatial-update stations, the mean residual standard deviation was 0.705, whereas the mean standard error of the station-mean correction was only 0.048. After spatial interpolation, the corresponding mean values were 0.663 and 0.032, respectively.

The Gaussian-process prediction standard deviation is now described separately as a model-based spatial-surface uncertainty benchmark. In the Monte Carlo analysis, the event-to-event residual variability and GP-based spatial uncertainty are combined in quadrature to form a conservative predictive-dispersion layer; we have clarified that this combined quantity should not be interpreted as purely epistemic uncertainty.

A sensitivity check further showed that separating the aleatory component from the combined predictive-dispersion formulation does not alter the principal scenario hierarchy, with the Tuzla Fault scenario remaining dominant. The Methods and Results have been revised accordingly, and the distinction between residual variability and correction uncertainty has been clarified throughout the manuscript.

6. Magnitude-range representativeness and possible small-event dominance: The monitoring database spans approximately M 2.2‑7.6, whereas the scenario forecasts considered in the manuscript correspond to Mw 6.5‑7.0 events. Please clarify the magnitude distribution of the calibration dataset and discuss whether the fitted reference model may be disproportionately influenced by the large number of small- and moderate-magnitude events. A magnitude-distribution summary and, if feasible, a sensitivity analysis using a moderate-to-large-magnitude subset would strengthen confidence in the applicability of the reference model to the scenario range.

Author's response: Thank you for the comment. We added a magnitude-distribution summary and repeated the complete modelling workflow using moderate-to-large-event subsets.

Of the 37,077 modelling records, 17,402 (46.9%) correspond to M ≥ 4.0, 7,944 (21.4%) to M ≥ 4.5, and 3,631 (9.8%) to M ≥ 5.0. The corresponding models retained sufficient spatial coverage, with 108, 107, and 100 stations, respectively, satisfying the minimum-record requirement for residual updating.

The sensitivity analysis showed that the principal spatial interpretations are highly stable. Spatial correlations between the full-catalogue and subset-based updated PGA fields exceeded 0.98 for all five scenarios, the scenario hierarchy remained unchanged, and the Tuzla Fault remained the dominant scenario. Persistence patterns were similarly stable.

At the same time, the subset models produced approximately 31–49% higher mean PGA values, indicating that absolute demand levels are sensitive to the magnitude composition of the calibration dataset. We have therefore clarified that the relative spatial organization and scenario ranking are robust, whereas absolute PGA values should be interpreted as conditional on the adopted calibration

dataset and reference-model formulation. The magnitude-distribution and sensitivity results are now reported in the manuscript.

7. Spatial-block validation of the residual correction field: The current leave-one-station-out validation assesses interpolation performance at omitted individual stations; however, this procedure may still benefit from nearby monitoring stations and may therefore provide an optimistic representation of performance in sparsely instrumented areas. Please consider adding a spatial-block cross-validation or an equivalent analysis in which groups of neighboring stations are withheld simultaneously. Alternatively, prediction error could be evaluated as a function of distance to the nearest available monitoring station. This would provide a more realistic assessment of the spatial transferability of the residual correction field.

Author's response: Thank you for the comment. We added a spatial-block cross-validation in which the 108 spatial-update stations were divided into six coordinate-based spatial groups and each group was withheld in turn. The residual surface was reconstructed from the remaining stations using the same IDW procedure.

The spatial-block validation yielded an RMSE of 0.402 and an MAE of 0.320, which are very close to the leave-one-station-out values of 0.405 and 0.319. The overall bias decreased to 0.005. However, the Spearman correlation decreased from 0.227 under leave-one-station-out validation to 0.050 under spatial-block withholding, indicating reduced ability to reproduce local residual rankings when contiguous monitoring areas are unavailable.

As an additional diagnostic, prediction errors were examined as a function of distance to the nearest available monitoring station. Stations more than 20 km from the nearest monitoring location showed a higher RMSE of 0.447. We have therefore clarified that the residual surface is suitable as a low-bias regional correction layer, whereas local residual prediction becomes less reliable in areas with larger monitoring gaps.

8. Sensitivity of station-level residual corrections to the number of records per station: The station-level correction term is calculated as the mean residual over all usable records at each station. However, the statistical reliability of this quantity may vary substantially among stations depending on the number of available recordings. Please report the distribution of records per station and discuss how differences in sample size influence the estimated station-level residual tendencies. A sensitivity analysis using alternative minimum-record thresholds, or an uncertainty measure such as the standard error of the station mean residual, would help demonstrate the robustness of the spatial correction field.

Author's response: Thank you for the comment. We examined the distribution of usable records per station and evaluated the sensitivity of the spatial correction field to alternative minimum-record requirements.

Across the 119 stations, the number of records ranged from 1 to 724, with a median of 323. The baseline criterion of at least 10 records retained 108 stations. Increasing the threshold to 20, 30, and 50 records retained 105, 102, and 100 stations, respectively, but produced an unchanged spatial residual correction field. Even under a substantially more restrictive 100-record criterion, retaining

93 stations, the resulting field remained highly consistent with the baseline representation (Pearson r = 0.990; Spearman ρ = 0.981; RMSE = 0.025 log units).

We also quantified the sampling uncertainty of each station-mean residual using:

$$\mathrm{SE}(\bar{r}_i) = \frac{s_{r,i}}{\sqrt{n_i}}$$

As expected, the standard error decreased strongly with increasing record count (Spearman ρ = −0.902). These results indicate that the adopted 10-record minimum prevents very poorly sampled stations from entering the spatial update while the resulting correction field is robust to substantially more restrictive station-selection criteria. The corresponding clarification and sensitivity results have been added to Sections 3.1 and 4.1.

9. Separation of archive growth from monitoring-network expansion in the model-version analysis: The cumulative model-version experiment shows increasing stability as the archive expands from 1977‑2010 to 1977‑2025. However, the number of stations and spatial-update stations also increases substantially between model versions. Please clarify to what extent the observed stabilization reflects accumulation of additional recordings versus expansion of the monitoring network itself. If feasible, a supplementary analysis using a common set of stations across model versions would help distinguish temporal learning from changes in spatial monitoring coverage.

Author's response: Thank you for the comment. We agree that the original cumulative model-version experiment simultaneously reflects growth of the monitoring archive and expansion of station coverage.

To separate these effects, we performed an additional fixed-network sensitivity analysis using the same 19 stations across all four model versions. These stations satisfied the minimum-record requirement in the earliest 1977–2010 version and remained available throughout the subsequent periods. Both the reference model and spatial residual correction were reconstructed using only this common station set.

The late-stage stabilization remained clearly evident. For the 2020–2025 transition, the fixed-network analysis yielded a residual-field delta RMSE of 0.017, an MPI delta RMSE of 0.019, an MPI rank Spearman correlation of 0.995, and a top-10% MPI overlap of 0.840. These values are consistent with the strong stabilization observed in the original expanding-network analysis.

The results therefore indicate that the convergence of the regional demand representation is not solely a consequence of monitoring-network expansion. Accumulation of additional observations within a fixed monitoring network also contributes substantially to the late-stage stabilization. We have added this sensitivity analysis to the revised Methods, Results, and Discussion.

10.Explicit form of the reference ground-motion model: Please provide the explicit functional form of the reference ground-motion model.

Author's response: Thank you for the comment. We have now provided the explicit functional form and fitted coefficients of the reference ground-motion representation. The model is expressed as:

$$\ln(PGA_H) = \beta_0 + \beta_1 M + \beta_2 \ln() + \beta_3 R_{eff} + \beta_4 \ln(\frac{VS_{30}}{760})$$

with fitted coefficients:

$$\ln(PGA_H) = -1.6648 + 1.6252M - 1.2441\ln() - 0.00427R_{eff} - 0.4474\ln(\frac{VS_{30}}{760})$$

The model was fitted using Huber regression after standardization of the predictors, and the coefficients reported in the manuscript have been back-transformed to the original predictor scales. The corresponding methodological description has been added to Section 3.1 to improve reproducibility.

11.Difference between the initial and usable number of records: Please clarify the difference between the initial number of records and the final number of usable records.

Author's response: Thank you for the comment. The original monitoring database contains 37,130 event–station records. During data-quality screening, 53 records were excluded from reference-model fitting: 33 lacked a usable horizontal PGA component, 19 contained a non-positive horizontal PGA component, and one lacked magnitude information. This resulted in the final modelling dataset of 37,077 usable records.

We have added this clarification to the dataset description in Section 2.

12.Minimum-record criterion for station residuals: The minimum-record criterion used for station-level residual updating should be specified.

Author's response: Thank you for the comment. The baseline spatial residual update required a minimum of 10 usable event records per station. This criterion retained 108 of the 119 monitoring stations.

We have now stated this threshold explicitly in Sections 3.1 and 4.1. In addition, as part of the revised robustness analysis, we evaluated more restrictive minimum-record thresholds of 20, 30, 50, and 100 records per station. The resulting spatial correction field remained essentially unchanged for thresholds up to 50 records and remained highly consistent even at the 100-record threshold, supporting the robustness of the adopted 10-record baseline criterion.

13.Details of the IDW interpolation: Please provide additional details on the spatial interpolation procedure.

Author's response: Thank you for the comment. We have expanded the description of the spatial interpolation procedure in Section 3.1. IDW interpolation was performed in UTM Zone 35N (EPSG:32635) using an inverse-distance power of 2 and the eight nearest available stations, without a fixed search radius. The prediction domain was represented by a regular 2-km grid covering 26.55–27.45°E and 38.05–38.80°N. The same interpolation settings were applied to the residual correction, residual-variability, and station-informed VS30 fields.

These implementation details have now been added to the revised manuscript to improve reproducibility.

14.Gaussian-process benchmark and the choice of IDW: The Gaussian-process benchmark should be described in greater detail, and the reason for retaining IDW should be clarified.

Author's response: Thank you for the comment. We have expanded the description of the Gaussian-process benchmark in Section 3.1. The GP was fitted in projected UTM coordinates scaled to kilometres using a constant × radial-basis-function kernel with an additional white-noise term. The initial RBF correlation length was 15 km, with bounds of 1–200 km, and the kernel hyperparameters were optimized using three optimizer restarts with normalized response values. The full-data fit yielded an optimized correlation length of approximately 33.3 km.

We have also clarified the rationale for retaining IDW as the primary updating method. The GP benchmark produced slightly lower cross-validated RMSE and MAE than IDW, and we do not interpret IDW as the statistically superior local predictor. Instead, IDW was retained because it provides a simple, deterministic, transparent, and readily regenerable interpolation operator that can be applied consistently to the residual, residual-variability, VS30, and cumulative model-version fields. The GP is therefore used as an independent covariance-based benchmark and uncertainty reference.

The manuscript has been revised accordingly.

15.Normalization used in MPI: Please clarify the normalization procedure used in the monitoring-priority indicator.

Author's response: Thank you for the comment. We have clarified the normalization procedure used in the Monitoring Priority Indicator. The mean updated logarithmic demand across the five scenarios and the residual-variability field are independently transformed to the interval [0,1] using min–max normalization over the prediction grid before applying the respective weights of 0.60 and 0.40.

$$\tilde{y}_{\mathrm{upd}}(x) = \frac{\bar{y}_{\mathrm{upd}}(x) - \min(\bar{y}_{\mathrm{upd}})}{\max(\bar{y}_{\mathrm{upd}}) - \min(\bar{y}_{\mathrm{upd}})}$$

$$\tilde{\sigma}_s(x) = \frac{\sigma_s(x) - \min(\sigma_s)}{\max(\sigma_s) - \min(\sigma_s)}$$

$$\mathrm{MPI}(x) = 0.60 \times \tilde{y}_{\mathrm{upd}}(x) + 0.40 \times \tilde{\sigma}_s(x)$$

This clarification has been added to Section 3.3.

16. 'Monitoring / inspection priority' terminology: The terminology 'monitoring / inspection priority' could be clarified because actual inspection priority would normally require exposure and vulnerability information.

Author's response: Thank you for the comment. We agree that direct inspection prioritization would require information beyond seismic demand, including exposure, vulnerability, asset criticality, and consequences.

We have therefore standardized the terminology throughout the manuscript and now refer to MPI as the Monitoring Priority Indicator. We have also clarified that MPI is intended as a demand-side

screening indicator for identifying locations where intensified monitoring or more detailed site-specific assessment may be warranted, rather than as a complete inspection-priority or risk metric.

17.Connectivity definition in the connected-component analysis: Please clarify the connectivity definition used in the connected-component analysis.

Author's response: Thank you for the comment. The connected-component analysis uses four-neighbour grid connectivity. Accordingly, high-demand cells are considered connected only when they share a horizontal or vertical grid edge; diagonal adjacency is not included.

This definition has now been added to Section 3.4 to make the spatial concentration analysis fully reproducible.

18.Interpretation of mean persistence: Please clarify the interpretation of mean persistence because the threshold is based on the pooled 80th percentile.

Author's response: Thank you for the comment. We agree that because the common high-demand threshold is defined as the pooled 80th percentile of all scenario-grid PGA values, the domain-averaged persistence is approximately 0.20 by construction and should not be interpreted as an independent empirical finding.

We have clarified this point in Sections 3.3 and 4.3. The revised manuscript now emphasizes the spatial distribution and maximum persistence, which provide the informative measures of repeated high-demand behaviour across scenarios. In particular, the maximum persistence of 0.80 indicates locations that exceed the common threshold in four of the five scenarios.

19.Rationale for scenario magnitudes: The rationale for the assigned scenario magnitudes should be stated more explicitly.

Author's response: Thank you for the comment. We have clarified the basis of the assigned scenario magnitudes. The adopted Mw values are representative scenario magnitudes informed by mapped fault or representative segment lengths, fault mechanisms, and empirical magnitude–rupture-length scaling relationships, particularly Wells and Coppersmith (1994).

For example, the approximately 20-km İzmir Fault rupture adopted in previous İzmir scenario studies is consistent with a magnitude of about Mw 6.5, while the approximately 50-km Tuzla Fault yields a strike-slip rupture magnitude close to Mw 7.0. The adopted magnitudes for the Seferihisar, Gülbahçe, and Yenifoça scenarios are similarly consistent with representative mapped rupture lengths and corresponding fault mechanisms.

We have also clarified that these values are comparative scenario assumptions rather than unique estimates of the maximum credible magnitude of each fault. Accordingly, the relative scenario hierarchy is interpreted as conditional on the adopted magnitude and source-geometry assumptions.

20.Monte Carlo spatial dependence: Please clarify whether the Monte Carlo perturbations are generated independently at individual grid points or whether spatial covariance is incorporated.

Author's response: Thank you for the comment. In the present Monte Carlo implementation, stochastic perturbations are generated independently at individual grid locations conditional on the local predictive-dispersion value. No additional spatial covariance structure is imposed between grid cells.

We have clarified this assumption in Section 3.4 and now explicitly state that the resulting exceedance probabilities should be interpreted as pointwise uncertainty measures rather than as realizations of a fully spatially correlated ground-motion random field. We have also identified spatially correlated residual simulation as a future extension for applications requiring joint regional exceedance or portfolio-level interpretation.

21.Consistent use of 'digital-twin-oriented': Please maintain consistent use of the term 'digital-twin-oriented' throughout the manuscript.

Author's response: Thank you for the comment. We have reviewed the manuscript and standardized the terminology throughout to consistently use "digital-twin-oriented" when referring to the proposed framework, updating process, and regional demand representation.

We have also retained and reinforced the explicit statement in Section 5.4 that the present implementation should be interpreted as a digital-twin-oriented regional seismic-demand forecasting framework rather than as a fully operational real-time digital twin. This distinction is now applied consistently in the Abstract, Methods, Results, Discussion, Conclusions, tables, and figure captions.

22.Figure readability: Figures 3–7 could be improved slightly for readability.

Author's response: Thank you for the comment. Figures 3–7 have been revised to improve readability. Font sizes, axis and coordinate labels, color-bar labels and tick marks, panel identifiers, and relevant map symbols have been enlarged and standardized. The figures have also been regenerated at higher resolution to improve readability in both electronic and printed formats.

23.Terminology and language check: Please perform a final terminology and language check. Expressions such as 'dissolved scenario sources' may be confusing.

Author's response: Thank you for the comment. We performed a final language and terminology review throughout the manuscript.

In particular, the expression "dissolved scenario sources" has been replaced with "selected scenario sources," "representative scenario sources," or "mapped scenario fault traces," depending on the context. We also standardized the terminology used for the Monitoring Priority Indicator, residual correction field, residual-variability field, scenario sources, and digital-twin-oriented framework.

Minor grammatical and formatting inconsistencies were corrected throughout the manuscript without altering the technical content or results.