https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Worth waiting for, hardly worth waiting: data latency and the timing of regional landslide warnings in China

Hongyun Zhang[1]*, Jie Liu[2], Jin Liu[3], Jianchuang Ding[1], Yuanhao Zhu[1]

[1] School of Disaster Prevention and Mitigation Engineering, University of Emergency Management, Langfang 065201, China

[2] School of Marine Technology and Geomatics, Jiangsu Ocean University, Lianyungang 222005, China

[3] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China

Correspondence to: Hongyun Zhang, zhanghongyun@whu.edu.cn

Abstract. Landslide warning services must weigh the better evidence that later data bring against the lead time lost while waiting for them. We studied this trade-off for 280 rainfall-induced landslides in China (2015–2024) with minute-, hour- or period-level occurrence times reported in the news, using the Global Precipitation Measurement (GPM) mission's Integrated Multi-satellitE Retrievals for GPM (IMERG) precipitation and as-issued Global Forecast System (GFS) forecasts in a latency-aware retrospective replay. In a case-crossover design, observed rainfall replacing forecasts added substantial information (matched area under the receiver operating characteristic curve +0.058 [0.034, 0.081]), whereas later satellite runs and newer forecast cycles added little or none. In a rolling replay of 24 h warnings, waiting one 6 h step before issuing gave no stable gain in detection at warning loads of 0.5–10% of cell-time. In a cost–loss analysis, issuing at once with satellite rainfall plus forecast reached a relative economic value (REV, dimensionless) of 0.55, close to a latency-free benchmark (0.57) and above a rainfall threshold (0.34) and, on matched 2015–2020 decisions, the archived version 2.0.0 Landslide Hazard Assessment for Situational Awareness (LHASA) nowcast (0.34). A value-of-information waiting rule added +0.002 [−0.005, +0.007] over the best fixed policy, averaged across the exploratory cost–loss range. The pre-specified cost–loss test could not discriminate, because almost no rule warns at its cost–loss ratios, but a second test, fixed before an independent 2008–2014 sample was assembled, detected a value of waiting when no forecast was available (+0.03 [0.01, 0.05] in REV). Brackets give 95% confidence intervals. With both satellite rainfall and forecasts available, waiting rarely repaid its lead-time cost.

Keywords: landslide early warning; data latency; value of information; cost–loss model; IMERG; GFS; case-crossover

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

1 Introduction

Regional landslide early warning systems translate rainfall observations and forecasts into warnings for large areas (Piciullo et al., 2018; Guzzetti et al., 2020), from national services such as Norway's (Krøgli et al., 2018) to global satellite-based nowcasts (Kirschbaum and Stanley, 2018; Stanley et al., 2021). Most rely on rainfall thresholds or data-driven models (Guzzetti et al., 2008; Segoni et al., 2018) that are calibrated as if the triggering rainfall were known when the warning is issued. In practice the data arrive in batches. The Integrated Multi-satellitE Retrievals for GPM (IMERG) of the Global Precipitation Measurement (GPM) mission are released about 4 h after observation (Early run), about 14 h after (Late run) and, gauge-adjusted, about 3.5 months after the observation month (Final run; Huffman et al., 2023a). Global forecasts such as the Global Forecast System (GFS) are issued every six hours and become available a few hours later. At any decision time a warning service can issue now with the evidence at hand, or wait for the next release and lose lead time. China's national geo-hazard meteorological warning service, which issues 24 h regional warnings from rainfall observations and forecasts (Xiao et al., 2024), faces exactly this choice.

Waiting can improve the evidence in three ways: the next forecast cycle may be more skilful (channel 1), a later, better-processed version of the same observation may arrive (channel 2, e.g. IMERG Late replacing Early), and part of the period that could only be forecast becomes observed (channel 3). Jewson et al. (2021) extended the cost–loss model (Murphy, 1977) to the question "decide now or wait for the next forecast?" and showed that the answer depends on the predictive distribution of the next forecast, and run-to-run changes carry information of their own (Marchal et al., 2025). Hazard warnings have been framed as Bayesian decisions (Economou et al., 2016), and the economic value of forecasts for users with different cost–loss ratios is well established (Katz and Murphy, 1997; Zhu et al., 2002). For landslides, the gain from waiting has to be weighed against the risk left unprotected during the delay.

Two obstacles have kept this question out of landslide research. It needs occurrence times precise to hours, whereas most inventories record dates at best, and landslide catalogues are incomplete and biased by reporting (Kirschbaum et al., 2015; Froude and Petley, 2018), so comparisons across places confound rainfall with the chance of a report. We use a recent catalogue of landslides in China extracted from news text (Zhao et al., 2026), whose records carry minute-, hour- or period-of-day times. A time-stratified case-crossover design (Maclure, 1991; Janes et al., 2005) measures the information in each data stream by comparing each location with itself, which removes reporting differences between places. A replay of six-hourly decisions at each location measures the decision value. Because it pools locations, its values refer to reported landslides.

We ask when newer rainfall information justifies delaying a regional landslide warning. The study separates the three sources of information gained by waiting within one latency-aware reconstruction and follows them through to warning decisions, with a value-of-information (VOI) rule in a cost–loss setting and a rolling warning replay at matched warning budgets. A rainfall threshold and an archived version 2.0.0 LHASA global nowcast serve as baselines, and an independent 2008–2014 test without forecasts checks that the method detects a value of waiting where one exists. Unlike recent work on

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

satellite-driven warning platforms (e.g. Yunus et al., 2026), we address the timing of decisions within an existing information chain.

2 Data

2.1 Landslide records

We used the news-derived landslide catalogue of Zhao et al. (2026) (figshare, doi:10.6084/m9.figshare.29603420), which lists 1582 landslide events in China in 2008–2024, extracted from news reports with a large language model and checked in several stages. Its evaluation reports 82.4% accuracy for extracted time fields. We kept rainfall-triggered records in April–October 2015–2024, which hold 97% of the qualifying records, located at county level or finer and timed to the hour (T3), the minute (T4) or a named period of the day (T2). This gives 280 records (213 T3/T4, 67 T2) at 263 grid cells of 0.1° in 23 provinces (Fig. 1a–c). The occurrence time T is the reported start for T3/T4 records and the midpoint of the period for T2 records (Appendix A); 62 of the 67 periods span 6 h or less, which places the midpoint within 3 h of any time in the period, and five (night, since morning) span 12–18 h. Times were read as Beijing time and converted to Coordinated Universal Time (UTC; Beijing time is UTC+8). Records within 72 h and 300 km of each other form one rainfall event (193 events in the main sub-daily sample). Available event clusters are the resampling units for the case-crossover, lead-time and detection intervals; the daily resolution control retains the grouping described in Appendix A. The cost–loss analysis resamples 7-day blocks of all cells. Date-only records (T1) entered only a daily-scale sensitivity analysis, and the 354 qualifying records of 2008–2014 (215 events at 321 cells) only the case-crossover replication and the independent Early-only test.

News reports over-represent populated places and damaging events, and the absence of a record does not mean the absence of a landslide. All labels are therefore reported landslides; Sect. 5.4 discusses what this implies for each analysis.

2.2 Rainfall observations and forecasts

IMERG V07 half-hourly precipitation at 0.1° (Early, Late and Final runs; Huffman et al., 2023b, c, d) was extracted as point series at the record cells for 2008–2024, and daily fields over China (Huffman et al., 2023e, f, g) served the resolution control. Daily totals from the two sources agree to within 1.3×10^{-4} mm at all 1914 extracted cells. Over mainland China, IMERG outperforms its predecessor, the Tropical Rainfall Measuring Mission Multi-satellite Precipitation Analysis, at sub-daily scales, and the Final run outperforms the near-real-time runs, with skill declining from the humid southeast to the arid northwest (Tang et al., 2016; Wang et al., 2018). GFS 0.25° forecasts of all four daily cycles in April–October 2015–2024 came from the Geoscience Data Exchange of the National Center for Atmospheric Research for 2015–2020 (NCEP, 2015) and from the archive of the National Oceanic and Atmospheric Administration on Amazon Web Services (AWS) for 2021–2024 (NOAA, 2026), as six-hourly precipitation buckets at leads 6–30 h (9 of 42 800 fields missing). The AWS time stamps give a median publication delay of 3.6–3.7 h (99% within 4.8 h), so at decision time t the latest available cycle is t − 6 h.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

The GFS archive holds the forecasts as issued, but the IMERG record does not. V07 Early and Late entered real-time production only in June 2024, and the 2015–2023 record was produced afterwards by retrospective processing, including the climatological calibration of Early and Late to Final (Huffman et al., 2024; NASA GPM, 2024). The runs that operators received in 2015–2023 were earlier versions, which NASA advises against mixing with V07 (NASA GPM, 2023). The replay is therefore faithful to latency but not to product version, except in June–October 2024.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

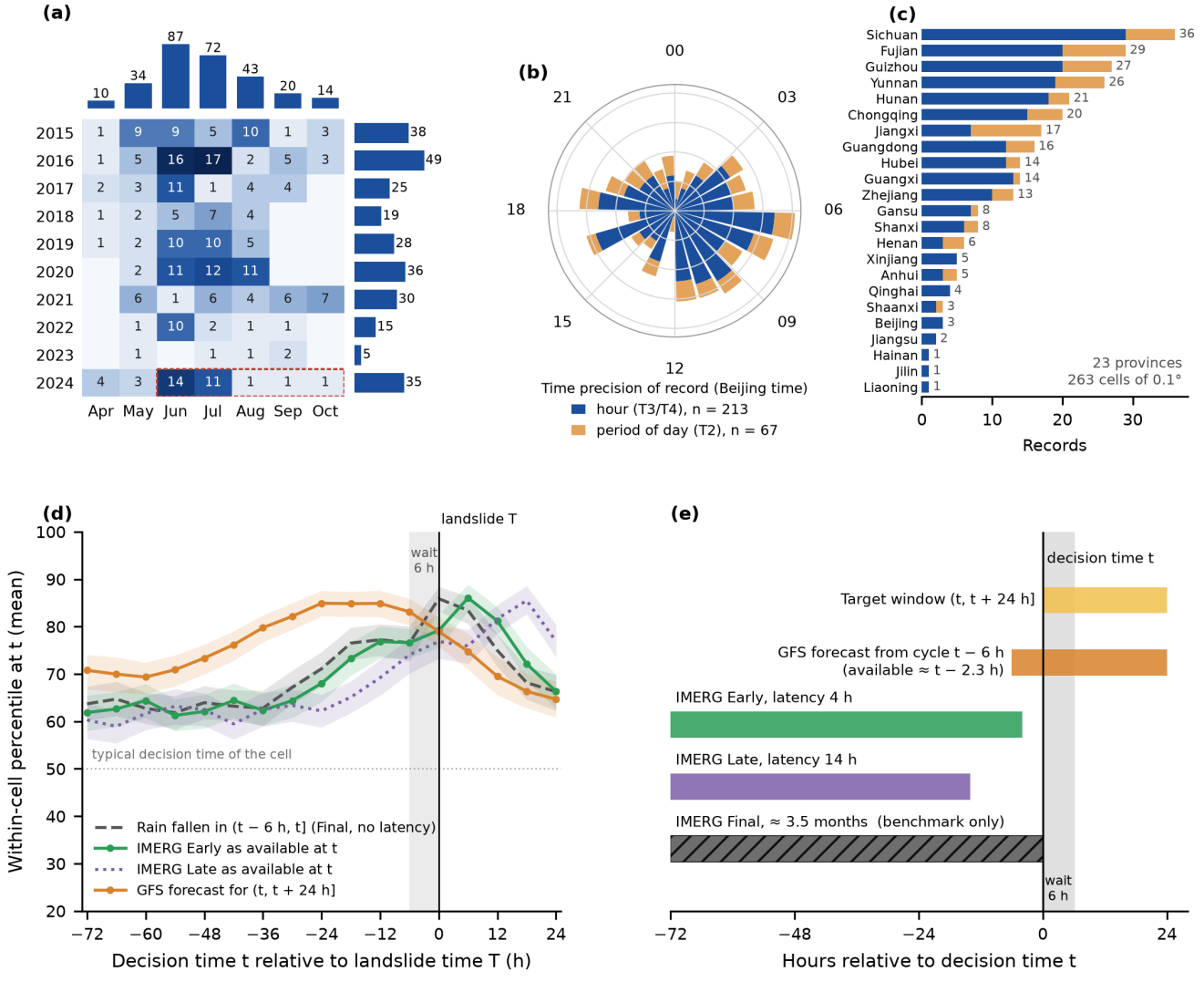


Figure 1. The landslide records and the data available at decision time. (a) Qualifying rainfall-induced records by year and month (April–October 2015–2024); dashed box: June–October 2024, when IMERG V07 Early and Late were produced in real time. (b) Occurrence time of day (Beijing time): hour- and minute-level (T3/T4) records by reported hour, period-level (T2) records spread over their period. (c) Records by province and time precision. (d) Superposed-epoch composite of the 280 records: mean within-cell percentile (over all April–October decision times of the cell) of what each data stream shows at decision time t, against t − T in 6 h bins, for rain fallen in (t − 6 h, t] (IMERG Final, no latency), IMERG Early and Late 6 h totals as available at t, and the GFS forecast for (t, t + 24 h]; bands are pointwise 95% confidence intervals from 1000 resamples of the 193 rainfall events, and grey marks the 6 h before the landslide. (e) Data latency at t: IMERG Early and Late totals end 4 h and 14 h before t, the latest GFS cycle (t − 6 h) forecasts (t − 6 h, t + 24 h], and IMERG Final serves only as a latency-free benchmark. Grey marks the 6 h wait. With the 4 h Early latency, observations available at t + 6 h extend to t + 2 h, so only the first 2 h of (t, t + 6 h] have become observable.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

2.3 Data latency at a decision time

Decisions are made at 00, 06, 12 and 18 UTC. At decision time t, each data stream covers a known period (Fig. 1e; Table 1): IMERG Early totals over 6, 24, 72 and 168 h ending at t − 4 h, IMERG Late totals ending at t − 14 h, and GFS from cycle t − 6 h for (t − 6 h, t], which fills the hours that Early has not yet reported, and for (t, t + 24 h]. IMERG Final totals ending at t serve only as a latency-free benchmark, fitted with the same model form. The replay reconstructs when data became available. It does not reproduce the historical IMERG versions or the warnings an operational service issued.

Table 1. Data streams and their latency at decision time t.

Stream	Resolution	Latency	Window used at t	Role
IMERG V07 Early	0.1°, 30 min	~4 h	6/24/72/168 h ending t − 4 h	available (retrospective V07 for 2015–2023)
IMERG V07 Late	0.1°, 30 min	~14 h	6/24/72/168 h ending t − 14 h	available (retrospective V07 for 2015–2023)
IMERG V07 Final	0.1°, 30 min	~3.5 months	6/24/72/168 h ending t	benchmark only
GFS 0.25°	0.25°, 6 h buckets	cycle + 3.6–3.7 h (median)	cycle t − 6 h: (t − 6 h, t] and (t, t + 24 h]	available (as issued)

3 Methods

Figure 2 summarizes the design: the reconstruction of the information available at each decision time (a), the case-crossover comparison of the three channels (b), the probability models and the two decision evaluations (c), and the validation and sensitivity analyses (d).

3.1 Case-crossover comparison of data streams

For each record, the case is the rainfall predictor at the record's cell and time, and the referents are the same cell and clock time on the other days of the same calendar month at least 3 days away, a time stratification that controls for season and long-term trends (Janes et al., 2005; Carracedo-Martínez et al., 2010). Each predictor is scored by the matched area under the receiver operating characteristic curve (AUC), the case's percentile among its referents averaged over records (0.5 = no information), and by the log-likelihood of a conditional logistic model, the case-crossover analogue of the logarithmic score (Roulston and Smith, 2002). Log-likelihoods of models with more than one predictor are computed out of fold over ten folds of whole rainfall events. Confidence intervals (CIs, 95%) come from 5000 resamples of rainfall events. A placebo shifts every case and referent 7 days earlier, and a sensitivity analysis restricts referents to ±10 days.

Three channels are compared over identical windows aligned with the six-hourly GFS buckets: a 24 h window around each record (all records) and the containing 6 h window (T3/T4). They compare, respectively, the GFS cycle issued at the window

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

start with the cycle 6 h older; IMERG Late with Early; and Early with the older GFS cycle. A daily control averages IMERG over 0.3° and 0.5° neighbourhoods before comparison with 0.25° GFS; its event grouping is described in Appendix A.

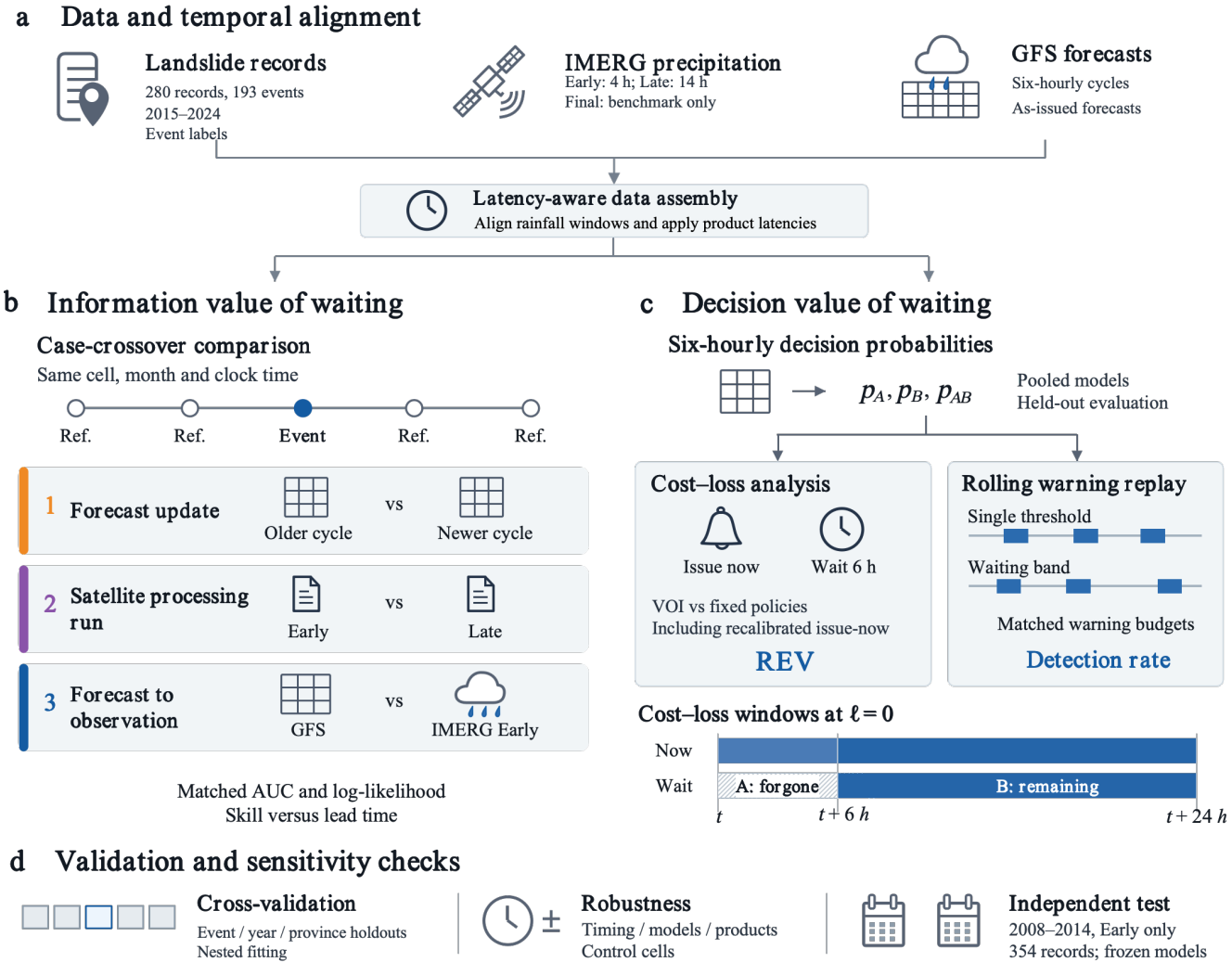


Figure 2. Study design. (a) Reported landslides (280 records, 193 rainfall events, 2015–2024), IMERG precipitation and as-issued GFS forecasts aligned at six-hourly decision times with their latencies (IMERG Early 4 h, Late 14 h; Final only as a latency-free benchmark). (b) Case-crossover comparisons within the same cell, month and clock time separate forecast updates, Early-to-Late processing updates and observations replacing forecasts (Ref.: referent time). (c) Pooled probability models feed the cost–loss evaluation of the value-of-information (VOI) policy against fixed policies and the rolling replay of threshold waiting bands against single thresholds at matched warning budgets. With required lead ℓ = 0, waiting from t to t + 6 h forgoes A = (t, t + 6 h] and keeps B = (t + 6 h, t + 24 h]; p_A, p_B and p_{AB} are the current probabilities of a reported landslide in A, B and their union AB. (d) Event, year and province hold-outs, nested fitting, sensitivity analyses and the independent Early-only test on 354 records from 2008–2014.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

3.2 Lead time and skill

For each T3/T4 record and required lead ℓ = 0, 6, …, 48 h, the decision time is $d = \text{floor}_6(T) - \ell$, where $\text{floor}_6(T)$ is the last six-hourly decision time at or before T. Multi-predictor conditional logistic models were fitted with the data available at d for five information sets: all available (Early + GFS), Early only, GFS only, Late + GFS and the latency-free benchmark Final + GFS. The out-of-fold log-likelihood gain over the no-information model measures the information each additional 6 h of waiting adds. Only records with complete data at every lead were used, so the comparisons are paired, and pointwise CIs come from 5000 rainfall-event resamples.

3.3 Decision probabilities and predictive performance

For the decision analysis, every record cell contributes six-hourly decisions in April–October 2015–2024, 2 247 335 in total after removing the 0.15% with missing GFS data. The label for a lead threshold ℓ marks a reported landslide at the cell in (t + ℓ, t + 24 h], requiring a lead strictly greater than ℓ and no greater than 24 h. A record exactly at t + ℓ is excluded. The positive-label base rate is 4.8×10^{-4} per decision at ℓ = 0. Logistic regressions on log-transformed rainfall totals were fitted across pooled cells, one per information set and target window, and evaluated by leave-one-year-out cross-validation, which holds out each of the 273 records with complete features once (Appendix A). We also report the pre-specified split (train 2015–2019, validate 2020–2021, test 2022–2024) with LightGBM and logistic regression. At this base rate the Brier score barely separates models, so log loss is the main proper score (Gneiting and Raftery, 2007), with the pre-specified Brier score also reported. Skill scores use the training base rate as reference. Intervals for scores and economic values come from resampling 7-day blocks of all cells, which keeps spatial dependence within each block. Detection rates are the share of records warned more than ℓ hours ahead at a given warned share of decision steps, with intervals from resampling rainfall events.

3.4 Baseline warning models

The conventional baseline B1 is a frequentist cumulated event rainfall–duration (E–D) threshold (Guzzetti et al., 2008; Brunetti et al., 2010; Segoni et al., 2018) built from the same IMERG Early data. Rain events were reconstructed at each cell from hourly totals, with wet hours of at least 0.2 mm and events separated by at least 48 dry hours (cf. Melillo et al., 2015). At decision time t, E (mm) and D (h) are the cumulative rainfall and duration of the current event up to the last wet hour before t − 4 h. The threshold $\log_{10}E = \log_{10}a + b \cdot \log_{10}D$ uses a as the power-law scale coefficient and b as the exponent ($\log_{10}a$ is the intercept and b the slope in log space, with E in mm and D in h). It was fitted by least squares to the triggering conditions of the 277 records with a rain event at their occurrence time and lowered to the 5th percentile of the residuals. The distance above the threshold was mapped to probabilities by logistic regression, with threshold and mapping fitted without the held-out year (Appendix A), so B1 enters both evaluations like the other information sets.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

The second baseline (B2) uses the archived version 2.0.0 global nowcast of the Landslide Hazard Assessment for Situational Awareness (LHASA; Stanley et al., 2021), driven by daily IMERG Early rainfall and Soil Moisture Active Passive (SMAP) soil moisture. The retrospective archive (Stanley, 2023) covers April 2015 to February 2021. For each 0.1° cell we sampled a 3 × 3 array of 1 km pixels at four-pixel spacing and used their valid maximum and mean from the latest UTC day treated as available under a 5 h delay after day end. These two predictors were logit-transformed and mapped to probabilities by leave-one-year-out logistic regression, with Early + GFS refitted on the same six years (Appendix A). The paired evaluation uses 1 344 193 decisions at 263 cells and 195 records over April–October 2015–2020, starting on 5 April 2015. A separate fixed-threshold comparison uses a sampled-pixel maximum above 0.5, a threshold selected for this study rather than an official LHASA operating point. This comparison concerns the retrospective daily archive under the stated sampling and latency assumptions, rather than historical operational warnings.

3.5 Cost–loss evaluation of selective waiting

Each decision is treated as a cost–loss problem (Murphy, 1977). The cost–loss ratio α is the cost of issuing a warning divided by the loss from an unprotected reported landslide. Normalizing that loss to 1 gives a warning cost of α; 1/α is therefore the number of warnings one would accept to avoid one missed landslide. A warning issued at decision time t protects landslides in AB = (t + ℓ, t + 24 h]. Waiting one six-hourly step before issuing protects only B = (t + ℓ + 6 h, t + 24 h], leaving A = (t + ℓ, t + ℓ + 6 h] unprotected. Let p_A, p_B and p_{AB} be the probabilities, given information at t, of at least one reported landslide in A, B and AB, respectively, and let $q = P$ (B | information at t + 6 h) be the probability available after waiting. Here P denotes probability and 1[·] is an indicator equal to 1 when its argument is true and 0 otherwise; 1[A], 1[B] and 1[AB] indicate a record in the corresponding window. The realized normalized expense is α for issuing now, 1[AB] for not warning, and α·1[q > α] + max(1[A], 1[B]·1[q ≤ α]) for waiting. The VOI policy uses the following myopic approximation to the value of waiting:

$$\mathrm{VOI} \approx \min(\alpha, p_{AB}) - [p_A + \mathbb{E}\min(\alpha, q)], \quad (1)$$

where $\mathbb{E}$ denotes expectation over the preposterior distribution of q given the current information (Raiffa and Schlaifer, 1961; Jewson et al., 2021). We estimated this distribution from the other years by grouping decisions with similar current probability $\boldsymbol{p_B}$ and similar 24 h GFS forecast and taking the distribution of q within each group. In a fully nested version, the held-out year entered neither the models nor the groups (Appendix A). The VOI policy (the waiting band of the cost–loss plots) waits when VOI > 0 and otherwise issues if $\boldsymbol{p_{AB}} > \alpha$. It is compared with fixed policies: issuing now with all available information, with Late + GFS or with GFS only, and always waiting one step. Issue-now policies with Early only, B1 and the latency-free benchmark, and VOI policies with Early only or GFS only, were also evaluated. Equation (1) adds the expected loss in A to the expected warning-or-loss expense in B. It neglects the overlap correction when both windows contain a record and no warning is issued after waiting, whereas the realized expense caps the missed-event loss at one. This discrepancy affects 2–3 of the 2.2 million evaluated decisions.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Two variants test whether gains of the VOI policy arise from the probabilities rather than from waiting: an issue-now policy recalibrated within the cross-validation, added to the fixed policies, and window probabilities derived coherently from shared 6 h sub-window models. An oracle reference that knows the next-step probability q before waiting illustrates the value of that knowledge. Appendix A gives the definitions and the calibration and coherence diagnostics.

The pre-specified H2 analysis assessed the Spearman rank correlation coefficient ρ between predicted VOI and the realized gain of waiting over all decisions. The implemented analysis estimates ρ at each α and lead, with pointwise 95% intervals from 2000 resamples of 7-day blocks (Appendix E). Separately, an exploratory analysis compared realized gains across deciles of predicted VOI among decisions where the issue-now rule warns, a contrast that does not depend on the realized q through the prediction. This subset analysis was specified after the main results and is reported separately from the original H2 test.

Relative economic value (REV; Richardson, 2000) compares the expense of a policy with that of the climatological and perfect-forecast references:

$$\mathrm{REV} = \frac{e_{\mathrm{clim}} - e}{e_{\mathrm{clim}} - e_{\mathrm{perfect}}} \tag{2}$$

where e is the mean normalized expense of the evaluated policy, $e_{\mathrm{clim}} = \min(\alpha, s)$ is the expense of the better of always warning and never warning, $e_{\mathrm{perfect}} = \alpha s$ is the expense of a perfect forecast, and s is the frequency of positive labels in AB. Thus Eq. (2) is dimensionless: REV = 0 denotes the climatological reference, REV = 1 denotes a perfect forecast, and REV < 0 indicates greater expense than the climatological reference. We evaluated α on a 23-point logarithmic grid from 10^{-4} to 0.2 and required leads ℓ = 0, 6 and 12 h. Intervals come from 2000 paired resamples of 7-day blocks, with the best fixed policy re-selected within each resample.

3.6 Rolling warning replay

Evaluating each decision separately can misrepresent the cost of waiting when warning windows overlap. We therefore also replayed each cell as a warning service, in which a warning issued at t remains active over (t, t + 24 h]. A record is detected with lead > ℓ if an issue time t satisfies T − 24 h ≤ t < T − ℓ; warning load is the share of cell-time under an active warning. For each lead threshold ℓ, p_t is the current probability of a record in (t + ℓ, t + 24 h]. A single-threshold policy issues if $p_t > \tau$, where τ is the issuance threshold. A threshold waiting band issues immediately if $p_t > \tau_{hi}$, or issues after one step if the probability available at the preceding decision time t − 6 h lay in $(\tau_{lo}, \tau_{hi}]$ and $p_t > \tau_c$. The parameters τ_{lo} and τ_{hi} delimit the deferral band, and τ_c is the confirmation threshold after waiting; both τ_{lo} and τ_c are below τ_{hi}. For each warning budget (0.5–10% of cell-time), parameters were chosen on nine years to maximize detection within the budget and evaluated on the tenth. The candidate family also includes all single-threshold policies, giving both families the same immediate-issuance candidates (Appendix A). We repeated the replay with a denser band grid and nested threshold selection, and evaluated the conventional B1 operating point, which warns whenever current event rainfall exceeds its threshold. These threshold waiting bands are distinct from the VOI policy.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

3.7 Robustness and sensitivity analyses

We repeated the decision analyses with province hold-outs (23 folds, every model and preposterior group refitted), with labels from T3/T4 records only, with occurrence times perturbed by uniform random errors of up to ±1 h and ±3 h, and with all times shifted by −3 h and +3 h, as if reports timed landslides late (for example at the storm peak) or early. The case-crossover channel comparisons were repeated with the perturbed and shifted times, and every logistic model was also replaced by LightGBM (Appendix A).

Two analyses address the retrospective IMERG record. The leave-one-year-out models for 2024, trained on the retrospective record, were evaluated on decisions from 1 June 2024, when V07 Early and Late were produced in real time (28 records, 160 693 decisions). Early and Late were also perturbed by a mean-one lognormal factor per cell and day (log-scale standard deviation $\sigma = 0.3$ or 0.6, shared by the two runs), with Final unchanged and all models refitted.

A third check addresses the spatial population. The decision table contains only cells with a qualifying record, and only five of the 263 cells have records in more than one year, so a held-out year's positive cells are present mostly because of that year's landslide. We therefore added 1000 control cells drawn at random from the 52 919 cells of mainland China with mean slope ≥ 5° (Copernicus GLO-90 digital elevation model) and no rainfall-triggered record in any year or time precision. Record cells have sampling weight 1 and sampled control cells weight 52 919/1000. The weighted target population is the union of the qualifying record cells and the 52 919 no-record cells with mean slope ≥ 5°. It excludes 583 cells (1.1%) whose only records are date-only, from 2008–2014 or outside April–October, and retains 22 qualifying record cells with mean slope below 5°; it therefore does not represent every cell above the slope threshold. Unweighted results are also reported. Because the weighted base rate is about 200 times lower, the α grid was extended down to 10^{-7}, with the exploratory range set to 0.2–20 times the base rate (close to 10^{-4}–10^{-2} for record cells alone). Province hold-outs were not repeated with control cells.

A REV difference is called material when its absolute mean exceeds 0.01 and its 95% interval excludes zero; 0.01 is about 2% of peak REV and comparable to the spread of peak REV across validation variants (Table C1). Because the intervals are pointwise, counts of α values or budgets with intervals excluding zero are descriptive. The overall tests are the mean difference over the α values of the exploratory range, with its block-bootstrap interval, and a simultaneous 95% band over that range from the bootstrap maximum of the standardized differences. The rolling replay is summarized in the same way over its five budgets. The materiality criterion and the overall tests were set after the main results (Appendix B).

3.8 Pre-specified tests and exploratory analyses

An internal analysis plan with hypotheses and stopping criteria was written before any model was trained. It was not publicly registered, and its dates derive from internal records. Appendix B summarizes the plan and its deviations; Appendix D gives the independent-test protocol and execution record. The original plan stated four hypotheses: H1, IMERG Late improves the Brier score over Early for the same valid period; H2, predicted VOI correlates positively with the realized gain of waiting; H3, the VOI rule has higher REV than the best fixed policy over at least half of $\alpha \in [0.01, 0.2]$; and H4, the H3 advantage

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

does not reverse under province hold-outs. Leave-one-year-out validation and the log loss replaced the pre-specified split and Brier score after the first decision-model results, and all other analyses were specified later (Appendix B).

Because the pre-specified H3 test could not discriminate, we fixed a second confirmatory test after all other analyses but before any decision data from 2008–2014 were assembled (Appendix D). These records had entered no decision analysis, and the GFS archive used here does not reach back to them, so the test concerns information without a forecast, for which waiting had value in 2015–2024. Early-only window models and preposterior groups (on $\boldsymbol{p}_{\mathbf{B}}$ alone) were fitted on 2015–2024 and applied unchanged. The primary hypothesis C1, a positive control, was that the VOI policy outperforms issuing at once with Early only, averaged over the seven α values from 10^{-4} to 8×10^{-4} at $\ell = 0$ (95% lower limit above zero, 2000 resamples of 7-day blocks). The secondary hypothesis C2 was that always waiting one step does worse.

H1, H2, C1 and C2 were pre-specified. H2 was a pre-specified hypothesis, but its pointwise implementation does not constitute a single fully specified confirmatory test. The correlation over all decisions is distinguished from the later exploratory warning-subset analyses; a positive association in that subset does not by itself confirm the original H2 criterion. H3 was pre-specified but could not discriminate, because the evaluated cost–loss range produced few warnings, and H4, which presupposed an H3 advantage, is read as a check on the stability of the decision conclusions. All other analyses are exploratory, and the decision conclusions rest on their agreement and confidence limits.

4 Results

4.1 Observation replacing forecast is the informative channel

The three channels differ by an order of magnitude (Fig. 3). Over the same window, IMERG Early discriminates landslide times better than the GFS forecast by +0.058 [+0.034, +0.081] in matched AUC and +0.38 [+0.24, +0.52] in log-likelihood per case, about five times the gain from a later satellite run, and once Early is known the forecast for the same window adds no detectable gain (+0.027 [−0.006, +0.058]). IMERG Late improves on Early by +0.009 [+0.003, +0.016] in AUC and +0.055 [+0.020, +0.092] in log-likelihood (24 h windows), consistently across daily and half-hourly analyses and in the independent 2008–2014 sample, and Final adds little beyond Late. A newer GFS cycle shows no detectable gain over the cycle issued 6 h earlier, over 24 h windows (+0.001 [−0.013, +0.014]) or 6 h windows (−0.012 [−0.029, +0.004]), and skill was flat across forecast ages of 6–30 h.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

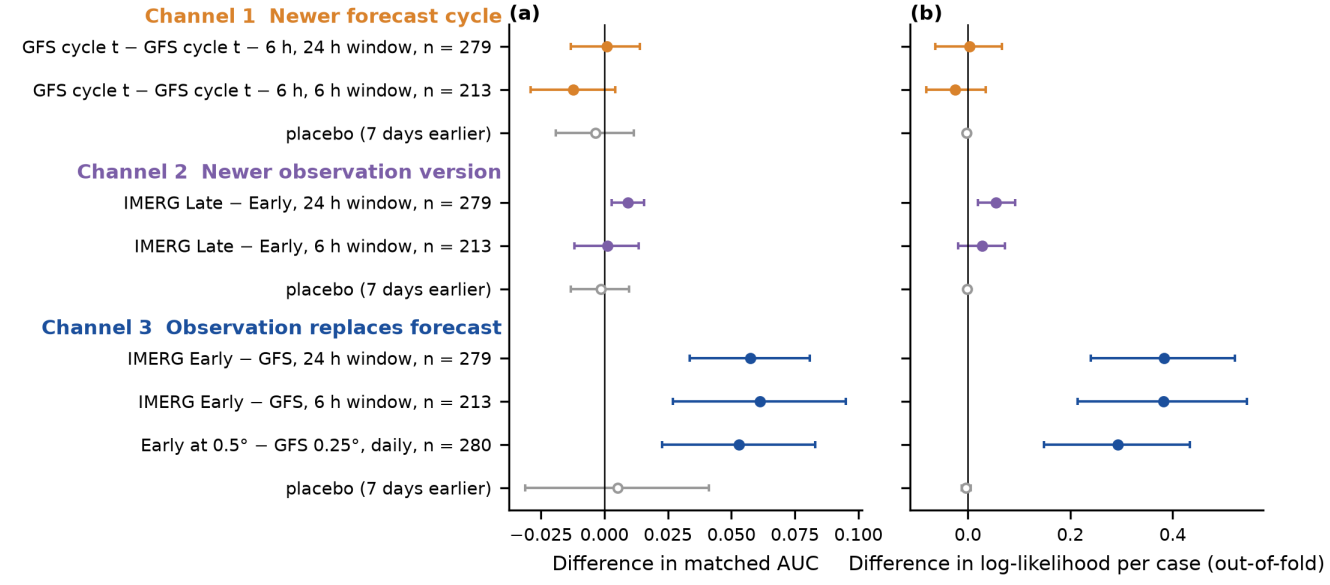


Figure 3. Three channels of the value of waiting, from the case-crossover design: differences in (a) matched AUC and (b) out-of-fold log-likelihood per case, over identical windows aligned with GFS buckets. Channel 1, the GFS cycle issued at the window start against the cycle 6 h older; channel 2, IMERG Late against Early; channel 3, IMERG Early against GFS, and Early averaged to 0.5° against GFS 0.25° (daily resolution control). Open grey symbols: placebo with all times shifted 7 days earlier. Points: estimated mean differences. Horizontal error bars: pointwise 95% confidence intervals from 5000 bootstrap resamples of the event clusters available in each comparison: 192 events for 279 records (24 h), 156 for 213 records (6 h), and 183 for 280 records (daily resolution control). The plotted 24 h placebo uses 275 records in 190 events. The daily control retains event membership assigned before the time-precision restriction (Appendix A). The vertical line marks zero difference.

The advantage of the observation is not an artefact of resolution or timing. It persists when IMERG is averaged to 0.5° (+0.053 [+0.023, +0.083] in daily AUC), all placebo differences are close to zero, and it survives random timing errors of ±3 h (+0.061 [+0.037, +0.085]) and systematic shifts of 3 h earlier or later (+0.074 [+0.049, +0.101] and +0.059 [+0.038, +0.081]). Under these perturbations channel 1 stays non-significant and channel 2 stays at +0.007 to +0.008.

4.2 Skill rises as the decision approaches the landslide

With only the data available at each decision time, skill falls monotonically with lead time (Fig. 4a): the log-likelihood gain of all available information is 0.91 per case at ℓ = 0–6 h, 0.60 at 24–30 h and 0.21 at 48–54 h. Within 24 h of the landslide the GFS-only skill is roughly flat (0.58–0.63), whereas the Early-only skill climbs from 0.15 at 24–30 h to 0.62 at 0–6 h as the triggering rainfall becomes observed. Waiting 6 h raises the log-likelihood of all available information by 0.04–0.11 per case at most leads (Fig. 4b). The latency-free benchmark exceeds it by only 0.03–0.10, and Late + GFS is indistinguishable from Early + GFS. The composite of the 280 records shows why (Fig. 1d). The rain at a record cell is most unusual for that cell in the 6 h before the landslide (mean within-cell percentile about 86), but IMERG Early shows that peak one decision step later and Late about 18 h later, while the GFS forecast for the next 24 h is already high (about 85) from 24 to 12 h before. The gain in information as the landslide approaches thus comes mainly from observations replacing forecasts.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

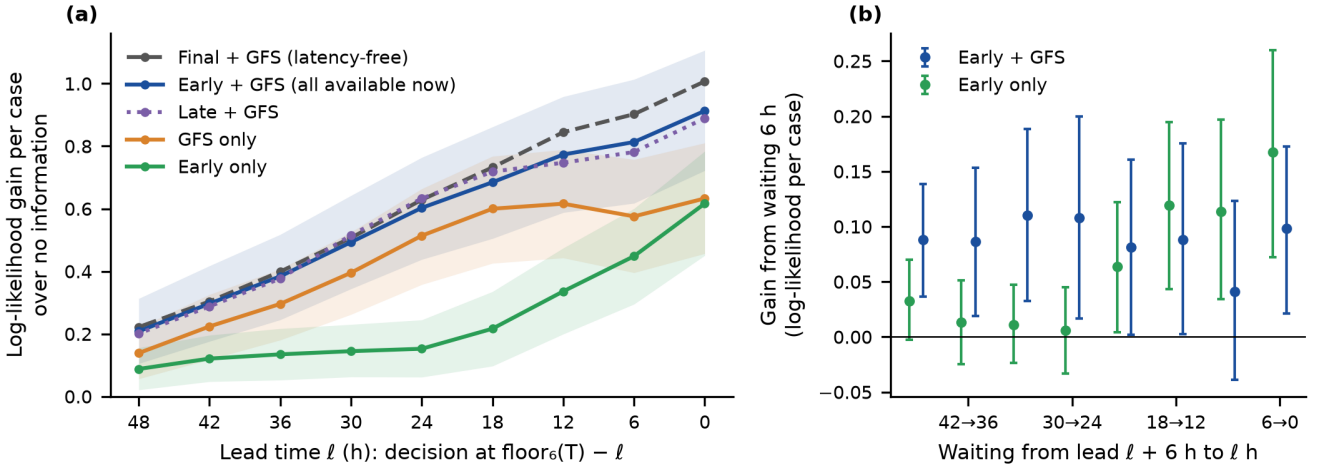


Figure 4. (a) Out-of-fold log-likelihood gain per case over no information, against lead time ℓ (decision at floor₆(T) − ℓ, using only data available then), for five information sets (hour-level T3 and minute-level T4 records, n = 213). Same-colour translucent bands show pointwise 95% confidence intervals for Early + GFS (blue), GFS only (orange) and Early only (green); bands are omitted for Late + GFS and Final + GFS. (b) Mean paired gain from waiting 6 h, from lead ℓ + 6 h to ℓ, for all available information and Early only, with vertical 95% confidence intervals and a horizontal zero-gain reference. All intervals use 5000 rainfall-event bootstrap resamples.

4.3 Observations improve the decision probabilities

On the pooled decision dataset (Table 2), all available information gives a held-out AUC of 0.875 and a log-loss skill of 0.154 [0.129, 0.178] at ℓ = 0, ahead of GFS only, Early only and the E–D threshold B1 (0.761 and 0.051 [0.037, 0.064]). Adding the observation to the forecast improves the log loss (-1.7×10^{-4} per decision [-2.3×10^{-4}, -1.0×10^{-4}]), the Brier score and the detection rate at a 2% warning budget (0.52 against 0.46; +0.07 [+0.02, +0.12]). Late + GFS is slightly worse than Early + GFS in log loss ($+2.6 \times 10^{-5}$ [+0.6, +4.5] $\times 10^{-5}$), as expected for data 10 h older. H1 is not supported: over the same windows, Late and Early differ by -6.3×10^{-6} [-1.9×10^{-5}, $+5.7 \times 10^{-6}$] in log loss and -2.9×10^{-8} [-6.9×10^{-8}, $+9.1 \times 10^{-9}$] in Brier score, because the many dry decision times dilute the case-crossover gain. Lead time is costly: at a 2% budget, detection falls from 0.52 (lead > 0 h) to 0.49, 0.44 and 0.33 (lead > 6, 12 and 18 h). The pre-specified split (55 test records) gives the same ordering with wide intervals. The archived LHASA version 2.0.0 baseline, evaluated on matched April–October 2015–2020 decisions, reached an AUC of 0.725 and a log-loss skill of 0.033 [0.011, 0.049] at ℓ = 0, against 0.853 and 0.137 [0.106, 0.167] for Early + GFS refitted on the same years, and adding it to Early + GFS produced no measurable log-loss improvement at the reported precision.

Table 2. Held-out performance on the pooled cell-level decision dataset (leave-one-year-out, ℓ = 0, 2 247 335 decisions, 273 records, 193 events). Detection = share of records warned more than ℓ h ahead when warnings occupy 1% or 2% of decision steps. The E–D baseline B1 is evaluated through its probability mapping; its rolling warning performance is reported separately.

Information set	AUC	Log-loss skill [95% CI]	Brier skill	Detection 1%	Detection 2% [95% CI]
Early + GFS (all available)	0.875	0.154 [0.129, 0.178]	0.0046	0.40	0.52 [0.45, 0.60]

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Information set	AUC	Log-loss skill [95% CI]	Brier skill	Detection 1%	Detection 2% [95% CI]
Late + GFS	0.871	0.147 [0.123, 0.171]	0.0038	0.40	0.52 [0.44, 0.59]
GFS only	0.831	0.114 [0.091, 0.134]	0.0030	0.36	0.46 [0.39, 0.53]
Early only	0.816	0.097 [0.076, 0.119]	0.0019	0.31	0.42 [0.35, 0.50]
E–D threshold (B1)	0.761	0.051 [0.037, 0.064]	—	—	—
Late, same windows as below	0.779	0.071 [0.052, 0.089]	0.0011	0.23	0.35 [0.28, 0.42]
Early over Late's windows	0.778	0.069 [0.050, 0.088]	0.0010	0.22	0.33 [0.26, 0.40]
Final + GFS (latency-free benchmark)	0.886	0.167 [0.141, 0.193]	0.0059	0.43	0.55 [0.47, 0.62]

4.4 Selective waiting provides limited economic gains

Pre-specified test (*H*3). Over the pre-specified range $\alpha \in [0.01, 0.2]$, every policy has REV below 0.07: at a base rate of ~5×10^{-4} the probabilities seldom exceed 0.01, so almost no rule warns. The VOI policy is never significantly better than the best fixed policy there and is significantly worse at four of nine α values, materially at one (−0.011; $\ell = 0$; Fig. 5). H3 is therefore not supported, but the test says little about the benefit of waiting: the range lies at 20–400 times the base rate, whereas the value of a forecast peaks where α equals the base rate (Richardson, 2000).

Exploratory range. Issuing at once with all available information reaches a peak REV of 0.55 at $\alpha \approx 4 \times 10^{-4}$ ($\ell = 0$), when one warning costs about 1/2500 of an unwarned landslide (Table 3); GFS only reaches 0.47, Early only 0.43, B1 0.34 and the latency-free benchmark 0.57. All available information beats B1 by 0.18–0.31 for $\alpha \leq 3 \times 10^{-3}$ (intervals excluding zero). At the peak the rule issues at 23% of decision times, keeps a third of the cell-time under warning and detects 91% of records (Table 4). Late + GFS reaches the same peak. Always waiting one step is worse at every α and strongly negative at small α, where most landslides in the next 6 h go unwarned.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

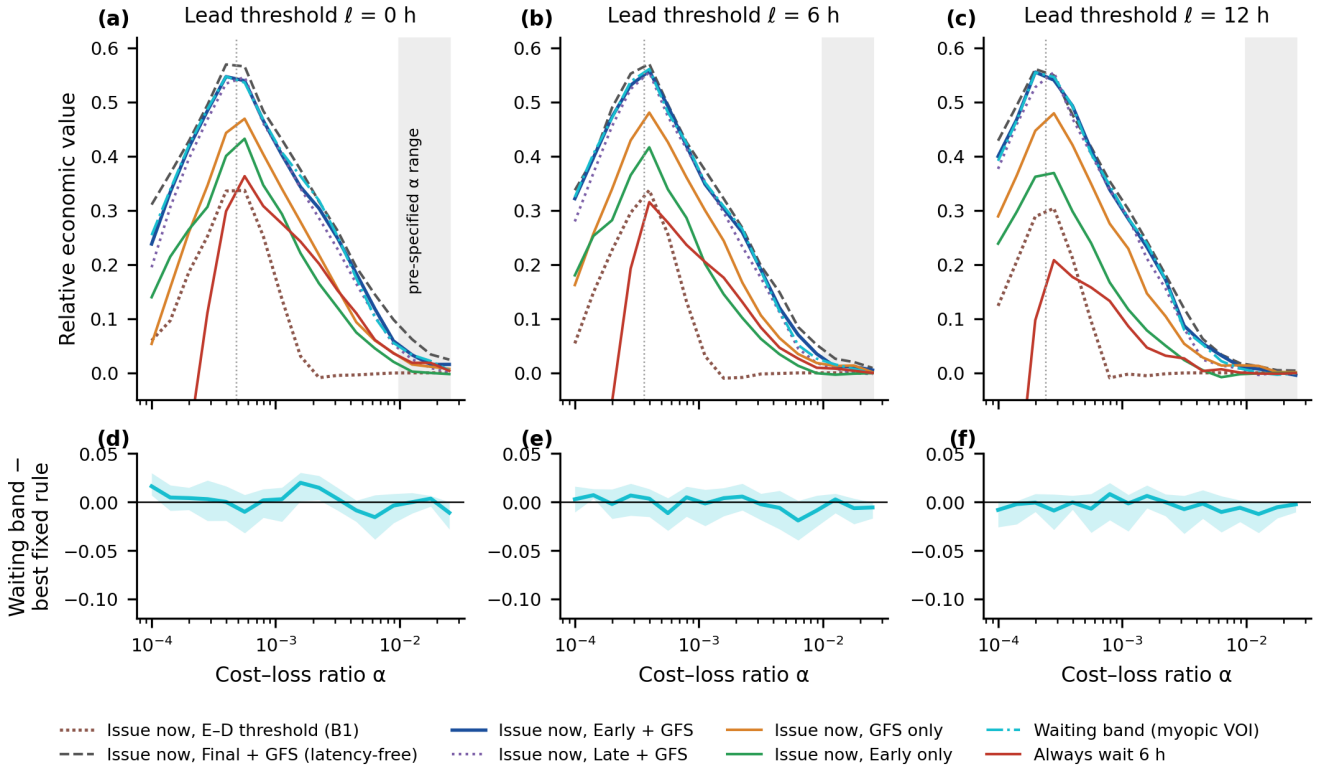


Figure 5. (a–c) Relative economic value (REV) against cost–loss ratio α for issue-now rules with five information sets and the E–D threshold B1, always waiting one step and the VOI policy, for required leads of 0, 6 and 12 h (leave-one-year-out; B1 on all decisions, the other rules on the 0.16% fewer decisions that have a successor 6 h later). Dotted vertical line: base rate; grey background: the pre-specified α ∈ [0.01, 0.2] (truncated at 0.03, beyond which all REV ≈ 0). REV of always waiting falls to −0.88 at α = 10^{-4} (ℓ = 0). (d–f) Mean REV difference between the VOI policy and the best fixed policy, with pointwise 95% confidence intervals (cyan) from 2000 paired resamples of 7-day blocks, the best fixed policy being re-selected in each resample.

Table 3. Decision performance under leave-one-year-out validation: peak relative economic value (REV) over cost–loss ratio α, and held-out detection at a 2% warned share of cell-time in the rolling replay, for required leads ℓ = 0 / 6 / 12 h. Selective waiting refers to the myopic VOI policy in the cost–loss analysis and the separately tuned threshold-band policy in the rolling replay; the two columns evaluate distinct policy families.

Policy	Peak REV, ℓ = 0 / 6 / 12 h	Rolling detection at 2%, ℓ = 0 / 6 / 12 h
Issue now, Early + GFS (all available)	0.55 / 0.56 / 0.56	0.39 / 0.36 / 0.28
Issue now, Early + GFS, recalibrated	0.55 / 0.56 / 0.55	—
Issue now, Late + GFS	0.55 / 0.56 / 0.56	0.39 / 0.36 / 0.28
Issue now, GFS only	0.47 / 0.48 / 0.48	0.33 / 0.27 / 0.22
Issue now, Early only	0.43 / 0.42 / 0.37	0.27 / 0.22 / 0.17
Issue now, E–D threshold (B1)	0.34 / 0.34 / 0.30	0.09 / 0.08 / 0.06
Always wait 6 h	0.36 / 0.32 / 0.21	—

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Policy	Peak REV, ℓ = 0 / 6 / 12 h	Rolling detection at 2%, ℓ = 0 / 6 / 12 h
Selective waiting	0.55 / 0.56 / 0.55	0.39 / 0.36 / 0.29
Issue now, Final + GFS (latency-free benchmark)	0.57 / 0.57 / 0.56	0.42 / 0.37 / 0.31

Table 4. Warning load and detection for the issue-now policy with all available information at selected cost–loss ratios α (ℓ = 0 unless stated). Columns give REV, the share of decision times with an issued warning, the share of cell-time under warning when each issue lasts 24 h (April–October), detection with lead > 0 h and > 6 h, and the share of decisions deferred by the VOI policy.

α	REV	Issues at (% of decisions)	Warned (% of cell-time)	Detection, lead > 0 h	Detection, lead > 6 h	Waiting band defers (% of decisions)
1.0×10^{-4}	0.24	51.2	61.0	0.97	0.96	1.9
4.0×10^{-4} (peak REV)	0.55	23.5	32.7	0.91	0.85	2.7
1.6×10^{-3}	0.34	6.6	11.0	0.67	0.61	1.2
4.5×10^{-3}	0.18	1.6	3.1	0.49	0.37	0.4
1.3×10^{-2} (pre-specified range)	0.03	0.3	0.6	0.19	0.10	0.2
5.0×10^{-2} (pre-specified range)	0.00	< 0.1	< 0.1	0.01	0.00	< 0.1

The VOI policy stays close to the best fixed policy (Fig. 5d–f). At ℓ = 0 it is materially better at three α values, by +0.016 [+0.007, +0.030] at $\alpha = 10^{-4}$, +0.020 [+0.001, +0.030] at 1.6×10^{-3} and +0.015 [+0.002, +0.027] at 2.2×10^{-3}, and never at ℓ = 6 or 12 h. Averaged over the exploratory range, the difference is +0.002 [−0.005, +0.007] at ℓ = 0 (−0.001 [−0.007, +0.002] and −0.003 [−0.010, 0.000] at ℓ = 6 and 12 h), and the simultaneous band lies above zero at no α (Table C2). The single gains move between validations (Table C1). With province hold-outs, T3/T4 labels, times shifted 3 h earlier and the 2024 real-time products the band is never materially better. In the other variants it is materially better at one or two α values (at most +0.023, fully nested) and materially worse at up to four, by up to 0.053. In no variant or lead does the mean difference over the range have an interval above zero, and the simultaneous band lies above zero at one α in 7 of 51 variant–lead combinations, with lower limits of at most 0.003. The upper 95% limit of the pointwise difference is at most 0.033 in the main and nested analyses and below 0.04 elsewhere, except in the small 2024 subset (0.058 at ℓ = 6 h).

Two of the three gains lie where the probabilities over-predict. Overall the window probabilities are calibrated (Table C3), but for $\boldsymbol{p_{AB}}$ between 10^{-3} and 3.2×10^{-3} the observed frequency is 0.76 [0.59, 0.93] of the predicted one, so issuing at once warns too readily there. Recalibrating the issue-now rule within the cross-validation brings this ratio to 0.87 [0.67, 1.07] without changing the log loss, and with the recalibrated rule among the fixed policies all three gains disappear. Only an immaterial gain of +0.007 remains at ℓ = 6 h. Coherent window probabilities from one sub-window model give the same picture (Table C1; Appendix C).

Figure 6a shows why waiting stops paying at larger α. On the decisions the VOI policy deferred, the realized cost of waiting exceeds the cost of issuing at once for $\alpha \geq 4.5 \times 10^{-3}$, by 3–24%, and about a third of it comes from landslides in the next 6 h. The models predicted the probability of such landslides well (2.9×10^{-3} against an observed 3.2×10^{-3} at $\alpha = 8.9 \times 10^{-3}$) but

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

underestimated the expected cost after waiting (4.2×10^{-3} against 6.7×10^{-3}). The option to wait is worth most when the current information lacks a forecast (Early only: up to +0.04, with intervals excluding zero at five $\alpha \leq 8 \times 10^{-4}$; Fig. 6b). With the GFS forecast it is worth at most 0.02, because the forecast already carries much of what waiting would reveal.

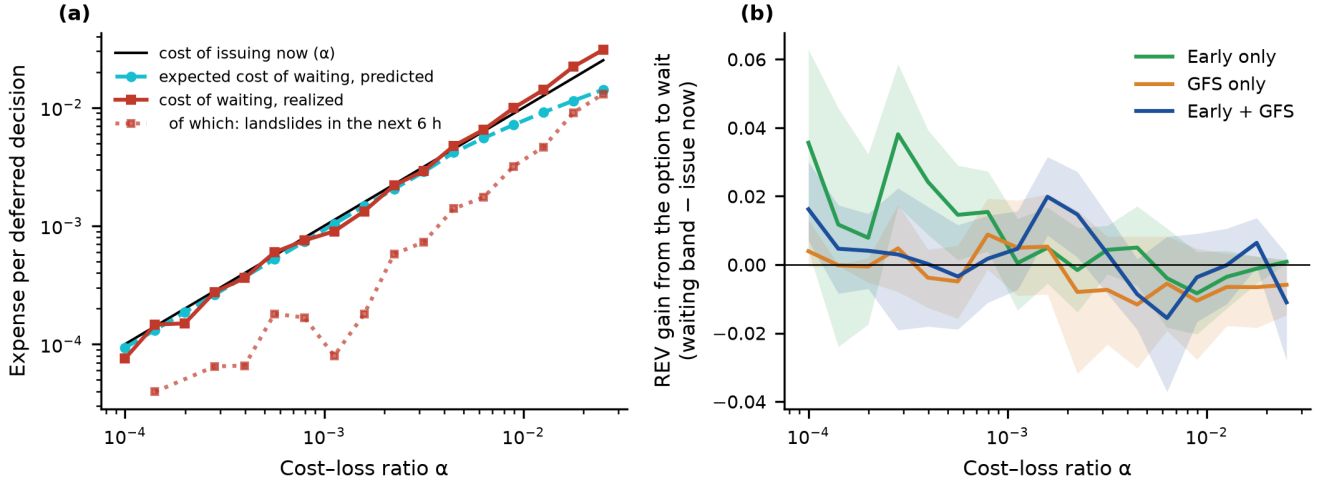


Figure 6. (a) On decisions the VOI policy deferred (ℓ = 0), the cost of issuing at once (α) against the predicted and realized expected cost of waiting, $p_A + \mathbb{E}\min(\alpha, q)$, and the realized part due to landslides in the next 6 h. (b) Value of the option to wait (VOI policy minus issue-now REV) when the information set is Early only, GFS only or Early + GFS. Same-colour translucent bands around the green (Early only), orange (GFS only) and blue (Early + GFS) curves show pointwise 95% confidence intervals for their paired REV differences, from 2000 bootstrap resamples of 7-day blocks. The horizontal line marks zero gain.

Among decisions where the issue-now rule warns, the realized gain rises across deciles of predicted VOI, with a top-minus-bottom difference whose interval excludes zero at all α up to 1.3×10^{-2} (ℓ = 0). This post hoc subset contrast provides exploratory evidence of ranking ability, rather than confirmation of the pre-specified H2 test over all decisions or accurate estimates of gain magnitude. For $\alpha \geq 4.5 \times 10^{-3}$, however, the realized gain is negative in most deciles (5–10 of 10), including some with positive predicted VOI. At ℓ = 0, the Spearman correlation among warning decisions is positive (ρ = 0.14–0.49). Over all 2 243 653 decisions it depends on α: ρ = −0.036 [−0.044, −0.027] at $\alpha = 10^{-4}$, −0.125 [−0.129, −0.120] at 5.6×10^{-4}, and +0.034 [+0.030, +0.038] at 1.3×10^{-2}. Table E1 gives all α–lead combinations, including undefined estimates at the largest α values. The original H2 p-value criterion was not evaluated as a single confirmatory test, and the post hoc subset findings do not establish it.

4.5 No stable detection advantage at matched warning budgets

The rolling replay shows no stable detection advantage for the waiting band with all available information (Fig. 7). With thresholds chosen on the other years, the band detects a similar share of records to the single threshold at every budget and lead (differences −0.011 to +0.014, none significant), although its candidate family contains the single rules. With a denser grid, nested threshold selection or both, the differences are −0.014 to +0.011, never significantly positive and twice

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

significantly negative. Averaged over the five budgets, the difference is +0.001 [−0.005, +0.007] at ℓ = 0 (−0.003 [−0.009, +0.003] and +0.001 [−0.006, +0.008] at ℓ = 6 and 12 h; Table C2), and single budgets have upper 95% limits of at most +0.041 (+0.028 with nested selection). For GFS only, the band is significantly better in one of 60 budget–lead–variant combinations and significantly worse in six.

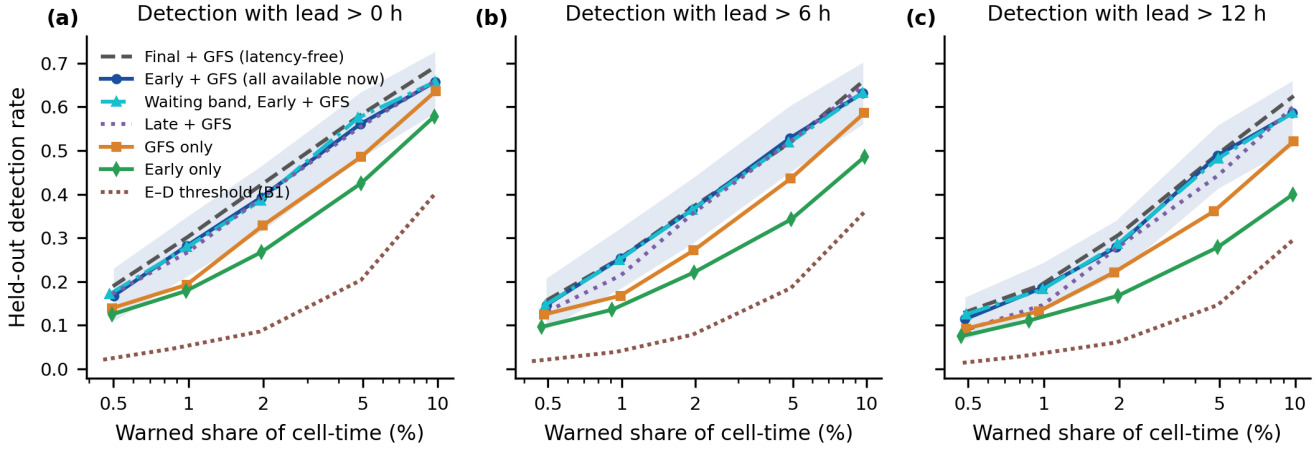


Figure 7. Rolling 24 h warning replay: held-out detection rate against warned share of cell-time, for lead thresholds ℓ = 0, 6 and 12 h; detection requires a lead strictly greater than ℓ. Thresholds were chosen on the other nine years for each budget (0.5, 1, 2, 5, 10%). The light-blue translucent area is the pointwise 95% confidence interval from 2000 rainfall-event bootstrap resamples for the Early + GFS single-threshold policy. Other policies are shown without confidence bands. Not shown (off scale): the 5% E–D threshold at its conventional operating point detects 0.92, 0.90 and 0.87 of records (ℓ = 0, 6, 12 h) with 60% of cell-time under warning.

The information sets keep their order. All available information detects 0.06–0.09 more records than GFS only at budgets of 1–5% (ℓ = 0 and 6 h), 0.08–0.19 more than Early only at budgets of 1–10%, and 0.15–0.36 more than B1 at budgets of 0.5–10% (ℓ = 0; 0.14–0.36 nested; all intervals excluding zero). Late + GFS and Early + GFS are indistinguishable at ℓ = 0, and the latency-free benchmark adds at most 0.04. At its conventional operating point, the 5% E–D threshold detects 0.92 of the records but keeps 60% of cell-time under warning, because long monsoon rain events exceed it for most of the season.

The archived daily LHASA version 2.0.0 baseline detected fewer records than Early + GFS at every tested budget and lead on the matched 2015–2020 decisions, by 0.14–0.29 at ℓ = 0 and 0.11–0.31 across leads (all intervals excluding zero), and its peak REV was 0.34 against 0.55 (ℓ = 0). Using the sampled-pixel maximum with the threshold of 0.5 selected for this study, it kept 1.1% of cell-time under warning and detected 0.04 [0.01, 0.06] of the records, whereas Early + GFS detected 0.26 within a 1% budget. Adding this LHASA baseline to Early + GFS changed detection by −0.02 to +0.02 across the tested settings, without a consistent gain. These results concern this archived daily configuration and do not evaluate the historical or current LHASA operational service.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

4.6 The conclusions hold across validations, products and populations

Province hold-outs, labels from the 213 hour- and minute-level (T3/T4) records only, random timing errors of ±1 h and ±3 h, systematic shifts of ±3 h and LightGBM in place of logistic regression leave the conclusions unchanged (Table C1): the log-loss gain of adding observations to forecasts stays significant, peak REV moves by less than 0.02, and the waiting band is materially better than the best fixed policy at no more than two α values. LightGBM gave no better probabilities (AUC 0.871 against 0.875 at $\ell = 0$).

On the real-time V07 products of June–October 2024 (28 records), the models trained on the retrospective record ranked decisions at least as well as in the main analysis (AUC 0.917 at $\ell = 0$) and the policies kept their order. The waiting band was never significantly better than the best fixed policy and was significantly worse at 10–15 of 23 α values. The log-loss gain of all available information over GFS only kept its sign but, with so few records, not its significance. Degrading the observations weakened them only slightly ($\sigma = 0.6$: AUC of Early only 0.796 instead of 0.804 at $\ell = 6$ h). The gain over GFS only stayed significant at every lead, and the waiting band was materially better at no more than two α values per lead.

Adding the 1000 control cells raises the number of decisions to 10.8 million and lowers the weighted base rate to 2.4×10^{-6} ($\ell = 0$), which moves the cost–loss analysis to smaller α. The ordering of the information sets is unchanged (Table C1): issuing at once with all available information reached an AUC of 0.929 and a peak REV of 0.69, against 0.57 for GFS only, 0.49 for B1 and 0.71 for the latency-free benchmark. REV is higher than for record cells alone, probably because the wider population adds contrast between wet and dry places. The waiting band was materially better than the best fixed policy at two of 43 α values (at most +0.017) and materially worse at four. In the rolling replay it again detected a similar share of records to the single threshold (differences −0.018 to +0.014), and all available information detected 0.05–0.11 more records than GFS only and 0.21–0.41 more than B1 ($\ell = 0$ and 6 h). Without sampling weights the results were the same (peak REV 0.68; band materially better at two α values, at most +0.012).

4.7 Independent test of waiting without a forecast

Both pre-specified hypotheses were supported (Table C4). With Early only, the VOI policy beat issuing at once by +0.029 [+0.009, +0.052] in REV averaged over $\alpha = 10^{-4}$–8×10^{-4} at $\ell = 0$ (C1), significantly at five of the seven α values and close to the +0.023 [+0.009, +0.037] of the same analysis in 2015–2024. Always waiting one step did much worse (−0.67 [−0.81, −0.53]; C2). The frozen models ranked the 2008–2014 decisions less well than in the main analysis (AUC 0.78 for $\boldsymbol{p}_{\mathbf{AB}}$, log-loss skill 0.06) and under-predicted their level (mean probability 0.56 of the base rate of 7.0×10^{-4}), so absolute REV values are lower and negative at the smallest α. At $\ell = 6$ and 12 h the advantage was smaller and not significant (+0.015 [−0.004, +0.036] and +0.002 [−0.008, +0.014]), as in 2015–2024. In an independent sample, the method thus detects a value of waiting of about 0.03 in REV when the current information lacks a forecast.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

5 Discussion

5.1 Why information gains yield limited decision benefits

The information and decision analyses answer different questions. The case-crossover and lead-time analyses condition on a landslide at T and ask how well the data available at T − ℓ single out T. The decision analysis asks what to do at t without knowing whether or when a landslide will occur. Information improves sharply over the last 6–12 h before a landslide as the triggering rainfall becomes observable. For the same reason, whenever the evidence is strong enough to consider a warning, much of the risk lies in the next six hours, and waiting improves the estimate while leaving that interval unprotected. On the decisions the waiting band deferred at $\alpha \geq 4.5 \times 10^{-3}$, about a third of the realized cost of waiting came from landslides in the next 6 h (Fig. 6a), a cost that was observed, not modelled.

Errors in estimating future costs also limit the benefit: on the same decisions the models underestimated the expected cost after waiting, and a better preposterior model could change the value of waiting at some α. The few cost–loss ratios with material gains are scattered over more than an order of magnitude, move between validations and vanish once the issue-now rule is recalibrated, so waiting there mainly corrected over-warning. The method can detect gains of this size: without a forecast, the same rule gained about 0.03 in REV in the independent 2008–2014 sample (Sect. 4.7). LightGBM extracted no more from the same rainfall predictors. New predictors such as soil moisture, or multi-step rules, may find more. Jewson et al. (2021) found that waiting for the next forecast can pay for temperature-sensitive decisions. For landslides, the next forecast cycle carried no measurable extra information, and the observations gained by waiting rarely offset the lead time lost.

5.2 Implications for rainfall inputs and warning lead time

Three practical points follow. First, the earliest satellite run is best merged with the current forecast: Early + GFS was the best feasible information set, or level with Late + GFS, in every analysis while being available 10 h earlier, and the GFS lead-6 bucket usefully fills the hours that Early has not yet reported. An E–D threshold on the same satellite data performed much worse, as a ranking (REV 0.34 against 0.55) and as a fixed threshold with a high warning load. Thresholds from denser gauge networks may do better (He et al., 2020, give national E–D thresholds for China), satellite products underestimate triggering rainfall unless thresholds are recalibrated (Brunetti et al., 2018), and data-driven satellite nowcasts have outperformed threshold-like rules elsewhere (Stanley et al., 2021). The archived LHASA version 2.0.0 daily baseline, which combines Early rainfall with soil moisture and static predictors in a globally trained model, ranked the matched 2015–2020 decisions less well than Early + GFS and gave no consistent additional improvement when combined with it. This result applies to the archived daily configuration tested here; a regionally calibrated or sub-daily version may do better. Second, the tested comparisons gave little evidence of a material benefit from waiting for IMERG Late or Final: even the latency-free benchmark improves REV by only about 0.02 and detection by at most 0.04. Third, lead time is the scarce

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

resource. Detection at a 2% budget falls by more than a third between lead thresholds of 0 and 18 h (0.52 to 0.33), so convection-permitting or ensemble forecasts merit evaluation as a source of lead time.

5.3 The operational warning chain

Our results bear on four links of the chain from forecast to warning to action. The first is the decision cycle: waiting for a later satellite run or a gauge-corrected product added at most about 0.02 in REV and cost hours of lead time. For China's national service, whose models increasingly use dynamic rainfall inputs (Xiao et al., 2024), this suggests testing the earliest run plus the current forecast against later products, using the real-time products the service actually receives.

The second link is the warning load. At the peak REV, warnings occupy a third of the cell-time (Table 4), a load suited only to low-cost actions such as heightened duty, patrols or public advisories. Costlier actions sit at larger α: at $\alpha = 1.6 \times 10^{-3}$ warnings cover 11% of cell-time and detect 67% of records (61% more than 6 h ahead), and at $\alpha = 4.5 \times 10^{-3}$ they cover 3% and detect 49% (37%). For loads a service can sustain, the rolling replay at budgets of 0.5–10% is the relevant evidence, and there waiting gave no stable detection advantage (Sect. 4.5).

Early action depends on where lead time comes from. Lead time comes from the forecast, and detection falls by more than a third between lead thresholds of 0 and 18 h, so actions that need many hours, such as evacuation or road closures, may require forecast-based probabilities with thresholds chosen for the required lead time and warning budget. The observation-driven update a few hours later can then escalate or cancel the warning, in the spirit of forecast-based action triggers (Coughlan de Perez et al., 2015). Combined with exposure, such cell-level probabilities would support impact-based warnings (WMO, 2015).

A latency-aware replay is a practical way to evaluate changes to the timing of a warning chain retrospectively, complementing event-based performance analyses that weigh warned time against detected landslides (Calvello and Piciullo, 2016). Testing its implications in a running service is the next step.

5.4 Limitations

The landslide labels come from news reports. Reporting bias that is constant at a cell cancels in the case-crossover comparisons, but reporting that increases with rainfall within a month would inflate all rainfall-based skill there. The decision models pool cells without a cell term, and the control cells keep unreported landslides as negatives, so absolute probabilities, detection rates and REV values describe reported landslides. For a fixed set of warning actions, if every true positive window were reported independently with the same probability r, the expected expense against reported events at α would be r times that against all events at α/r, and rankings of these fixed policies would carry over after rescaling α; normalized REV values need not be identical. Reporting that differs between cells or rises with rainfall breaks this correspondence and cannot be detected with these data.

The population matters as well. The weighted sample represents the qualifying record cells, including 22 with mean slope below 5°, together with the no-record terrain cells of mean slope ≥ 5°. It excludes cells whose only records fail the eligibility

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

criteria (Sect. 3.7). This reporting-stratified population is one of several possible frames and does not cover the full area a regional service warns, which also contains other flat, urban and arid cells and places whose landslides go unreported. The absolute values are therefore not the performance of China's operational service. The relative comparisons were consistent across both evaluated populations.

Random timing errors of up to ±3 h and systematic shifts of ±3 h did not change the results. Larger or location-dependent errors were not tested. With 193 rainfall events, uncertainty remains substantial: pointwise upper confidence limits on waiting gains in the main analyses were about 0.04 in REV or detection, while the small 2024 subset reached 0.058 in REV (Sects. 4.4–4.5). Smaller gains remain compatible with the data.

The IMERG record is the V07 retrospective reprocessing, not the products distributed in real time in 2015–2023 (Sect. 2.2). Its climatological calibration to the Final run may make the retrospective Early and Late more accurate than their real-time predecessors. A more accurate record raises the value both of the observations available now and of those that waiting would bring, so the direction of the bias on the value of waiting is unknown, but it may inflate channel 3 and the advantage of Early + GFS over GFS only. The earlier real-time versions are no longer distributed. On the real-time products of June–October 2024 the conclusions held, and with observations degraded by day-to-day errors the gain of observations shrank but did not vanish (Sect. 4.6). The 2024 sample is small (28 records), and degradation does not reproduce systematic biases, so the absolute value of channel 3 for operators in 2015–2023 remains uncertain.

The probability models use a few time-varying rainfall predictors without static susceptibility factors; LightGBM gave the same conclusions, but other predictors could change absolute values. We studied the myopic one-step rule. Multi-step rules were not pursued after H3 was not supported, and GFS forecasts were extracted only up to 30 h. The cost–loss evaluation treats overlapping windows as separate decisions, and the rolling replay, which does not, measures cost as warned share rather than money. The models of China's national warning service are not public, so the comparison with the system that issues China's warnings remains open.

6 Conclusions

Using news-reported landslides with minute-, hour- or period-of-day times and a latency-aware replay, we separated three ways in which waiting can improve the evidence for a regional landslide warning in China. A newer forecast cycle adds nothing measurable, a later satellite run adds a little, and observation replacing forecast adds much more, so skill rises steeply as the landslide approaches. Waiting for these information gains gave no stable additional decision benefit when Early rainfall and GFS forecasts were already combined. At matched warning budgets, waiting with satellite rainfall plus forecast gave no stable detection advantage. In the cost–loss analysis, the few gains of a value-of-information waiting rule, at most 0.02 in REV, did not survive province hold-outs, hour- and minute-level records only, a gradient-boosting model or the 2024 real-time products, were no longer material relative to the best fixed policy after recalibration, and averaged about zero. The pre-specified cost–loss test could not discriminate, because almost no rule warns in its range, but a second test, fixed

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

before an independent 2008–2014 sample was assembled, confirmed that the same rule finds value in waiting when no forecast is available. Issuing at once with satellite rainfall plus forecast outperformed a conventional rainfall threshold and, on matched 2015–2020 decisions, the archived daily LHASA version 2.0.0 baseline under the tested configuration. Under the tested inputs and one-step policies, better rainfall information alone provided no stable empirical basis for delaying a warning when satellite rainfall and forecasts were already available.

The decision results rest on 193 rainfall events and on reported landslides, and they come from a retrospective replay with a reprocessed satellite record rather than from an operational trial (Sect. 5.4). For warning chains like the one replayed here, combining the earliest satellite rainfall with the current forecast and deciding at once is the default to test in operation, and lead time is better sought in forecasts than in later data.

Appendix A: Model fitting and policy evaluation

Occurrence-time alignment. For T2 records, the reported part of day was assigned a fixed, deliberately broad local-time window before taking its midpoint: midnight 00:00–01:00; pre-dawn 00:00–06:00; dawn or early morning 05:00–09:00; morning 06:00–12:00; "since morning" 06:00–24:00; noon 11:00–14:00; afternoon 12:00–18:00; dusk 17:00–20:00; evening 18:00–24:00; night 18:00–06:00 the following day; and late night 22:00–02:00 the following day. T3 records were represented by the start of the reported hour and T4 records by the reported minute.

Decision-window boundaries. Probability labels and detection use the same open lower boundary: a warning at t can detect a record at T only when $\ell < T - t \leq 24$ h. This convention excludes a record exactly at $t + \ell$, including one simultaneous with issuance when $\ell = 0$. An audit of the 280 occurrence timestamps identified 15 exactly on the six-hourly UTC decision grid. The implemented exclusion is retained for these boundary cases in all decision and rolling-replay results.

Event grouping for paired comparisons. The complete-data 24 h and 6 h comparisons retain 279 records in 192 events and 213 records in 156 events, respectively. The daily resolution control retains 280 records in 183 events because it assigns event membership on the larger rainfall-record pool, including date-only T1 records, before selecting the 2015–2024 T2–T4 sample. It uses the same 72 h and 300 km rule, but additional records can connect clusters; its cross-validation and intervals retain that grouping. The plotted 24 h placebo retains 275 records in 190 events after the time shift and complete-data restriction.

Probability models. Each rainfall total x is transformed as $\log(1 + x)$ and standardized before fitting a logistic regression. Here log is the natural logarithm. Models use an L2 penalty with inverse regularization strength $C = 1$ and are optimized by the limited-memory Broyden–Fletcher–Goldfarb–Shanno method (L-BFGS), with gradient tolerance 10^{-10}. A stricter tolerance than the default 10^{-4} is needed for convergence at base rates of 10^{-4}–10^{-6}. With control cells, sampling weights are rescaled to mean 1 for fitting. The LightGBM variant is constrained to increase monotonically with every rainfall total (learning rate 0.03, 15 leaves, at least 200 decisions per leaf) and is trained leaving one year out, with the following year used for early stopping.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

LHASA mapping. Each 0.1° cell spans 12 × 12 native LHASA pixels. Sampling every fourth row and column (offsets 0, 4 and 8 from the northern and western edges) gives nine spatially distributed samples, rather than nine contiguous pixels. Values outside [0, 1] are discarded; the maximum and mean use the remaining valid samples, and a cell without valid samples is missing. The latest available LHASA day is the UTC day preceding the day containing t − 5 h. Each aggregate probability is clipped to $[10^{-6}, 1 - 10^{-6}]$, logit-transformed and standardized using the training fold. Logistic models use C = 1 and tolerance 10^{-10}, with six held-out-year folds for 2015–2020. Early + GFS uses its usual log(1 + x) rainfall transformation on exactly the same decisions. Only common complete-feature decisions between 5 April 2015 and 31 October 2020 at 18:00 UTC enter; no additional decisions in this period were lost to missing LHASA values. The LHASA-only, Early + GFS and combined models are fitted separately for each target window.

Preposterior distribution. For a held-out year k, decisions from the other years were grouped into 40 quantile bins of logit p_B (where logit p = log[p/(1 − p)] for probability p) and four classes of the 24 h GFS forecast (below 0.1 mm, then terciles of the wet decisions), and $\mathbb{E}\min(\alpha, q)$ is the mean of min(α, q) within the group. In the main analysis the probabilities of the other years come from their own leave-one-year-out models, which saw the held-out year, and the GFS classes are cut on the forecasts (not the labels) of all years. In the fully nested version, for each held-out year k the bins and q of every other year j (j ≠ k) come from models trained without both j and k (90 models per window), and the GFS classes are cut on the other years' forecasts.

B1 mapping. For a held-out year k, the training decisions of every other year j were scored with an E–D threshold fitted without both j and k, so that year k's records entered neither the threshold nor the mapping.

Recalibrated issue-now rule. For each held-out year k, a logistic regression on a cubic B-spline of logit p_{AB} (knots at probabilities p = 10^{-5}, 10^{-4}, 3 × 10^{-4}, 10^{-3}, 3 × 10^{-3}, 10^{-2} and 10^{-1}, linear extrapolation) was fitted to the labels of the other years, using their probabilities from models trained without both that year and k, and applied to year k. It changed the log loss by +2.8 × 10^{-6} [−5.6, +10.9] × 10^{-6} per decision ($\ell = 0$). An isotonic map gave the same conclusions. A spline map fitted on the ordinary leave-one-year-out probabilities, whose models saw the held-out year, worsened the log loss (+1.6 × 10^{-5} [+0.6, +2.6] × 10^{-5} per decision) and left one material gain of the waiting band.

Sub-window model. For each 6 h sub-window S_m = (t + 6(m − 1) h, t + 6m h], with index m = 1–4, a logistic model using all available information estimates h_m, the probability of at least one reported landslide in S_m given information at t. A window probability P is constructed as $P = 1 - \prod_m (1 - h_m)$, where $\prod_m$ denotes the product over the sub-windows belonging to that window. This coherence-enforcing construction uses a conditional-independence approximation between sub-window events given the information at t; the fitted probabilities are not survival-conditioned hazards. The next-step probability q is constructed from the same models evaluated at t + 6 h. At $\ell = 0$ these probabilities have a slightly lower log loss than the separate window models (−3.2 × 10^{-6} [−7.1, −0.1] × 10^{-6} per decision). The VOI policy, oracle reference and fixed policies were re-evaluated with both the main and fully nested preposterior groups.

Oracle reference and diagnostics. The oracle reference waits when $p_A + \min(\alpha, q) < \min(\alpha, p_{AB})$, using the realized next-step probability q as if it were known before waiting. It is neither feasible nor an upper bound on decision value: it combines

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

probabilities conditioned on different information, and separate window models need not satisfy $p_A \le p_{AB}$, $p_B \le p_{AB}$ and $p_{AB} \le p_A + p_B$. We compared observed and predicted frequencies overall and in half-decade probability bins, and counted violations of these conditions (Table C3).

Rolling replay. Single thresholds were searched over 200 quantiles of the current probability p_t, and band parameters over a 12-level quantile grid (0.90–0.9995), or 20 levels in the denser variant. In the nested variant the thresholds for each held-out year were chosen on probabilities of the other years from models trained without that year. For B1, the E–D threshold and the mapping were also fitted without that year.

Overall tests. For a paired difference D(α) with bootstrap replicates D_b(α) and bootstrap standard deviation s(α), the simultaneous band over a range of α values is D(α) ± c s(α), where c is the 95th percentile over resamples of max over α of |D_b(α) − D(α)|/s(α); the mean over the range uses the same resamples. The rolling-replay summaries treat the five budgets in the same way, with rainfall-event resamples.

Perturbations. Random timing errors add a draw from the uniform distribution U(−h, h) to each occurrence time, where h is the maximum absolute error, set to 1 or 3 hours; systematic perturbations shift all times by ±3 h. Labels are reconstructed for the perturbed occurrence times, rainfall features remain unchanged, and all models are refitted. Observation degradation multiplies every half-hourly Early and Late rate for a cell and UTC day by exp(σz − σ²/2), where exp is the exponential function, z is a standard normal variate (z ~ N(0, 1)) and σ is the log-scale standard deviation. The same factor is applied to both runs because they derive from the same retrievals. IMERG and GFS series for control cells are extracted in the same way as for record cells.

Appendix B: Analysis plan and deviations

B1 Original plan and stopping criteria

The internal plan was first written on 9 October 2026 (version v0.1), before data download or model training; the dates below are internal records. This appendix reports the criteria and analysis history needed to distinguish pre-specified from exploratory evidence.

The four hypotheses are stated in Sect. 3.8. The original H1 criterion was a 95% rainfall-event bootstrap interval for the Brier-score improvement of Late over Early excluding zero. H2 specified a positive Spearman correlation (ρ > 0, with a significance-test p value below 0.05), but did not identify a primary α–lead combination or an across-grid confirmation rule. The implemented analysis reports pointwise block-bootstrap intervals rather than that nominal p-value test (Appendix E). H3 required higher REV than the best fixed policy over at least half of α ∈ [0.01, 0.2], and H4 required that the direction of this advantage not reverse when a province was held out. The planned temporal split was training in 2015–2019, validation in 2020–2021 and testing in 2022–2024, initially with monotone LightGBM and isotonic calibration.

Five stopping or restriction criteria were recorded. First, the primary design required at least 200 independent rainfall events with period-level or better times, county-level or better locations, and coverage of at least five provinces; the catalogue met it

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

exactly (287 records, 200 events in 23 provinces before the April–October restriction). Second, if the 95% interval for the Late–Early Brier improvement included zero and a newer GFS cycle also showed no significant gain, the design was to be reconsidered or restricted to reporting the information comparison. Third, failure to outperform the best fixed policy over at least half of the original α range would stop extension to multi-step rules. Fourth, loss of an advantage after removing overlap with the National Aeronautics and Space Administration (NASA) Global Landslide Catalog (GLC) or stratifying by population density would restrict its interpretation. Fifth, identified future-information leakage would require correction and repetition of the affected analyses.

B2 Dated revisions and changes in scope

On 10 October 2026, version v0.2 recorded the feasibility audit, measured GFS publication delays, April–October restriction and time-precision classes. Version v0.3 added time-stratified case-crossover comparisons, a seven-day placebo shift, the alternative referent window and cell-level decision replay. The proposed cell-by-month conditional risk model was subsequently replaced by pooled logistic probability models. The final main sample included both T2 and T3/T4 records, with the T3/T4-only sample used for sensitivity analysis, reversing the primary and sensitivity roles proposed in version v0.3. Version v0.4, also dated 10 October, organized the study around the three channels of information gained by waiting, with decision evaluation focused on observations replacing forecasts. This third channel was not covered by the original stopping criterion for processing-run and forecast-cycle improvements.

Version v0.5, dated 11 October 2026 after H3 was not supported, retained the one-step analysis, reinterpreted H4 as stability of the decision conclusions, and added the time-precision check and rolling warning replay. The later independent test was frozen separately on 14 October (Appendix D).

The initial design considered antecedent wetness from the ERA5-Land land reanalysis and static susceptibility predictors from elevation, land cover and lithology. These were not included in the final rainfall-only probability models; terrain was used only to define the control-cell frame. GLC records added few usable daily records and did not become additional labels for the six-hourly decision analysis. Population-density stratification was deferred and was not performed; the control-cell analysis does not substitute for it. The initial preposterior proposal used quantile regression, whereas the final analysis used empirical probability groups. The original always-wait baseline referred to waiting for Late; the implemented baseline instead waits one six-hour step, so it tests the next decision opportunity rather than guaranteed arrival of the Late run. The metrics ultimately emphasized proper probability scores, REV and warning-budget detection, rather than the full set of discrimination, calibration and categorical diagnostics proposed initially.

B3 Departures after inspecting results

The following departures from the initial analysis plan are reported together with their timing:

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

1. Leave-one-year-out became the main evaluation after the first decision-model results on the pre-specified split had been seen; that test period contains only 55 records and its intervals could not separate the information sets. Results on the pre-specified split are reported.
2. The log loss became the main score after the same results showed that Brier-score differences were of order 10^{-8} at a base rate of ~5×10^{-4}. The Brier score is reported.
3. The pre-specified range $\alpha \in [0.01, 0.2]$ was set without reference to the base rate and kept when the cell-level decision design was adopted, before any decision model was trained; at about 5×10^{-4} per decision, the base rate lies far below the range. The α range was widened to 10^{-4}–0.2 after the decision-model results and before the policy results; H3 is reported on both ranges.
4. After the case-crossover results, the analysis was organized around the three channels, with the decision analysis centred on channel 3, which the pre-specified stop conditions did not cover.
5. After H3 was not supported, the analysis remained focused on one-step decisions without extending to multi-step policies, and H4 is read as the stability of the decision conclusions under leave-one-province-out validation. The rolling warning replay and the T3/T4-only labels were specified at the same point, before they were run.
6. The following analyses were specified after the main results were known, in this order: (a) the B1 baseline (pre-specified but not run until then), the oracle reference, the exploratory warning-subset decile analysis related to H2 and the random timing perturbation; (b) the 2024 real-time subset, the observation degradation, the control cells and the materiality criterion for REV differences (Sect. 3.7); (c) the systematic timing shifts, the LightGBM variant of the decision analysis and the translation of cost–loss ratios into warning load (Table 4); (d) the fully nested evaluation of the preposterior bins and of the rolling-replay thresholds, the single-threshold rules within the band family and the denser band grid (Sects. 3.5–3.6); (e) the calibration and coherence diagnostics of the window probabilities (Table C3), the recalibrated issue-now rule and the sub-window model (Sect. 3.5).
7. After all of these, and before the 2008–2014 decision data were assembled, the second confirmatory test of Sect. 3.8 was written and frozen; it was evaluated once according to the frozen plan.
8. After the 2008–2014 test, the LHASA baseline B2 on 2015–2020 (Sect. 3.4) and the overall tests across α values and budgets (Sect. 3.7) were added. A subsequent end-to-end recheck covered the analyses listed in the archived comparison record. Most checked result files matched exactly; small numerical differences occurred in the pre-specified-split Early-only LightGBM results at a 12 h lead threshold. The archived comparison record marks the LHASA baseline as not rerun; its separately archived script and results document the original calculation. The recheck therefore does not cover a repeat of every analysis (see Code and data availability).

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix C: Robustness of the decision conclusions

Table C1. Robustness of the decision conclusions (ℓ = 0). Log-loss differences per row with 95% CIs; peak REV over α; number of α values (of 23; of 43 with control cells, whose grid extends to 10^{-7}) at which the VOI policy is significantly better or significantly worse than the best fixed policy (95% interval excluding zero), and in brackets how many of these differences are also material (|mean| > 0.01). The largest material gain in any row is +0.023 (fully nested), the largest material loss −0.053 (2024 subset). In the nested row only the preposterior bins change, so the log-loss differences equal those of the main analysis; in the recalibration row only the set of fixed policies changes, and in the sub-window rows only the probabilities of all available information. With control cells, log-loss differences are on the scale of the much lower weighted base rate.

Validation	Log loss: all available − GFS only	Log loss: Late + GFS − all available	Peak REV: all available / GFS only / Early only / always wait / VOI policy / latency-free benchmark	VOI policy vs best fixed policy: α values significantly better (material) / significantly worse (material)
Leave one year out	−1.66e-4 [−2.28e-4, −1.06e-4]	+2.6e-5 [+5.7e-6, +4.7e-5]	0.55 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	3 (3) / 4 (1)
Leave one year out, fully nested bins	−1.66e-4 [−2.28e-4, −1.06e-4]	+2.6e-5 [+5.7e-6, +4.7e-5]	0.55 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	2 (2) / 4 (1)
Leave one year out, recalibrated issue-now rule among the fixed policies	−1.66e-4 [−2.28e-4, −1.06e-4]	+2.6e-5 [+5.7e-6, +4.7e-5]	0.55 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	0 (0) / 5 (2)
Leave one year out, coherent sub-window probabilities	—	—	0.55 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	0 (0) / 3 (0)
Leave one year out, coherent sub-window probabilities, fully nested bins	—	—	0.55 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	1 (1) / 3 (0)
Leave one province out	−1.58e-4 [−2.19e-4, −9.8e-5]	+2.3e-5 [+3.0e-6, +4.4e-5]	0.54 / 0.47 / 0.42 / 0.36 / 0.55 / 0.57	0 (0) / 4 (1)
T3/T4 labels only	−1.16e-4 [−1.64e-4, −6.8e-5]	+1.4e-5 [−1.7e-6, +3.0e-5]	0.54 / 0.45 / 0.42 / 0.35 / 0.54 / 0.56	0 (0) / 1 (0)
LightGBM instead of logistic regression	−1.39e-4 [−2.04e-4, −7.5e-5]	+2.5e-5 [−2.0e-6, +5.1e-5]	0.54 / 0.47 / 0.42 / 0.37 / 0.53 / 0.55	0 (0) / 7 (4)
Times ± 1 h	−1.67e-4 [−2.29e-4, −1.05e-4]	+2.6e-5 [+5.8e-6, +4.6e-5]	0.55 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	2 (2) / 3 (0)
Times ± 3 h	−1.59e-4 [−2.21e-4, −1.00e-4]	+2.6e-5 [+6.4e-6, +4.7e-5]	0.55 / 0.48 / 0.42 / 0.36 / 0.55 / 0.57	1 (1) / 4 (1)
Times shifted −3 h	−1.44e-4 [−2.02e-4, −8.8e-5]	+2.4e-5 [+4.9e-6, +4.3e-5]	0.54 / 0.47 / 0.40 / 0.36 / 0.54 / 0.55	0 (0) / 1 (0)
Times shifted +3 h	−1.84e-4 [−2.52e-4, −1.18e-4]	+3.1e-5 [+8.9e-6, +5.3e-5]	0.55 / 0.47 / 0.45 / 0.36 / 0.55 / 0.57	1 (1) / 4 (1)
Observations degraded, σ = 0.3	−1.60e-4 [−2.23e-4, −1.00e-4]	+2.4e-5 [+4.6e-6, +4.4e-5]	0.54 / 0.47 / 0.43 / 0.36 / 0.55 / 0.57	2 (2) / 3 (0)

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Validation	Log loss: all available − GFS only	Log loss: Late + GFS − all available	Peak REV: all available / GFS only / Early only / always wait / VOI policy / latency-free benchmark	VOI policy vs best fixed policy: α values significantly better (material) / significantly worse (material)
Observations degraded, σ = 0.6	−1.44e-4 [−2.02e-4, −8.6e-5]	+1.9e-5 [+2.0e-6, +3.7e-5]	0.54 / 0.47 / 0.41 / 0.36 / 0.54 / 0.57	2 (1) / 5 (1)
Real-time V07, Jun–Oct 2024 only	−3.01e-4 [−7.18e-4, +9.1e-5]	+4.1e-5 [−8.1e-5, +1.4e-4]	0.62 / 0.53 / 0.52 / 0.39 / 0.61 / 0.61	0 (0) / 10 (4)
With control cells, weighted	−1.40e-6 [−1.83e-6, −9.7e-7]	+1.7e-7 [+4.7e-8, +3.0e-7]	0.69 / 0.57 / 0.59 / 0.47 / 0.70 / 0.71	3 (2) / 7 (4)
With control cells, unweighted	−5.06e-5 [−6.74e-5, −3.4e-5]	+6.5e-6 [+1.7e-6, +1.2e-5]	0.68 / 0.56 / 0.57 / 0.47 / 0.68 / 0.69	3 (2) / 6 (1)

Table C2. Overall tests of selective waiting. Cost–loss rows: mean REV difference between the VOI policy and the best fixed policy over the α values of the exploratory range (14 values from 10^{-4} to 10^{-2}; with control cells, 0.2–20 times the weighted base rate), with its 95% interval from 2000 paired resamples of 7-day blocks, and in parentheses the number of these α values at which the simultaneous 95% band lies above / below zero. Rolling rows: mean difference in held-out detection between the threshold waiting band and the single threshold (Early + GFS) over the five warning budgets, with its 95% interval from 2000 rainfall-event resamples, and the number of budgets with the simultaneous band above / below zero.

Validation	ℓ = 0	ℓ = 6 h	ℓ = 12 h
Leave one year out	+0.002 [−0.005, +0.007] (0/0)	−0.001 [−0.007, +0.002] (0/0)	−0.003 [−0.010, 0.000] (0/0)
Leave one year out, fully nested bins	+0.003 [−0.004, +0.008] (1/0)	0.000 [−0.006, +0.003] (0/0)	−0.002 [−0.009, +0.001] (0/0)
Recalibrated issue-now rule among the fixed policies	−0.002 [−0.009, +0.002] (0/0)	−0.005 [−0.012, −0.001] (0/0)	−0.005 [−0.013, −0.002] (0/0)
Coherent sub-window probabilities	0.000 [−0.006, +0.004] (0/0)	0.000 [−0.006, +0.004] (0/0)	−0.003 [−0.010, 0.000] (0/0)
Coherent sub-window probabilities, fully nested bins	+0.002 [−0.005, +0.006] (1/0)	−0.001 [−0.007, +0.004] (0/0)	−0.002 [−0.009, +0.001] (0/0)
Leave one province out	+0.002 [−0.006, +0.006] (0/0)	−0.002 [−0.008, +0.002] (0/0)	−0.001 [−0.008, +0.002] (1/0)
T3/T4 labels only	+0.003 [−0.005, +0.007] (0/0)	+0.002 [−0.007, +0.005] (0/0)	0.000 [−0.008, +0.002] (0/0)
LightGBM instead of logistic regression	−0.002 [−0.012, +0.003] (0/0)	−0.003 [−0.013, +0.002] (0/0)	−0.001 [−0.013, +0.002] (0/0)
Times ± 1 h	+0.003 [−0.004, +0.008] (0/0)	0.000 [−0.006, +0.004] (0/0)	−0.003 [−0.010, +0.001] (0/0)
Times ± 3 h	+0.004 [−0.003, +0.008] (0/0)	0.000 [−0.006, +0.004] (0/0)	−0.004 [−0.010, −0.002] (1/0)

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Validation	ℓ = 0	ℓ = 6 h	ℓ = 12 h
Times shifted −3 h	+0.002 [−0.005, +0.007] (0/0)	+0.001 [−0.006, +0.005] (0/0)	−0.002 [−0.009, +0.001] (0/0)
Times shifted +3 h	+0.002 [−0.005, +0.006] (0/0)	+0.002 [−0.005, +0.006] (0/0)	−0.001 [−0.009, +0.002] (0/0)
Observations degraded, σ = 0.3	+0.002 [−0.004, +0.006] (1/0)	−0.001 [−0.007, +0.003] (1/0)	+0.001 [−0.006, +0.004] (0/0)
Observations degraded, σ = 0.6	+0.002 [−0.004, +0.005] (0/0)	−0.001 [−0.007, +0.002] (0/0)	−0.001 [−0.008, +0.002] (0/0)
Real-time V07, Jun–Oct 2024 only	−0.012 [−0.102, −0.001] (0/0)	−0.003 [−0.083, +0.002] (0/3)	−0.009 [−0.088, −0.006] (0/2)
With control cells, weighted	+0.003 [−0.004, +0.006] (0/0)	−0.003 [−0.010, 0.000] (0/0)	−0.002 [−0.009, +0.001] (0/0)
With control cells, unweighted	0.000 [−0.007, +0.004] (0/0)	+0.001 [−0.005, +0.005] (0/0)	0.000 [−0.007, +0.003] (1/0)
Rolling replay	+0.001 [−0.005, +0.007] (0/0)	−0.003 [−0.009, +0.003] (0/0)	+0.001 [−0.006, +0.008] (0/0)
Rolling replay, nested selection	−0.002 [−0.007, +0.003] (0/0)	−0.001 [−0.006, +0.003] (0/0)	+0.003 [−0.005, +0.011] (0/0)
Rolling replay, denser grid	+0.001 [−0.006, +0.008] (0/0)	−0.004 [−0.012, +0.004] (0/0)	+0.001 [−0.008, +0.010] (0/0)
Rolling replay, nested, denser grid	−0.004 [−0.011, +0.003] (0/0)	−0.009 [−0.017, −0.001] (0/0)	−0.001 [−0.010, +0.007] (0/0)
Rolling replay, control cells, weighted	−0.006 [−0.013, −0.001] (0/0)	−0.002 [−0.009, +0.004] (0/0)	−0.002 [−0.010, +0.006] (0/0)
Rolling replay, control cells, unweighted	−0.003 [−0.007, +0.001] (0/0)	−0.002 [−0.008, +0.003] (0/0)	−0.004 [−0.011, +0.004] (0/0)

Table C3. Calibration and coherence of the window probabilities for all available information (Early + GFS), and the oracle reference. O/E: ratio of observed to predicted frequency over all rows, with 95% intervals from 7-day blocks. Slopes: logistic recalibration slopes of p_A, p_B, p_{AB} and q (q against the label of B). Incoherent decisions: share of decisions with $p_{AB} > p_A + p_B$, and in brackets with $p_B > p_{AB}$, both excluded by a single probability measure. Local O/E: bins of p_{AB} (half-decade wide, at least 20 positive rows) whose interval excludes 1, or nearly does. Oracle reference: largest and smallest mean REV difference from issuing now over α. Last column: α values at which the VOI policy has a higher mean REV than the oracle reference, and the largest such difference.

Validation, lead	O/E of p_{AB}	Slopes	Mean q / mean p_B	Incoherent decisions	Local O/E of p_{AB}	Oracle reference − issue now: max / min	VOI policy above oracle reference
Leave one year out, ℓ = 0	1.00 [0.85, 1.15]	0.98–0.99	1.001	74% (1.7%)	10^{-3}–3.2×10^{-3}: 0.76 [0.59, 0.93]	+0.024 / −0.007	7 of 23 (+0.005)

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Validation, lead	O/E of p_{AB}	Slopes	Mean q / mean p_B	Incoherent decisions	Local O/E of p_{AB}	Oracle reference − issue now: max / min	VOI policy above oracle reference
Leave one year out, ℓ = 6 h	1.00 [0.85, 1.15]	0.98–0.99	0.998	53% (< 0.1%)	10^{-3}–3.2×10^{-3}: 0.83 [0.66, 1.00]	+0.022 / −0.006	7 of 23 (+0.008)
Leave one year out, ℓ = 12 h	1.00 [0.85, 1.15]	0.98	1.002	30% (< 0.1%)	—	+0.020 / −0.005	6 of 23 (+0.004)
Real-time V07, Jun–Oct 2024, ℓ = 0	1.11 [0.65, 1.49]	0.96–1.02	1.054	76% (1.6%)	—	+0.031 / −0.066	9 of 23 (+0.058)
With control cells, weighted, ℓ = 0	1.00 [0.85, 1.15]	1.03–1.10	1.001	17% (2.2%)	10^{-5}–3.2×10^{-5}: 0.73 [0.57, 0.91]	+0.023 / −0.012	9 of 43 (+0.044)
With control cells, unweighted, ℓ = 0	1.00 [0.85, 1.15]	0.99	1.001	77% (1.6%)	3.2×10^{-4}–10^{-3}: 0.72 [0.56, 0.88]	+0.023 / −0.318	12 of 43 (+0.027)
Coherent sub-window probabilities, ℓ = 0	1.00 [0.85, 1.14]	0.99	1.000	0% (0%)	10^{-3}–3.2×10^{-3}: 0.76 [0.59, 0.93]	+0.024 / −0.008	6 of 23 (+0.004)

The separate window models of the main analysis give $p_{AB} > p_A + p_B$ in 74% of decisions at ℓ = 0 (53% and 30% at ℓ = 6 and 12 h), typically by a few percent (Table C3). Deriving all window probabilities from one sub-window model removes this without changing calibration and reveals no stable advantage for waiting: the band is significantly better than the best fixed policy at no α in the exploratory range at any lead (upper limits at most 0.029), and only the fully nested version gains materially, once, at $\alpha = 1.6 \times 10^{-3}$ (+0.022 [+0.005, +0.032]), where the sub-window p_{AB} over-predicts as before. The oracle reference gains at most +0.024 [+0.013, +0.036] over issuing at once ($\alpha = 1.6 \times 10^{-3}$, ℓ = 0; +0.023 and +0.020 at ℓ = 6 and 12 h) and does not raise the peak REV. With incoherent probabilities, small inconsistencies decide many of its choices, so its predicted gain exceeds its realized gain at every α and the feasible band beats it at 6–7 of 23 α values per lead; with coherent probabilities the band never exceeds it by more than 0.008.

Table C4. Pre-specified test on the independent 2008–2014 sample, information set Early only. D1: mean over $\alpha = 10^{-4}$–8×10^{-4} (seven grid values) of the REV difference VOI policy − issue now (C1 at ℓ = 0); D2: the same for always waiting one step − issue now (C2 at ℓ = 0). 2008–2014: models and preposterior bins fitted on 2015–2024 and applied unchanged; 2015–2024: the same analysis with leave-one-year-out probabilities, for reference. 95% intervals from 2000 resamples of 7-day blocks. AUC of the Early-only p_{AB}.

Sample, lead	Records / events	AUC	D1: VOI policy − issue now	D2: always wait − issue now
2008–2014 (test), ℓ = 0	354 / 215	0.779	+0.029 [+0.009, +0.052]	−0.67 [−0.81, −0.53]
2008–2014 (test), ℓ = 6 h		0.765	+0.015 [−0.004, +0.036]	−0.63 [−0.77, −0.50]
2008–2014 (test), ℓ = 12 h		0.754	+0.002 [−0.008, +0.014]	−0.56 [−0.69, −0.44]
2015–2024 (reference), ℓ = 0	273 / 193	0.816	+0.023 [+0.009, +0.037]	−0.46 [−0.56, −0.37]
2015–2024 (reference), ℓ = 6 h		0.804	+0.009 [−0.003, +0.021]	−0.45 [−0.54, −0.35]

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Sample, lead	Records / events	AUC	D1: VOI policy − issue now	D2: always wait − issue now
2015–2024 (reference), ℓ = 12 h		0.793	+0.008 [−0.000, +0.018]	−0.40 [−0.50, −0.30]

Appendix D: Protocol for the independent 2008–2014 test

D1 Timing and scope

The independent test was written and frozen on 14 October 2026, after the 2015–2024 analyses but before the 2008–2014 decision dataset was assembled. The older records had been used for a case-crossover replication, but had not entered a decision analysis, probability-model fit or policy evaluation. The protocol tested Early-only information because the GFS 0.25° archive used here starts in 2015. It was designed as a positive control for detecting waiting gains when current information lacks a forecast, rather than as an independent test of the main Early + GFS comparison.

D2 Data and frozen evaluation

The eligible sample comprised 354 rainfall-triggered records in April–October 2008–2014, with coordinates, county-level or finer locations and T2–T4 occurrence times, grouped into 215 rainfall events at 321 cells. Occurrence times and clustering followed Sect. 2.1. Decisions were made every 6 h at these cells. Features were IMERG Early totals over 6, 24, 72 and 168 h ending at t − 4 h. Rows with incomplete features or without a corresponding decision at t + 6 h were excluded.

For each required lead ℓ = 0, 6 or 12 h, the target windows were AB = (t + ℓ, t + 24 h], A = (t + ℓ, t + ℓ + 6 h] and B = (t + ℓ + 6 h, t + 24 h]. At the next decision, q targets the same remaining window B, corresponding to offsets (ℓ, 18 h] relative to t + 6 h. The Early-only logistic models used the transformation, standardization and regularization in Appendix A (C = 1, tolerance 10^{-10}). One model per target window was fitted once to all 2015–2024 decisions and applied unchanged to 2008–2014; no parameter was tuned on the older sample.

The preposterior distribution used 40 quantile bins of logit p_B. Bin edges were determined from the 2015–2024 leave-one-year-out probabilities, and bin contents were the corresponding next-step probabilities q from those years. For each older-sample decision, $\mathbb{E}\min(\alpha, q)$ was estimated from the 2015–2024 decisions in its bin. No GFS forecast classes were used. The compared policies were issuing immediately if $p_{AB} > \alpha$, always waiting one step and then issuing if $q > \alpha$, and waiting selectively according to Eq. (1), otherwise using the immediate rule. Expenses followed Sect. 3.5, and REV in Eq. (2) used the positive-label base rate of the 2008–2014 sample. Confidence intervals used 2000 paired bootstrap resamples of 7-day blocks across all cells, with random seed 0.

D3 Pre-specified hypotheses and interpretation

The primary and secondary tests used ℓ = 0 and seven points of the main α grid: 1.00×10^{-4}, 1.41×10^{-4}, 2.00×10^{-4}, 2.82×10^{-4}, 3.98×10^{-4}, 5.63×10^{-4} and 7.95×10^{-4} (values rounded here for presentation). D_1 denotes the mean across these points

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

of REV for selective waiting minus REV for issuing immediately; primary hypothesis C1 required its lower 95% confidence limit to exceed zero. D_2 denotes the corresponding mean difference for always waiting minus issuing immediately; secondary hypothesis C2 required its upper 95% confidence limit to be below zero. Each bootstrap resample averaged the differences across the same seven points, preserving their dependence.

Support for C1 would show that the method can detect a waiting benefit in an independent Early-only sample. Failure to support C1 would weaken that positive-control argument. Neither outcome by itself determines the value of waiting when GFS is available. Per-α differences, leads of 6 and 12 h, AUC and log-loss skill, sample counts and the 2015–2024 reference analysis were declared descriptive outputs, not additional confirmatory tests. In that reference analysis, each year was excluded from its own preposterior groups.

D4 Execution record

Internal execution records indicate that the decision dataset was assembled and the test run once after the protocol had been frozen and hashed, without changes to the protocol. The frozen Chinese protocol, its English translation and checksum, the execution log, and the results are archived with the code in release v1.0.0 (see Code and data availability): protocol/S2_original_zh_frozen.md, protocol/S2_prereg_2008_2014.md, results/e5/confirm0814.log, and results/e5/confirm_2008_2014.json, respectively. The archive also includes SHA256SUMS. These internal records document the reported sequence; the later public deposit does not independently establish the timing. Both hypotheses and the descriptive outputs are reported in Sect. 4.7 and Table C4.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Appendix E: Overall rank correlations for H2

The original H2 plan stated $\rho > 0$ and $p < 0.05$, without selecting a primary α–lead combination or an across-grid confirmation rule. The implemented analysis reports pointwise correlations and intervals, rather than the nominal p-value test. Table E1 presents all archived results for the main analysis; selected positive associations do not provide a single confirmatory verdict for H2.

Ranks, including average ranks for ties, are calculated over all 2 243 653 decisions with complete features and an available next decision. Intervals use 2000 resamples of 7-day blocks across all cells, retaining the full-sample ranks. They are pointwise and not adjusted for multiple comparisons.

Table E1. Spearman correlation ρ between predicted VOI and the realized gain of waiting over all decisions in the main leave-one-year-out Early + GFS analysis, by cost–loss ratio α and lead threshold ℓ. Brackets give pointwise 95% block-bootstrap intervals. A dash denotes a non-finite point estimate; † marks a finite point estimate whose archived interval bounds are non-finite. Neither is interpreted as a zero effect. Values are rounded; the full-precision results are in results/e3/voi_policy.json in the code archive.

α	ρ [95% CI], ℓ = 0 h	ρ [95% CI], ℓ = 6 h	ρ [95% CI], ℓ = 12 h
1.00×10^{-4}	−0.036 [−0.044, −0.027]	−0.129 [−0.138, −0.119]	−0.291 [−0.301, −0.280]
1.41×10^{-4}	−0.071 [−0.078, −0.063]	−0.163 [−0.170, −0.155]	−0.333 [−0.341, −0.324]
2.00×10^{-4}	−0.093 [−0.099, −0.086]	−0.189 [−0.195, −0.182]	−0.358 [−0.364, −0.351]
2.82×10^{-4}	−0.106 [−0.111, −0.100]	−0.206 [−0.211, −0.200]	−0.366 [−0.371, −0.361]
3.98×10^{-4}	−0.118 [−0.122, −0.113]	−0.213 [−0.218, −0.209]	−0.354 [−0.358, −0.349]
5.63×10^{-4}	−0.125 [−0.129, −0.120]	−0.213 [−0.216, −0.209]	−0.323 [−0.329, −0.318]
7.95×10^{-4}	−0.124 [−0.128, −0.121]	−0.200 [−0.203, −0.196]	−0.275 [−0.281, −0.268]
1.12×10^{-3}	−0.122 [−0.125, −0.118]	−0.178 [−0.182, −0.173]	−0.237 [−0.244, −0.229]
1.59×10^{-3}	−0.113 [−0.116, −0.108]	−0.147 [−0.152, −0.142]	−0.186 [−0.193, −0.177]
2.24×10^{-3}	−0.088 [−0.092, −0.083]	−0.131 [−0.136, −0.125]	−0.167 [−0.175, −0.158]
3.17×10^{-3}	−0.112 [−0.117, −0.107]	−0.120 [−0.126, −0.113]	−0.078 [−0.085, −0.070]
4.47×10^{-3}	−0.049 [−0.054, −0.045]	−0.010 [−0.015, −0.006]	−0.013 [−0.018, −0.009]
6.32×10^{-3}	+0.021 [+0.018, +0.025]	+0.035 [+0.031, +0.039]	+0.012 [+0.009, +0.016]
8.93×10^{-3}	+0.037 [+0.034, +0.041]	+0.040 [+0.035, +0.045]	+0.017 [+0.014, +0.020]
1.26×10^{-2}	+0.034 [+0.030, +0.038]	+0.032 [+0.027, +0.037]	+0.016 [+0.013, +0.020]
1.78×10^{-2}	+0.025 [+0.021, +0.028]	+0.023 [+0.019, +0.027]	+0.008 [+0.004, +0.011]
2.52×10^{-2}	+0.018 [+0.014, +0.021]	+0.012 [+0.009, +0.014]	+0.006 [+0.003, +0.009]
3.55×10^{-2}	+0.008 [+0.006, +0.010]	+0.007 [+0.005, +0.009]	+0.004 [+0.001, +0.006]
5.02×10^{-2}	+0.005 [+0.003, +0.006]	+0.004 [+0.002, +0.006]	+0.003†
7.09×10^{-2}	+0.002 [+0.001, +0.004]	+0.002†	—
0.1	+0.002†	—	—
0.142	—	—	—
0.2	—	—	—

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Code and data availability

The study code, analysis plans, frozen protocol, execution logs and result files are available at https://gitee.com/hhgxylj/landslide-warning (release v1.0.0; commit 85b74de4c24ee4279b202662fb98191d70bdba6d), in zenodo/landslide-warning-archive.zip. Paths cited in the appendices are relative to the extracted archive directory landslide-warning-archive/. The end-to-end comparison record is results/e5/recheck_compare.json.

The landslide catalogue is available from Zhao et al. (2026). IMERG V07 is available from the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC; Huffman et al., 2023b–g), GFS from the NCAR GDEX archive (NCEP, 2015) and the NOAA archive on AWS (NOAA, 2026), the LHASA version 2.0.0 retrospective archive from GES DISC (Stanley, 2023), and the Copernicus DEM GLO-90 from the European Space Agency (ESA, 2022). The terrain frame of the control cells was produced using Copernicus WorldDEM-90 © DLR e.V. 2010–2014 and © Airbus Defence and Space GmbH 2014–2018 provided under COPERNICUS by the European Union and ESA; all rights reserved.

Author contributions

Conceptualization, Hongyun Zhang; methodology, Jie Liu; resources, Jin Liu; writing—original draft preparation, Hongyun Zhang; writing—review and editing, Hongyun Zhang; validation, Jianchuang Ding; visualization, Yuanhao Zhu.

Competing interests

The authors declare no conflict of interest.

Acknowledgements

The authors used OpenAI Codex to assist with code development, benchmark evaluation, literature review, manuscript organization, and language editing. The authors reviewed and edited the generated material and take full responsibility for the content of this article.

Financial support

This research was funded by the Fundamental Research Funds for the Central Universities, grant number 3142024048, and the National Key Research and Development Program of China, grant number 2024YFF1306200.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Brunetti, M. T., Peruccacci, S., Rossi, M., Luciani, S., Valigi, D., and Guzzetti, F.: Rainfall thresholds for the possible occurrence of landslides in Italy, Nat. Hazards Earth Syst. Sci., 10, 447–458, https://doi.org/10.5194/nhess-10-447-2010, 2010.

Brunetti, M. T., Melillo, M., Peruccacci, S., Ciabatta, L., and Brocca, L.: How far are we from the use of satellite rainfall products in landslide forecasting?, Remote Sens. Environ., 210, 65–75, https://doi.org/10.1016/j.rse.2018.03.016, 2018.

Calvello, M. and Piciullo, L.: Assessing the performance of regional landslide early warning models: the EDuMaP method, Nat. Hazards Earth Syst. Sci., 16, 103–122, https://doi.org/10.5194/nhess-16-103-2016, 2016.

Carracedo-Martínez, E., Taracido, M., Tobias, A., Saez, M., and Figueiras, A.: Case-crossover analysis of air pollution health effects: a systematic review of methodology and application, Environ. Health Perspect., 118, 1173–1182, https://doi.org/10.1289/ehp.0901485, 2010.

Coughlan de Perez, E., van den Hurk, B., van Aalst, M. K., Jongman, B., Klose, T., and Suarez, P.: Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts, Nat. Hazards Earth Syst. Sci., 15, 895–904, https://doi.org/10.5194/nhess-15-895-2015, 2015.

Economou, T., Stephenson, D. B., Rougier, J. C., Neal, R. A., and Mylne, K. R.: On the use of Bayesian decision theory for issuing natural hazard warnings, Proc. R. Soc. A, 472, 20160295, https://doi.org/10.1098/rspa.2016.0295, 2016.

ESA: Copernicus DEM – Global and European Digital Elevation Model, European Space Agency [data set], https://doi.org/10.5270/ESA-c5d3d65, 2022.

Froude, M. J. and Petley, D. N.: Global fatal landslide occurrence from 2004 to 2016, Nat. Hazards Earth Syst. Sci., 18, 2161–2181, https://doi.org/10.5194/nhess-18-2161-2018, 2018.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Guzzetti, F., Peruccacci, S., Rossi, M., and Stark, C. P.: The rainfall intensity–duration control of shallow landslides and debris flows: an update, Landslides, 5, 3–17, https://doi.org/10.1007/s10346-007-0112-1, 2008.

Guzzetti, F., Gariano, S. L., Peruccacci, S., Brunetti, M. T., Marchesini, I., Rossi, M., and Melillo, M.: Geographical landslide early warning systems, Earth-Sci. Rev., 200, 102973, https://doi.org/10.1016/j.earscirev.2019.102973, 2020.

He, S., Wang, J., and Liu, S.: Rainfall event–duration thresholds for landslide occurrences in China, Water, 12, 494, https://doi.org/10.3390/w12020494, 2020.

Huffman, G. J., Bolvin, D. T., Joyce, R., Kelley, O. A., Nelkin, E. J., Tan, J., Watters, D. C., and West, B. J.: Integrated Multi-satellitE Retrievals for GPM (IMERG) Technical Documentation, NASA Goddard Space Flight Center, Greenbelt, MD, https://gpm.nasa.gov/sites/default/files/2023-07/IMERG_TechnicalDocumentation_final_230713.pdf (last access: 14 October 2026), 2023a.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Early Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V07, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERG/3B-HH-E/07, 2023b.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Late Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V07, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERG/3B-HH-L/07, 2023c.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V07, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERG/3B-HH/07, 2023d.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Early Precipitation L3 1 day 0.1 degree x 0.1 degree V07, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERGDE/DAY/07, 2023e.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Late Precipitation L3 1 day 0.1 degree x 0.1 degree V07, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERGDL/DAY/07, 2023f.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V07, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/GPM/IMERGDF/DAY/07, 2023g.

Huffman, G. J., Bolvin, D. T., Joyce, R., Kelley, O. A., Nelkin, E. J., Portier, A., Stocker, E. F., Tan, J., Watters, D. C., and West, B. J.: IMERG V07 Release Notes, NASA Goddard Space Flight Center, Greenbelt, MD, https://gpm.nasa.gov/sites/default/files/2024-12/IMERG_V07_ReleaseNotes_241126.pdf (last access: 14 October 2026), 2024.

Janes, H., Sheppard, L., and Lumley, T.: Case–crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias, Epidemiology, 16, 717–726, https://doi.org/10.1097/01.ede.0000181315.18836.9d, 2005.

Jewson, S., Scher, S., and Messori, G.: Decide now or wait for the next forecast? Testing a decision framework using real forecasts and observations, Mon. Weather Rev., 149, 1637–1650, https://doi.org/10.1175/MWR-D-20-0392.1, 2021.

Katz, R. W. and Murphy, A. H. (Eds.): Economic Value of Weather and Climate Forecasts, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511608278, 1997.

Kirschbaum, D. and Stanley, T.: Satellite-based assessment of rainfall-triggered landslide hazard for situational awareness, Earth's Future, 6, 505–523, https://doi.org/10.1002/2017EF000715, 2018.

Kirschbaum, D., Stanley, T., and Zhou, Y.: Spatial and temporal analysis of a global landslide catalog, Geomorphology, 249, 4–15, https://doi.org/10.1016/j.geomorph.2015.03.016, 2015.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Krøgli, I. K., Devoli, G., Colleuille, H., Boje, S., Sund, M., and Engen, I. K.: The Norwegian forecasting and warning service for rainfall- and snowmelt-induced landslides, Nat. Hazards Earth Syst. Sci., 18, 1427–1450, https://doi.org/10.5194/nhess-18-1427-2018, 2018.

Maclure, M.: The case-crossover design: a method for studying transient effects on the risk of acute events, Am. J. Epidemiol., 133, 144–153, https://doi.org/10.1093/oxfordjournals.aje.a115853, 1991.

Marchal, H., Bouttier, F., and Nuissier, O.: Is considering (in)consistency between runs so useless for weather forecasting?, Nat. Hazards Earth Syst. Sci., 25, 2613–2628, https://doi.org/10.5194/nhess-25-2613-2025, 2025.

Melillo, M., Brunetti, M. T., Peruccacci, S., Gariano, S. L., and Guzzetti, F.: An algorithm for the objective reconstruction of rainfall events responsible for landslides, Landslides, 12, 311–320, https://doi.org/10.1007/s10346-014-0471-3, 2015.

Murphy, A. H.: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation, Mon. Weather Rev., 105, 803–816, https://doi.org/10.1175/1520-0493(1977)105<0803:TVOCCA>2.0.CO;2, 1977.

NASA GPM: IMERG V07 Final Processing has Begun, NASA Global Precipitation Measurement Mission, https://gpm.nasa.gov/data/news/imerg-v07-final-processing-has-begun (last access: 14 October 2026), 2023.

NASA GPM: Status of IMERG Version 06 and 07 Data, NASA Global Precipitation Measurement Mission, https://gpm.nasa.gov/mission/news/status-imerg-version-06-and-07-data (last access: 14 October 2026), 2024.

NCEP: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, NSF National Center for Atmospheric Research [data set], https://doi.org/10.5065/D65D8PWK, 2015.

NOAA: NOAA Global Forecast System (GFS), Registry of Open Data on AWS [data set], https://registry.opendata.aws/noaa-gfs-bdp-pds/ (last access: 10 October 2026), 2026.

Piciullo, L., Calvello, M., and Cepeda, J. M.: Territorial early warning systems for rainfall-induced landslides, Earth-Sci. Rev., 179, 228–247, https://doi.org/10.1016/j.earscirev.2018.02.013, 2018.

Raiffa, H. and Schlaifer, R.: Applied Statistical Decision Theory, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Q. J. R. Meteorol. Soc., 126, 649–667, https://doi.org/10.1002/qj.49712656313, 2000.

Roulston, M. S. and Smith, L. A.: Evaluating probabilistic forecasts using information theory, Mon. Weather Rev., 130, 1653–1660, https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2, 2002.

Segoni, S., Piciullo, L., and Gariano, S. L.: A review of the recent literature on rainfall thresholds for landslide occurrence, Landslides, 15, 1483–1501, https://doi.org/10.1007/s10346-018-0966-4, 2018.

Stanley, T.: Global Landslide Nowcasts from LHASA L4 1 day 1 km x 1 km V2.0.0, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], https://doi.org/10.5067/8VKQDQFFOTS3, 2023.

Stanley, T. A., Kirschbaum, D. B., Benz, G., Emberson, R. A., Amatya, P. M., Medwedeff, W., and Clark, M. K.: Data-driven landslide nowcasting at the global scale, Front. Earth Sci., 9, 640043, https://doi.org/10.3389/feart.2021.640043, 2021.

https://doi.org/10.5194/egusphere-2026-6400
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Tang, G., Ma, Y., Long, D., Zhong, L., and Hong, Y.: Evaluation of GPM Day-1 IMERG and TMPA Version-7 legacy products over Mainland China at multiple spatiotemporal scales, J. Hydrol., 533, 152–167, https://doi.org/10.1016/j.jhydrol.2015.12.008, 2016.

Wang, C., Tang, G., Han, Z., Guo, X., and Hong, Y.: Global intercomparison and regional evaluation of GPM IMERG Version-03, Version-04 and its latest Version-05 precipitation products: similarity, difference and improvements, J. Hydrol., 564, 342–356, https://doi.org/10.1016/j.jhydrol.2018.06.064, 2018.

WMO: WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services, WMO-No. 1150, World Meteorological Organization, Geneva, 2015.

Xiao, R., Liu, Y., Chen, C., Su, Y., Wang, H., Xu, W., Fang, Z., and Liang, H.: 20-year early warning for regional geo-hazards risk in China: 2003–2022, The Chinese Journal of Geological Hazard and Control, 35, 1–9, https://doi.org/10.16031/j.cnki.issn.1003-8035.202401028, 2024 (in Chinese with English abstract).

Yunus, A. P., Faizah, E., Mukundan, S., Sajinkumar, K. S., Subramanian, S. S., Attada, R., and Achu, A. L.: ALERT: A scalable cloud-native framework for satellite rainfall-driven landslide early warning in data-scarce regions, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-4588, 2026.

Zhao, B., Zhang, L., Liu, Z., Ma, W., Wang, J., Sun, Q., Luo, W., Yu, Z., and Yuan, L.: A high-precision catalogue of landslide events in China based on news text mining with large language model, Sci. Data, 13, 722, https://doi.org/10.1038/s41597-026-07066-w, 2026.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The economic value of ensemble-based weather forecasts, B. Am. Meteorol. Soc., 83, 73–83, https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2, 2002.