# Referee Report

**Manuscript:** ***"Trends in geo-hydrological risk to the population of Italy."***

**Manuscript identifier:** EGUsphere preprint 2026-4342

**Recommendation: Major revision**

## General assessment

This manuscript examines an important question using a valuable, long-term national catalogue of flood and landslide fatalities. Its key strengths include distinguishing fatalities from fatal days, analyzing floods and landslides separately, and comparing alternative temporal models. The evidence of an early decline in recorded landslide-related mortality merits further investigation. However, several conclusions extend beyond what the analyses support. In particular, the socioeconomic explanation for the historical decline, the interpretation of policy effects, and the claim that fatal days are more predictable need stronger evidence or more cautious wording. Important numerical results and methodological details are also missing. I recommend major revision to address the following concerns.

## Major comments

1. **Distinguish observed mortality from hazard occurrence and population risk**

   *Sections 1, 5.1, 6, and 7*

   The principal outcomes are annual fatality counts and days with at least one fatal event. These are useful indicators of realized human consequences, but neither directly measures the frequency of hazardous conditions independently of exposure and vulnerability. A flood or landslide can occur without producing a fatal day, and improvements in protection could reduce fatal days even if hazard frequency remained unchanged.

   Please revise statements suggesting that fatal days isolate the occurrence of dangerous conditions. Similarly, clarify whether "risk" refers to absolute national mortality, individual mortality risk, or expected future losses. Although Table 1 includes population-normalized mortality, the principal models concern counts. A supplementary analysis using population as an offset would help distinguish national fatality burden from population-normalized mortality, while acknowledging that national population is an imperfect measure of exposed population.

2. **Provide stronger evidence for catalogue homogeneity and clarify the event-selection rules**

   *Section 3, lines 204–219*

   The claim that the 1950–2024 catalogue is sufficiently homogeneous is central to the analysis but needs supporting evidence. Please describe changes in information sources, reporting practices, verification procedures, and coverage over this period. A comparison with EM-DAT establishes broader coverage but does not establish temporal consistency within the Italian catalog.

   Excluding Vajont and Stava has a defensible process-based rationale, but these events account for a substantial share of recorded deaths. Please specify consistent inclusion criteria for engineering-related and other non-rainfall-triggered events, and present sensitivity analyses with and without the two exclusions. Testing sensitivity to other influential years would also help establish the robustness of the inferred transition.

3. **Report the numerical evidence underlying the predictor and predictability conclusions**

   *Sections 5.2, 5.4, and 5.5*

   Several central findings are presented almost entirely in narrative form. The manuscript provides no detailed tables of predictor-model results, policy coefficients, or forecast scores, making it difficult to evaluate the conclusions.

   Provide tables reporting sample sizes, model specifications, effective degrees of freedom, dispersion estimates, deviance explained, AIC differences, and relevant uncertainty estimates. For nonlinear predictor effects, include partial effect plots with uncertainty bands and the distribution of observed predictor values. Policy results and predictability metrics could be presented in supplementary tables.

4. **The time-adjusted models cannot establish the proposed socioeconomic explanation**

   *Sections 4.4, 5.4, and 6, particularly lines 428–450*

   The finding that associations between GDP and population and calendar time weaken after adjustment for calendar time does not establish that socioeconomic development caused the historical decline. The time smooth can absorb variation shared with these predictors, leaving little information to separate their effects. The official mgcv documentation specifically identifies this problem as “concurvity” in models combining smooth time trends with time-varying covariates.

   Report concurvity diagnostics and sensitivity to the flexibility of the time smooth. Distinguish the question of explaining long-term change from explaining residual interannual variation. Unless you provide additional evidence, present post-war urbanization, improved housing, and infrastructure as plausible explanations rather than demonstrated mechanisms.

5. **Limit the policy conclusions to what the policy-timing design can test**

   *Section 4.4, lines 290–298; Section 5.4, lines 391–398*

   Binary windows around policy dates test temporal alignment, but do not directly measure implementation, enforcement, investment, geographic coverage, or delayed and cumulative effects.

Symmetric windows also include years before policy adoption and therefore require a different interpretation from post-adoption windows.

Please explain the rationale for the selected windows, distinguish adoption from implementation dates, and report how overlapping policies were handled. Clarify the permutation procedure and how it preserves relevant temporal structure. With 378 policy-window models, multiple testing also needs explicit consideration.

The supported conclusion appears to be that the selected policy-date indicators add little explanatory value beyond the time trend. This should not be interpreted as evidence that risk-management policies were ineffective. Policies introduced after the initial decline could still have helped keep mortality lower.

6. **The comparison of raw predictive surprise does not establish greater temporal predictability**

*Sections 4.5 and 5.5*

Fatal days have a narrower distribution than fatality counts. Lower negative log predictive probabilities may therefore reflect lower marginal uncertainty, even if a model extracts no additional temporal information. Proper scoring rules evaluate predictive distributions, but interpreting differences across distinct response variables requires attention to their underlying uncertainty.

Do compare each model against an appropriate benchmark for the same outcome, such as a stationary or recent-window negative binomial model. Report the improvement over that benchmark before comparing forecast skill across outcomes.

The information-theoretic analysis also needs the number of quantile states, how ties are handled, bias corrections, the bootstrap procedure, and sensitivity to discretization. Given the short series and confidence intervals commonly including zero, claims about intrinsic predictability should remain cautious.

7. **Clarify model specification, forecast origins, and uncertainty**

*Section 4.3 and Figures 4–5*

For GAMs, report the spline basis, basis dimension, smoothing-selection method, model diagnostics, and method used to propagate uncertainty into predictions. For the Bayesian model, provide priors, initial-state assumptions, MCMC settings, convergence diagnostics, and posterior predictive checks.

Please clarify whether projections to 2030 remain conditional on data through 2020 or incorporate observations from 2021–2024. This distinction affects both forecast horizons and how you interpret recent fluctuations.

The reported increase in DLM projections also needs reconciliation with the stated drift-free random walk on the log mean. Explain whether the plotted central estimate is a mean or median and whether an increase on the count scale arises from widening latent uncertainty rather than an estimated upward trend.

Distinguish uncertainty in expected counts from uncertainty in future observations consistently across both modeling frameworks. Figure 5 should identify exactly which DLM distribution is displayed.

8. **Strengthen validation and qualify claims of stability**

*Sections 5.1 and 5.3*

Four annual observations within prediction intervals provide limited evidence of forecast performance, especially when intervals are wide. Please report quantitative scores, interval coverage, and interval widths over a longer rolling evaluation period, with benchmark comparisons. Clarify whether the 2025 observations were withheld from every stage of model selection.

Similarly, non-significant trends after 1975 do not establish that the series are stable. Please report the range of trends consistent with the uncertainty intervals and use wording such as "no statistically detectable monotonic trend."

For change-point analysis, report significance levels and assess uncertainty in breakpoint timing and sensitivity to influential years. Hurst estimates from short, nonstationary series should be interpreted cautiously because a structural shift can resemble persistence.

**Specific corrections and presentation comments**

- **Lines 224–225:** The GDP values and reported growth factor are inconsistent. An increase from €5.25 billion to €1.93 trillion is approximately 368-fold, rather than 7.8-fold. Verify the values, units, price basis, and Figure 2 axis.
- **Lines 227–229 and Figure 2D:** Clarify whether MOR represents annual deaths or a mortality rate. The definition, magnitude, and units need to agree.
- **Section 5.4:** Methods define improvement using negative ΔAIC, whereas results describe positive ΔAIC improvements. Adopt one convention throughout.
- **Section 5.2:** The reference to overdispersion "reported in Table 2" appears incorrect; Table 2 presents trend and dependence diagnostics.
- **Table 1:** Define how annual lethality is calculated in flood years with zero fatal days. Also explain that flood and landslide fatal-day totals overlap: their sum exceeds the combined total because both hazards can cause deaths on the same date.
- **Table 2:** Replace $p = 0.000$ with an appropriate inequality, define the confidence level explicitly, and provide the thresholds used to classify Hurst estimates.
- **Figures 2 and 5:** Enlarge labels and plotting areas. Correct the stray "Figure 1" text within Figure 2 and the inconsistent panel references in the Figure 5 caption.
- **Code and data availability:** Provide the annual analytical dataset and analysis scripts, including Stan code and policy indicators. Listing software versions and source websites alone does not make the analysis reproducible.

After these revisions, the manuscript could make a useful contribution. Its strongest contribution is the documentation of historical changes in recorded mortality and explanations of those changes should be more clearly separated from the descriptive evidence.