sustainability

MDPI

Article

An Audit Framework for Physics-Informed Multi-Hazard Seismic Models: Gradient-Verified Coupling, Protocol-Dependent Constants, and Baseline-Anchored Operating Envelopes for Municipal Disaster Risk Reduction

Boris Kriuk [1,*] and Fedor Kriuk [2]

[1] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR, China
[2] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
* Correspondence: bkriuk@connect.ust.hk

Abstract

Sendai Framework Target G and SDG Indicator 11.b.2 require multi-hazard risk information that local governments can write into procedures that they are able to defend, which in turn requires models whose claims can be audited. We propose an audit framework built on three questions—does the physics term reach the network, are the learned constants properties of seismology or of the protocol, and over which decisions is the score usable against which reference—and apply it to POSEIDON, a physics-informed three-head model, on 27,100 $M \geq 5$ ComCat events (2010–2024) with a chronological, embargoed split and tuned tabular baselines on the same partition. Automatic differentiation shows that the coupled physics term reaches the encoder (gradient norm 50.94) while the parallel likelihood terms give exactly zero, yet the decoupled ablation reproduces the learned constants to within 0.02. Constants move by 131–317 seed standard deviations between two protocols while remaining "admissible", which, for interval-bounded constants, is guaranteed in advance. The aftershock head has positive cost–loss value for $r \in [0.325, 0.850]$, peaking at the Peirce skill score 0.441; the tsunami head has none at any r, so no life-safety decision lies inside the envelope. Every quantitative claim is stamped by the inference that supports it. The framework gives Target G's "usable and relevant" an operational meaning that a purchasing authority can verify on ordinary hardware.

Keywords: disaster risk reduction; physics-informed machine learning; model auditability; forecast value; cost–loss ratio; seismic hazard

Check for updates

Academic Editors: Lin Lu, Weiqi Zhou and Mark A. Bonn

Received: 5 September 2026
Revised: 24 September 2026
Accepted: 18 October 2026
Published: 23 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Hazard Information as Public Infrastructure

A disproportionate share of recent urban growth has occurred in seismically active and low-elevation coastal zones [1–3], so seismic and tsunami hazard is a first-order constraint on the sustainability of urban development [4–7]. Sendai Target G commits signatories to "substantially increase the availability of and access to multi-hazard early warning systems and disaster risk information and assessments" [8–12]; SDG Indicator 11.b.2 counts the *proportion of local governments* that adopt and implement local disaster risk reduction strategies [13,14].

The unit of implementation is therefore the local authority. A municipal emergency manager consumes a hazard model as an input to a standing operating procedure, under a budget, with a duty to justify decisions afterwards. Three requirements follow: the information must be *multi-hazard* [9,15,16], *probabilistic* [17,18], and *auditable*. A large literature addresses the first two [19–22]. This paper is about the third. Our concern is not that published models are poorly built; it is that the field's reporting conventions systematically fail to answer the questions that a deploying authority must ask. We develop an audit framework that answers them, apply it to our own model, and report what it finds—including where our own evidence is weakest.

1.2. Auditability as a Sustainability Requirement

The sustainability case for auditability is that the *decision infrastructure* built on hazard information must itself be durable. The Sendai Framework and the 2030 Agenda are formally joined: SDG indicators 1.5.1, 11.5.1 and 13.1.1 are harmonised with Sendai indicators A-1, B-1 and E-1, so a model whose claims cannot be audited weakens all three reporting lines at once [23]. The 2023 Midterm Review found Target G among the targets furthest from trajectory and identified a "last-mile" gap in which risk information is produced but not used locally, the binding constraint being trust and interpretability rather than availability [24]; the Early Warnings for All initiative places *disaster risk knowledge* as its first pillar [25], and indicator G-5 asks that risk information be "accessible, understandable, usable and relevant" [9]. An unauditable model cannot be maintained across staff turnover, cannot be defended in an after-action inquiry, and must be re-procured rather than re-audited whenever the vendor, the catalogue or the task set changes. The framework is therefore offered as a contribution to the durability of the local strategies counted under Indicator 11.b.2, not only to model quality (Section 5).

1.3. Three Questions That an Audit Must Answer

(a) Does the physics term reach the network?

Statistical seismological laws are naturally written as functions of batch statistics: the Gutenberg–Richter likelihood depends on magnitude counts [26,27], Båth's law on the observed mainshock–largest-aftershock difference [28,29]. Both can be computed from labels and trainable constants alone. A physics loss written that way has *identically zero* gradient with respect to the network weights: the model is a conventional classifier trained alongside an in-loop maximum-likelihood fit that shares an optimiser. Its outputs, constants and loss curves are indistinguishable from a genuinely coupled model. Distinguishing them requires differentiating the physics term with respect to the network parameters. We make this a gate that must pass before training (Section 3.6) and report its outcome, with its limits, as a primary result (Section 4.1).

(b) Are the recovered constants properties of seismology, or of the protocol?

The standard interpretability evidence is that a fitted b-value or Omori exponent falls inside a published interval [27,30]. Two protocols applied to the same model family return $b = 1.069$ and $b = 0.752$, and $c = 0.0353$ d and $c = 0.1948$ d, and both pass that test. Worse, when the constants are sigmoid-bounded into their published intervals—as ours are (Section 3.5)—they cannot fail it. The honest uncertainty on a learned constant is the spread across protocols, not across seeds, and the honest evidence that a bounded constant was identified is its distance from a neutral initialisation (Section 4.2).

(c) Over which decisions is the score usable, and against what reference?

The cost–loss framework [31–33] says to act when $p > r$ for cost/loss ratio $r = C/L$. Reporting value at one ratio, or a threshold-free summary [34,35], leaves an authority unable

to tell whether *its* decision is one that the model serves. We report value as a function of r [32,36] and derive both envelope boundaries analytically (Appendix C). Murphy value is referenced to climatology, which is a weak opponent, so we also anchor to tuned tabular baselines on the same partition (Section 4.5). Class-imbalance weighting [37–40] inflates predicted probabilities; where the inflated minimum output exceeds a decision band, the rule fires on every event, value is zero by construction, and precision equals the base rate—numbers easily misread as discrimination [41]. Both executed decision bands sit in that regime (Section 4.6.1).

1.4. Positioning and Scope of the Framework

The framework is a general audit standard, not a new model or a new evaluation metric, and it is worth being precise about how it differs from the three literatures it borrows from. *Physics-informed learning* has studied gradient pathologies in PDE-constrained networks, where the residual gradient exists but is badly scaled relative to the data term [42,43]; the failure mode that we test for is prior to that—a physics penalty on batch statistics whose network gradient is not badly scaled but *identically zero*—and to our knowledge, no published seismological physics-informed model reports the per-term test that detects it. *Model documentation* (model cards, datasheets, factsheets) [44–46] specifies what to disclose; it does not specify a verification that the disclosed mechanism is the executed one, which is what the gradient gate and the manuscript–code reconciliation record (Appendix E) add. *Identifiability analysis* classically asks whether data determine a parameter; our decomposition asks the narrower operational question of whether a reported "admissible" constant carries any information once the reparameterisation has bounded it, and answers with quantities (distance from a neutral initialisation, same-sample maximum-likelihood value, identifying route) that any study can emit at no cost.

Its scope follows from what it tests. Question (a) applies to any model whose loss contains a physically named term with learned constants; (b) to any study that offers those constants as interpretability evidence; (c) to any model whose probabilistic output is thresholded for a decision. The three questions map onto exactly two features of a model—learned constants in a physics loss, and a thresholded probabilistic output—so any model with both is in scope, whichever hazard it addresses. Nothing in the gates depends on the network architecture, the seismological laws chosen, or the hazard; the seismic case study is the demonstration, not the boundary.

1.5. Contributions

1. **A gradient-verified coupling criterion with a reference implementation.** The physics objective is split into *coupled* terms that constrain model outputs and *params-only* likelihood terms that fit the constants; the gate differentiates each against the network and reports the norm per term (Table 5). The `no_physics` ablation is the decoupled architecture, measured (Section 4.4).
2. **An identifiability decomposition for bounded constants.** Each constant is reported with its bound, its neutral initialisation, the distance travelled, the standalone maximum-likelihood value on the same sample, and the route (coupled, params-only, or both) by which it is identified (Table 6, Figure 2).
3. **Protocol dependence as a general finding.** Two protocols differ by 131–317 seed standard deviations in b and p and a factor of 5.5 in c; both pass the published-range test (Section 4.3).

4. **A baseline-anchored operating envelope.** $V(r)$ for both heads, with analytic boundaries; and tuned tabular baselines on the same partition (Sections 4.6 and 4.5).
5. **Claim-class certification with an *absent* class, applied to our own draft.** Appendix E lists the discrepancies between drafts of this manuscript and the executed configuration, several of them load-bearing.

1.6. Relation to the Feasibility Study

This work builds on our earlier report of POSEIDON [47] ("the feasibility study"), which asked whether one architecture can carry multi-hazard heads, an energy-based latent and jointly fitted constants at all. Values quoted from it here are quoted, not recomputed; where a claim depends on them, we say so. The protocols differ in catalogue, magnitude range, declustering convention, loss weighting and sample (Table 8); having two protocols that we control is what lets us measure protocol sensitivity directly, and the direction and size of the shift that we observe is corroborated by independent published work on the effect of declustering on b [27,48]. Its baseline comparison does not transfer (Appendix F); we run our own.

1.7. Scope and Non-Claims

We make no claim of deterministic prediction [17]. All tasks are supervised classification with catalogue-derived labels and features computed strictly from preceding events [18]. Policy sections are normative and stamped accordingly. Table 1 consolidates the non-claims.

Table 1. Consolidated non-claims.

We Do *Not* Claim	**Discussed**
That POSEIDON outperforms, or is outperformed by, the tabular baselines beyond what Table 11 shows at the declared budget	Section 4.5
That any at-budget contrast with the baselines would persist at convergence	Section 4.5
That the reported constants are converged estimates	Sections 3.6 and 4.2
That admissibility of b, p, α or ΔM is evidence of anything (they are interval-bounded)	Section 3.5
That the seed standard deviations are uncertainties on the constants	Section 4.3
That the tsunami head forecasts tsunami occurrence, or that the aftershock head forecasts future aftershocks	Section 3.2
That precision and recall in the executed decision bands measure discrimination	Section 4.6.1
That the 1.8% residual to V_{pot} is a performance figure	Section 4.6.2
That the gradient norm of 50.94 describes the physics term's influence *during* training	Section 4.1
That the candidate cost ratios have been validated with stakeholders	Section 4.7
That physics-informing changed the compute required for a given performance	Section 3.7

2. Related Work

2.1. Machine Learning for Seismic Hazard

Deep learning is operationally competitive for detection and phase picking [49–52]. For aftershock forecasting, DeVries et al. [53] was shown to be matched by a single neuron on the same features [54]. That episode is the precedent for this paper: the original

was not in error, it was *under-anchored*, and the correction came from running the cheap alternative [55,56]. We run it on our own model (Section 4.5).

2.2. Forecast Evaluation Standards

Operational earthquake forecasting has an established verification tradition [17,18, 57–61] in which candidate forecasts are tested prospectively against a reference model with pre-registered statistics. The measures that we report—Brier skill score, reliability (ECE), the Peirce skill score and Murphy's relative economic value—are the standard forecast-verification toolkit [33,62,63], and the logistic-regression baseline of Table 11 is the conventional statistical comparator. Where this paper departs from the CSEP tradition is in asking, in addition, whether a forecast serves a *named decision*: an operating envelope is a statement about the decisions that a forecast can serve at a stated cost ratio, which is the question that a deploying authority asks of a candidate and which a likelihood test does not answer.

2.3. Physics-Informed Learning

PINNs embed governing equations as differentiable penalties [64–66]. In seismology, the relations are statistical [30,67–73] and usually imposed as a penalty on aggregate statistics with jointly learned constants. Two properties of that practice motivate this paper: a penalty on batch statistics does not constrain the network at all [42,43], and constants mapped through a bounded reparameterisation cannot fail the admissibility test that they are then reported as passing. That estimated seismological constants depend on catalogue window and declustering is itself well documented—Mizrahi et al. [48] show that declustering alone systematically shifts the mainshock b-value, and Marzocchi and Sandri [27] review the sensitivity of b to completeness and binning—so the protocol dependence that we measure in a learned model has an independent literature behind it.

2.4. Forecast Value and Skill

Murphy's cost–loss model [31], extended by Richardson [32], gives value relative to climatology as a function of r [33,36,62,63,74]. At $r = s$, the value equals the Peirce skill score [75] and its threshold optimum the Youden index [76]; we use these identities as constraints on interpretation. Decision-curve analysis [77] is the clinical analogue.

2.5. Auditability, Procurement and DRR Governance

Model cards, datasheets and internal-audit frameworks argue that metrics alone are insufficient for deployment [44–46,78–83]. The local-strategy literature asks for risk information that is understandable, locally owned and embedded in the authority's own planning cycle [84–87], and community-resilience frameworks add that the capacity to act on it is place-based and unevenly distributed [88]. The Midterm Review [24] and the Early Warnings for All plan [25] both identify the translation of risk knowledge into local action as the weakest link. Our contribution to that literature is narrow: it states, for one class of model, what a local authority would have to be shown before a probabilistic output can be written into a procedure, and shows that current reporting conventions do not supply it. Compute reporting is an established expectation [89–93]; we did not instrument it (Section 3.7).

3. Method

3.1. Data, Splits and Features

Table 2 gives provenance. The catalogue is USGS ComCat [94], 1 January 2010 to 1 January 2025, $M \geq 4.5$, 112,562 events, declustered with Gardner–Knopoff space–time

windows [95,96]. Triggers are the 27,100 events with $M \geq 5.0$, a threshold above the global completeness magnitude of ComCat for this period; per-year M_c and b-value series are emitted as diagnostics [97,98] and are released with the artefacts, but training uses the fixed threshold.

3.1.1. Split

Triggers are ordered by origin time. The split instant is defined as the origin time of the 21,681st trigger in that order: the earliest 80% (21,680 events) train and the latest 20% (5420 events) validate. At the catalogue's mean rate of roughly 1800 $M \geq 5$ triggers per year, the validation partition therefore spans approximately the final three years of the window, and its base rates are read directly from the validation predictions (Table 3). Training rows whose origin time falls within 90 days before the split instant are removed (embargo), so no training feature window overlaps the validation period; the embargo removes only rows whose 30- and 90-day lookbacks would otherwise straddle the boundary, and the exact post-embargo count and split timestamp are written to the released partition manifest. The random split of Section 4.4 uses the same 80/20 proportion with a seeded permutation and no embargo; it is reported only to quantify the optimism it introduces. All model and baseline validation numbers in this paper are computed on the same 5420 events.

Table 2. Dataset provenance and label definitions, as emitted by the pipeline.

Property	Value
Catalogue	USGS ANSS Comprehensive Catalog (ComCat), FDSN event service
Window	1 January 2010 T00:00:00 to 1 January 2025 T00:00:00, $M \geq 4.5$
Retrieved (UTC)	1 September 2026 T18:31:49
Events	112,562; SHA-256 `764ed0a6eae5b0e6af2e971edb3d2306`...
Declustering	Gardner & Knopoff [95] windows: $L(M) = 10^{0.1238M+0.983}$ km; $T(M) = 10^{0.032M+2.7389}$ d for $M \geq 6.5$, else $10^{0.5409M-0.547}$ d
Foreshock window	30 d before a larger event, within $1.0 \times L(M)$
Trigger threshold	$M \geq 5.0$: 27,100 samples
Base rates (all triggers)	aftershock 0.494, foreshock 0.225, tsunami flag 0.0352
Physics mask coverage	largest aftershock recorded for 0.346 of triggers; parent delay for 0.476
Tsunami label	ComCat operational `tsunami` flag

Table 3. Chronological partition. Validation-period base rates are read from the validation predictions; training-period rates are implied by arithmetic from the whole-sample rates of Table 2 and are stated before embargo.

Item	Training Partition	Validation Partition
Events	21,680 before the 90-day embargo	5420
Period	1 January 2010 to the split instant (origin time of trigger 21,681)	split instant to 31 December 2024
Aftershock rate	≈0.498 (implied)	0.4790
Tsunami-flag rate	≈0.034 (implied)	0.0408
Foreshock rate	0.225 (whole sample)	—

The tsunami-flag rate is higher in the validation period than in training (0.0408 vs. ≈0.034); the flag is an operational product whose issuance practice is not stationary (Section 3.2).

3.1.2. Stationarity Across the Split

The two periods are similar in the label that the primary head targets and differ in the label that is an operational product. The aftershock rate is ≈0.498 in training and 0.4790 in

validation, a difference of two percentage points on a rule that depends only on catalogue geometry. The tsunami-flag rate rises from ≈0.034 to 0.0408, a relative change of about 20%; the flag is issued by an operational procedure whose practice is not stationary (Section 3.2), and that non-stationarity is one reason the tsunami head is the weaker of the two under a chronological split. Because the trigger threshold ($M \geq 5.0$) sits above the completeness magnitude throughout the window, changes in network sensitivity over 2010–2024 affect the $M \geq 4.5$ catalogue used for feature construction but not the trigger population itself.

3.1.3. Features

Each trigger carries a 16-dimensional vector $\mathbf{x} \in \mathbb{R}^{16}$: magnitude, depth, latitude, longitude, day-of-year sine and cosine, hour, three depth-class indicators, and six local-context statistics from events within $\pm 2^\circ$ in the preceding 30 days (log counts at 7 and 30 d, local maximum magnitude, log energy, magnitude deficit, recent-activity trend). Three 6-channel grids $\mathbf{G}^{(7)}, \mathbf{G}^{(30)}, \mathbf{G}^{(90)} \in \mathbb{R}^{6\times 90\times 180}$ summarise global seismicity at 2° resolution over 7-, 30- and 90-day lookbacks (log count, maximum magnitude, log energy, mean depth, magnitude variance, activity trend). Causality is enforced by construction: every window uses a strict inequality on time, so no event at or after the trigger enters any channel [99,100]. The tabular baselines of Section 4.5 see $\mathbf{x}$ only; the network sees $\mathbf{x}$ and the grids.

3.2. Label Semantics

All three targets are catalogue-derived rules. We state each as a rule and say what discrimination against it measures (Table 4).

Aftershock (p_a): the trigger is itself classified as a dependent event by the Gardner–Knopoff window of a preceding larger event. Validation base rate $s_a = 0.4790$. Because the parent precedes the trigger, this label is in principle a function of the past catalogue; however, the GK time window reaches 268 d for an $M5.5$ parent and 885 d for $M6.5$, far beyond the 30/90-day feature windows, so it is only partly recoverable from the inputs. Operationally, the head answers "is this event part of an ongoing sequence", which is what governs cordon and inspection decisions. It does *not* forecast whether the trigger will produce aftershocks; that quantity is carried by the two auxiliary regression outputs of Section 3.4. An earlier draft described this label as "≥1 event with $M' \geq M - 1.0$ within 100 km and 30 d after the trigger"; that is not the executed rule (Appendix E, item 5).

Foreshock (p_f): a larger event follows within 30 d and within $L(M)$ of the larger event. Whole-sample base rate 0.225. This is a genuine forward-looking target.

Tsunami (p_t): the ComCat `tsunami` property, an operational screening flag whose documentation states that it "does not indicate if a tsunami actually did or will exist" [94]. Validation rate $s_t = 0.0408$. The flag is largely a deterministic function of magnitude, depth and offshore location, all of which are inputs; discrimination against it measures recoverability of a screening rule.

Table 4. Tasks, targets, base rates and loss weighting, as executed. Three heads, matching the emitted performance tables.

Head	Target Definition	Base Rate	Loss
p_a aftershock	trigger lies in the GK window of a preceding larger event	0.4790 (valid.)	BCE, label smoothing 0.05
p_f foreshock	a larger event follows within 30 d and $L(M)$	0.225 (all)	BCE, positive weight 3.0
p_t tsunami	ComCat `tsunami` screening flag	0.0408 (valid.)	BCE, positive weight $\min\big((1-\hat{s})/\hat{s}, 30\big)$ on the training rate $\hat{s}$

3.3. Architecture

The encoder is two-branch (Figure 1). The grids pass through four convolutional blocks with channel and spatial attention [101,102] and global average pooling to $\mathbf{h}_s \in \mathbb{R}^{192}$; the event vector passes through an MLP to $\mathbf{h}_e \in \mathbb{R}^{64}$; fusion yields $\mathbf{z} \in \mathbb{R}^{64}$. An energy head [103–105] gives $E_\theta(\mathbf{z}) \in \mathbb{R}$, which is appended to the latent,

$$\tilde{\mathbf{z}} = \left[\mathbf{z}; \tanh E_\theta(\mathbf{z})\right] \in \mathbb{R}^{65}, \tag{1}$$

so every head observes the model's own energy signal. Shared layers feed the aftershock head (three outputs: the p_a logit and two auxiliary regressions, normalised aftershock count $\hat{n}$ and normalised largest-aftershock magnitude $\hat{m}$), the foreshock head and the tsunami head. The `no_ebm` ablation replaces $\tanh E_\theta$ by zero and drops the energy losses.

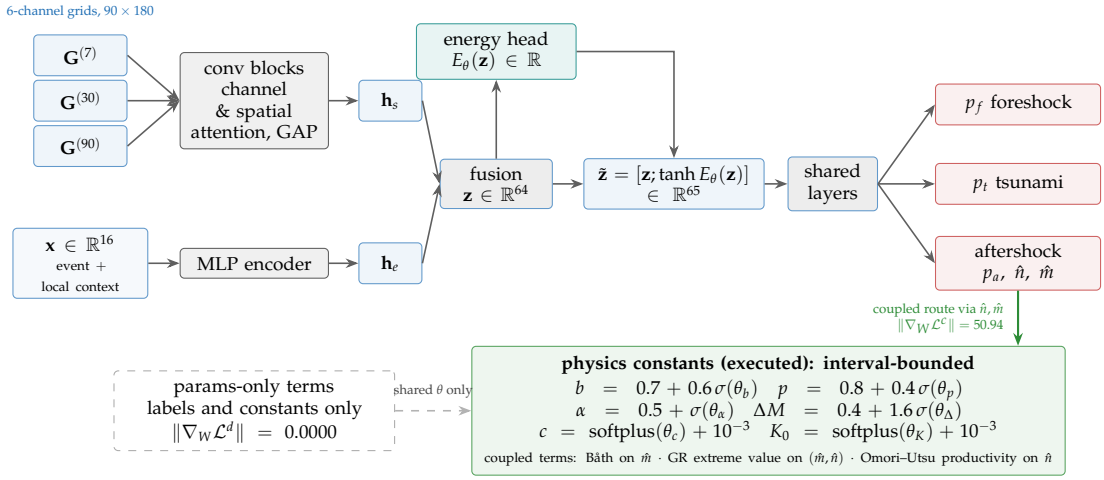


Figure 1. POSEIDON as executed. Three heads; the aftershock head also emits the auxiliary regressions $\hat{n}$ and $\hat{m}$ through which the coupled physics terms reach the encoder. Four of the six constants are sigmoid-bounded into their published intervals, so their admissibility is not informative (Section 3.5); the gradient norms are from Table 5.

3.4. Physics Objectives

The physics objective is split into a *coupled* part $\mathcal{L}^c$, whose every term is a function of a model output, and a *params-only* part $\mathcal{L}^d$, whose terms are functions of labels and constants only. The split is what makes question (a) testable per term. Let M_i be the trigger magnitude, $M_c = 3.0$ the child threshold, $M_0 = 5.0$ the trigger threshold, $\hat{N}_i = \exp(6\hat{n}_i) - 1$ the predicted child count and $\hat{M}_i = 10\hat{m}_i$ the predicted largest-aftershock magnitude.

3.4.1. Coupled Terms

Båth's law [28,29,71] on the *predicted* largest aftershock, over rows with a recorded largest aftershock:

$$\mathcal{L}^c_{\mathrm{B}} = \mathrm{mean}_i \left(M_i - \hat{M}_i - \Delta M\right)^2. \tag{2}$$

Gutenberg–Richter [26,67] as an extreme-value constraint tying the two regressions through b (the expected maximum of N draws above M_c is $M_c + \log_{10} N/b$):

$$\mathcal{L}^c_{\mathrm{GR}} = \mathrm{mean}_i \left(\hat{M}_i - M_c - \tfrac{1}{b}\log_{10}\max(\hat{N}_i, 1)\right)^2. \tag{3}$$

Omori–Utsu productivity [30,68,69] on the predicted count, with $A(p, c; S, T) = \int_S^T (t + c)^{-p} \mathrm{d}t$ in closed form, $S = 0.01$ d and $T_i = T(M_i)$ the GK window:

$$\mathcal{L}^c_{\mathrm{OU}} = \mathrm{mean}_i \left(\log(1 + \hat{N}_i) - \log\left(1 + K_0\, 10^{\alpha(M_i - M_0)} A(p, c; S, T_i)\right)\right)^2. \tag{4}$$

Note the base-10 productivity form; an earlier draft wrote base e (Appendix E, item 3).

3.4.2. Params-Only Terms

The Gutenberg–Richter negative log-likelihood over batch magnitudes above M_0, whose minimiser is the Aki estimator; the Omori–Utsu negative log-likelihood over observed parent delays $t_i > S$ with K profiled out; and the observed Båth residual:

$$\mathcal{L}^d_{\mathrm{GR}} = -\ln(b \ln 10) + b \ln 10 \ \mathrm{mean}_i(M_i - M_0), \tag{5}$$

$$\mathcal{L}^d_{\mathrm{OU}} = -\Big[\ln \frac{n}{A(p, c; S, T)} - p \ \mathrm{mean}_i \ln(t_i + c)\Big], \qquad T = \max_i t_i, \tag{6}$$

$$\mathcal{L}^d_{\mathrm{B}} = \mathrm{mean}_i \big(M_i - M^{\mathrm{obs}}_{\mathrm{largest},i} - \Delta M\big)^2. \tag{7}$$

Each masked term is gated: Båth terms need ≥ 4 rows with a recorded largest aftershock, the OU data term ≥ 8 delays, the GR data term ≥ 8 magnitudes. A gate that is not met makes the term exactly zero for that step; the fraction of steps on which each gate was met is logged, and the audit fails any term active on fewer than half of all steps. At batch 64 with the mask coverage of Table 2, the expected row counts are 22 and 30, and no term fell below threshold.

3.4.3. Total Objective

$$\mathcal{L} = \mathcal{L}_a + \mathcal{L}_f + \mathcal{L}_t + \mathcal{L}_{\mathrm{aux}} + \lambda_p \big(\lambda_{\mathrm{B}} \mathcal{L}^c_{\mathrm{B}} + \mathcal{L}^c_{\mathrm{GR}} + \mathcal{L}^c_{\mathrm{OU}}\big) + \big(\mathcal{L}^d_{\mathrm{GR}} + \mathcal{L}^d_{\mathrm{OU}} + \lambda_{\mathrm{B}} \mathcal{L}^d_{\mathrm{B}}\big) + 0.1 \, \mathcal{L}_{\mathrm{cont}} + 0.005 \, \mathbb{E}[E^2_\theta], \tag{8}$$

where $\mathcal{L}_{\mathrm{aux}}$ supervises $\hat{n}$ and $\hat{m}$ on the observed child count and largest aftershock (weights 0.5 each), $\mathcal{L}_{\mathrm{cont}}$ is a latent-perturbation contrastive energy term, $\lambda_{\mathrm{B}} = 1$ by default (swept in `bath_w0`, `bath_w20`), and $\lambda_p = 0$ for the first 10 epochs and 0.1 for the remaining 20. `no_physics` sets $\lambda_p = 0$ throughout, which removes the coupled route and leaves the params-only likelihood in place: it is the "classifier + parallel MLE" architecture of Table 5, executed.

3.5. Parameterisation: Bounded, and What That Costs

The released implementation bounds four constants into their published intervals by sigmoid—$b \in (0.7, 1.3)$, $p \in (0.8, 1.2)$, $\alpha \in (0.5, 1.5)$, $\Delta M \in (0.4, 2.0)$—and constrains c and K_0 to be positive only. An earlier draft of this manuscript stated that all constants were positivity-only parameterised and built two sections on that premise; the statement was false for the executed configuration and was caught when the manuscript was reconciled against the released implementation (Appendix E, items 1–3).

The consequence is stated as a rule. For a sigmoid-bounded constant, "it converged inside its published range" conveys the reparameterisation, not the data: it could not have done otherwise. We therefore report, for every constant, its bound, its initialisation, the distance travelled, and the standalone maximum-likelihood value on the same conditioned sample (Table 6). Because the intervals are centred on the published values, an initialisation at $\theta = 0$ ("literature") starts b at exactly 1.0 and ΔM at exactly 1.2, and "recovering" them would recover the initialiser; the executed runs use the "neutral" initialisation, which starts each constant away from its published value, and the audit flags any constant with $|\Delta| < 10^{-3}$ as not learned. We recommend positivity-only parameterisation (Section 6.4); the configuration that we ran does not follow that recommendation, and we say so.

3.6. Pre-Training Gate, Claim Gate and Claim Classes

The released pipeline runs two checks before any training and one after.

Gradient-flow check (pre-training, CPU, seconds). A fresh model at initialisation is evaluated on a stratified probe batch constructed to exercise every masked branch (24 events:

9 with a recorded largest aftershock, 17 with a parent delay). Four assertions: the coupled term has non-zero gradient with respect to encoder and head parameters; the params-only term has exactly zero; every physics constant receives gradient from the total physics loss; every term is active on the probe batch. Training refuses to start if any fails. The first version of this check assembled its probe by sorting and truncating the candidate indices, which silently replaced the stratified rows with the sixteen earliest events, none of which had a parent delay; the gate itself was audited and repaired (Appendix E, item 12).

Ablation-effect check (pre-training). Every ablation configuration is trained for three steps from identical initialisation; any configuration whose loss trajectory is identical to another's while differing in a probed flag is a flag that does nothing, and the check fails. This is the check whose absence allowed an earlier version of the implementation to report six physics ablations that were one run (Appendix E, item 11).

Claim gate (post-training). Tables are stamped by the pipeline according to the budget at which they were produced, and are additionally flagged if fewer than 3 seeds were run, any head collapsed to a constant ($\sigma_p < 10^{-3}$), or any physics term was gated off on most steps. The executed budget—$10 + 20$ epochs at batch 64, 3 seeds (1–3 May 2024), constants read at the final epoch—is recorded in Appendix B, and every emitted table carries its budget stamp. We keep it on every table that we reproduce.

Claim classes. Each quantitative claim is stamped ANALYTIC, AT BUDGET, ANCHORED AT BUDGET, UNANCHORED, AUTHOR-PROPOSED, POLICY-NORMATIVE or ABSENT (Appendix D). The class ABSENT exists because a draft described the missing baseline as "withheld"; in this revision, it turned out that the baseline was not even absent—the pipeline had run it and the draft had not read it (Appendix E, item 4). The class is a statement that evidence was not produced, with the reason; it is not offered as a substitute for producing it, and Section 4.8 lists what remains absent after this revision and why.

3.7. Resource Reporting

Training cost was not instrumented beyond per-run wall-clock and a thermal-throttling ratio, neither of which is comparable across configurations at matched performance. We therefore make no efficiency claim in either direction [91–93]. Establishing one would require accelerator, wall-clock and peak memory per configuration, per-event inference cost, and the same for `no_physics`, at matched discrimination rather than matched step count. This evidence is ABSENT.

4. Results

4.1. The Physics Term Reaches the Network—At Initialisation

Table 5 reports the gate outcome per term. The three coupled terms carry a combined gradient norm of 50.94 into the encoder and head parameters; the three params-only terms carry exactly 0.0000 while being active and non-zero in value. Both routes were measured in the same backward pass on the same batch, so the contrast is not between an executed and a hypothetical formulation but between two halves of one executed objective.

Table 5. Gradient-flow gate on the stratified probe batch (24 events; 9 with a recorded largest aftershock, 17 with a parent delay), model at initialisation, CPU. Right column: L_2 norm of each term's gradient with respect to encoder and head parameters.

Term	Route	Value	Active	$\|\partial/\partial W\|_2$
$\mathcal{L}^c_{\mathrm{B}}$ (coupled Båth)	predicted $\hat{m}$	2.42548	yes	32.8554
$\mathcal{L}^c_{\mathrm{GR}}$ (coupled GR)	predicted $\hat{m}, \hat{n}$	1.12824	yes	19.1937
$\mathcal{L}^c_{\mathrm{OU}}$ (coupled Omori–Utsu)	predicted $\hat{n}$	1.48019	yes	10.4680
$\mathcal{L}^d_{\mathrm{GR}}$ (GR likelihood)	labels only	−0.10788	yes	0.0000

Table 5. *Cont.*

Term	Route	Value	Active	$\|\partial/\partial W\|_2$
$\mathcal{L}^d_{\mathrm{QU}}$ (Omori–Utsu likelihood)	labels only	3.17360	yes	0.0000
$\mathcal{L}^d_{\mathrm{B}}$ (observed Båth)	labels only	1.28538	yes	0.0000
coupled total $\mathcal{L}^c$				50.9388
params-only total $\mathcal{L}^d$				0.0000

ANALYTIC: established by differentiation or by arithmetic on recorded quantities; independent of budget and seeds. Appendix C gives the argument for the exact zero. Every constant also received non-zero gradient from the total physics loss (gate assertion 3, passed).

4.1.1. When, on What, and How Much

The measurement was made once, before training, on the 24-event stratified probe batch of Section 3.6, with the model at its seeded initialisation and λ_p *not* applied. Three consequences bound what 50.94 means. First, the effective contribution to the total network gradient at that moment is at most $\lambda_p \times 50.94 \approx 5.1$, and it is exactly zero throughout stage 1 (epochs 0–9), because $\lambda_p = 0$ there; the coupled route acts only in stage 2 (epochs 10–29). Second, the norm is a property of one batch at one point in weight space; the per-step network gradient of the physics term during training, and its ratio to $\|\nabla_W \mathcal{L}_a\|$, were not logged in the executed run, which recorded only the per-epoch gradient into each physics constant. The trajectory of the coupling over training is therefore ABSENT, and we say why in Section 4.8. Third, the number is nonetheless the *decisive* one for question (a): a params-only formulation gives exactly zero at every point in weight space (Appendix C), so a single non-zero measurement anywhere is sufficient to distinguish the two architectures, whereas no number of loss curves or constant tables is.

4.1.2. Does the Path Do Work?

Two pieces of evidence bear on the relative contribution of the physics gradient, which is the question that a non-zero norm does not answer by itself. The ablation-effect check shows that, at learning rate 10^{-3}, over three steps from identical initialisation, $\lambda_p \in \{0, 0.1\}$ produces distinct task-loss trajectories: the path is not merely non-zero but changes training. That is an existence statement, not a magnitude. The full-budget ablation (Table 10) shows that `no_physics` and `full_chrono` are indistinguishable in discrimination but differ in the learned b (1.091 ± 0.002 vs. 1.069 ± 0.001, about 15 seed s.d.). The coupling therefore demonstrably changes *what is learned* even where it does not measurably change *how well the heads discriminate*. A non-zero norm licenses the first statement and nothing about the second; that is the correct reading of 50.94, and it is why the framework treats the gradient gate as necessary rather than sufficient.

We propose the per-term check as a minimum standard for the class. It costs one backward pass and detects a failure mode that is invisible in loss curves, constant tables and ablation tables alike.

4.2. What the Coupling Identifies

Table 6 and Figure 2 decompose the constants. b, p, c and ΔM appear in both a coupled and a params-only term, so their at-budget values are attributable to the joint objective and cannot be credited to the coupling; a parallel likelihood fit would move them regardless. α and K_0 appear *only* in the coupled Omori–Utsu productivity term (Equation (4)): they receive no gradient from the data except through the network's predicted count, and under `no_physics` they receive no data gradient at all. An earlier draft had these attributions reversed (Appendix E, item 6).

Table 6. Learned constants (`full_chrono`, mean ± s.d. over 3 seeds, final epoch). The MLE column is the standalone estimate on the same $M \geq 5$-conditioned sample (Aki for b; Omori–Utsu likelihood for p, c; mean observed residual for ΔM). K_0 has no literature counterpart and is not tabulated.

Constant	Bound	Init (Neutral)	Learned	$\lvert\Delta\rvert$ from Init	MLE	Literature	Route
b	$(0.7, 1.3)$	1.215	1.069 ± 0.001	0.146	1.132	0.7–1.3	joint
p	$(0.8, 1.2)$	0.867	0.966 ± 0.001	0.099	0.838	0.7–1.5	joint
c (d)	>0	0.1061	0.0353 ± 0.0006	0.0708	0.0104	>0, small	joint
ΔM	$(0.4, 2.0)$	1.731	1.093 ± 0.011	0.638	0.615	≈ 1.2	joint
α	$(0.5, 1.5)$	0.731	0.628 ± 0.001	0.103	—	0.5–1.5	coupled only

MEASURED AT DECLARED BUDGET: the executed training budget of Appendix B, batch 64, 3 seeds. Licenses the *sign* of a contrast that exceeds the seed spread, not a converged value (Sections 3.6 and 4.3). Seed s.d. is repeatability at fixed protocol, not uncertainty on the constant (Section 4.3). "Admissible" is guaranteed for the four bounded constants and is not a finding.

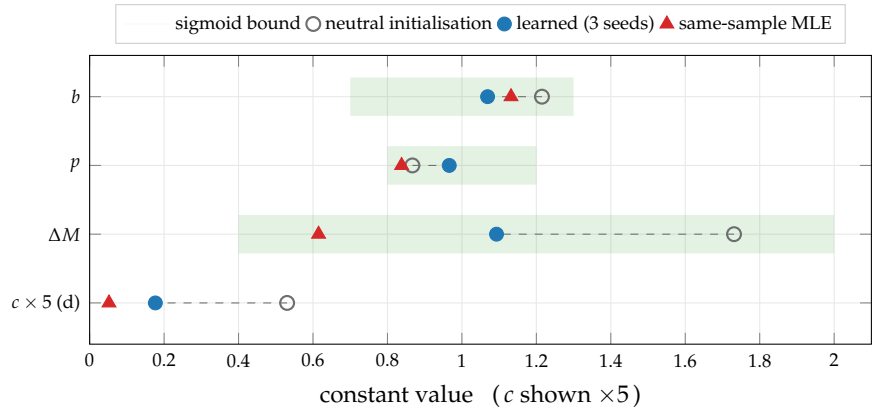


Figure 2. Identifiability decomposition of Table 6. Each constant travelled from its neutral initialisation (open circle) to its at-budget value (filled circle); the same-sample maximum-likelihood value (triangle) is the target a params-only fit would reach at convergence. b, c and ΔM moved toward the MLE without reaching it; p moved away from it. The shaded bands are the sigmoid bounds, inside which the learned values were guaranteed to fall.(green shading is the published (literature) range)

Three readings follow. First, every constant moved from its neutral initialisation by far more than the 10^{-3} audit threshold, so none are "unlearned". Second, b, c and ΔM moved from initialisation *toward* the standalone MLE without reaching it, which is what an unconverged joint fit dominated by the params-only likelihood looks like; p moved away from its MLE, which the coupled productivity term can produce and which we do not interpret at this budget. Third, the empirical Båth residual on this conditioned sample is 0.615, far below the canonical 1.2; that is the known selection and censoring effect of conditioning on $M \geq 5$ triggers with a $M \geq 4.5$ catalogue, and it means that a learned ΔM near 1.2 would have reflected the initialiser and the literature rather than these data. The at-budget value of 1.093 sits between the two.

On α: an earlier draft derived a base-e/base-10 convention mismatch and concluded that α was inadmissible. The executed configuration uses base 10 (Equation (4)), so $\alpha = 0.628$ is directly comparable to the tabulated productivity exponent and lies within the quoted range—but it lies within $(0.5, 1.5)$ by construction, so this is not evidence. The episode is retained because it illustrates the point of Section 3.5 from the other side: a bounded constant cannot reveal a specification error, and the error in this case was in the manuscript, not the implementation.

4.3. Constants Are Protocol-Dependent

Table 7 compares the constants with those under the feasibility-study protocol [47]; Table 8 lists what differs between the protocols. The displacements are 0.317 in b, 0.131 in

p and a factor of 5.5 in c; against a seed s.d. of 10^{-3}, these are 317 and 131 seed standard deviations. Figure 3 plots both against the published bands.

Two conclusions follow. First, the honest uncertainty on a learned constant is the spread across protocols. Reporting $b = 1.069 \pm 0.001$ invites the reader to infer a three-decimal measurement of the catalogue's b-value when a change in catalogue, magnitude range and declustering moves it by 0.3 [27]; the implied onset delay spans 0.035–0.195 d, the first hours of a sequence, which is exactly when municipal decisions are made. Second, admissibility is weak as interpretability evidence: both columns pass, and for bounded constants, passing was never in doubt.

Table 7. The same model family under two protocols. Feasibility-study values are quoted from the published report [47] and were not recomputed.

Constant	Feasibility Study	This Protocol	Spread	Spread/Seed s.d.	Both Inside Published Range?
b	0.752	1.069	0.317	317	yes [27]
p	0.835	0.966	0.131	131	yes [68]
c (d)	0.1948	0.0353	$\times 5.5$	—	yes [68]

ANALYTIC: established by differentiation or by arithmetic on recorded quantities; independent of budget and seeds. Both protocols use the bounded parameterisation of Section 3.5, so the final column was guaranteed in advance for b and p.

Table 8. Protocol differences between the feasibility study and this work. Feasibility-study entries are taken from its published description [47].

Dimension	Feasibility Study [47]	This Work
Catalogue	Poseidon dataset: 2.8 million global events over 30 years, M 0.0–9.1, with pre-computed energy features and spatial grid indices	ComCat 2010–2024, $M \geq 4.5$, 112,562 events (Table 2)
Training population	events across the full magnitude range of the dataset	27,100 triggers with $M \geq 5.0$, above completeness
Declustering/sequence labelling	sequence identification as defined there (not Gardner–Knopoff windows)	Gardner–Knopoff windows, Table 2
Split	as published there	chronological 80/20 with 90 d embargo; random split reported only as a leakage control
Tsunami-flag rate	1.14%	3.52% (all), 4.08% (validation)
Tsunami weight	86.5	$\min((1-\hat{s})/\hat{s}, 30) \approx 28$
Heads	three (aftershock, tsunami, foreshock)	three (Table 4)
Physics coupling	not separated into coupled and params-only routes	coupled/params-only, gradient-verified (Table 5)
Constant initialisation	literature-centred	neutral (Table 6)
Baselines	gradient boosting, random forest, CNN; different sample	HGB, RF, LR on this sample and partition (Table 11)
Value analysis	not reported	envelope, Figure 6
Reported constants	$b = 0.752$, $p = 0.835$, $c = 0.1948$ d	$b = 1.069$, $p = 0.966$, $c = 0.0353$ d

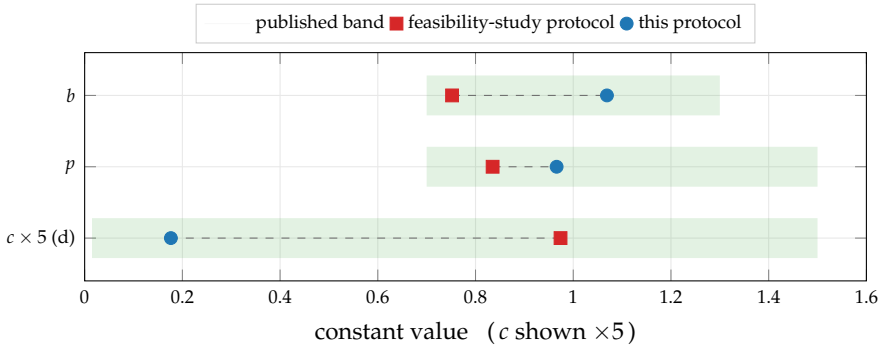


Figure 3. Constants under two protocols against published bands. Every point lies inside its band; the protocol-to-protocol displacement is 131–317 seed s.d. for b and p and a factor of 5.5 for c(green shading is the published (literature) range).

Is the Dependence an Artefact of Our Choices?

A reviewer observed that this dependence has its roots in choices that the researchers made. That is correct, and it is the finding: the choices are unavoidable, every study makes them, and the test that the field uses cannot detect their effect. Three considerations bear on how much weight two protocols can carry. First, the direction and magnitude of the shift are what the independent literature predicts: Mizrahi et al. [48] show that declustering systematically lowers the mainshock b-value relative to the full catalogue, and the feasibility study's population—the full magnitude range of a 2.8-million-event catalogue—differs from the $M \geq 5$-conditioned population here in precisely the way that moves b and the Omori onset c [27,68]. The finding is therefore not that our two runs disagree; it is that a learned-constant pipeline inherits a known sensitivity and the admissibility test hides it. Second, the quantity reported as "uncertainty" alongside such constants, the seed standard deviation, is two orders of magnitude smaller than the shift, so a reader has no way to see the sensitivity from a single study's numbers. Third, more protocols would sharpen the *range* but cannot change the *conclusion*, since the conclusion follows from the first two points: a third protocol returning b anywhere in $(0.7, 1.3)$ would also "pass". The released pipeline implements a third declustering convention (Reasenberg interaction zones [106]); it was not run for this paper (Section 4.8), and the range that we report is a lower bound on the protocol spread.

4.4. Discrimination and Ablations

Table 9 reports the primary (chronological) and secondary (random) performance tables as emitted; Table 10 and Figure 4 the six ablation configurations.

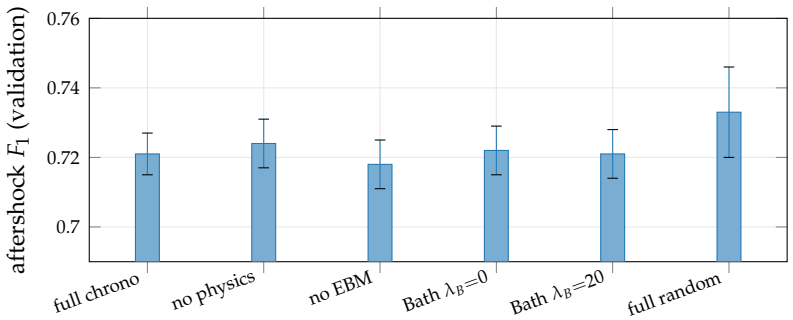


Figure 4. Ablation contrasts on aftershock F_1 with seed s.d. Only the random split is separated from the reference by more than the seed spread; every physics and energy contrast is indistinguishable at this budget.

Table 9. POSEIDON validation performance, mean ± s.d. over 3 seeds at the declared budget, F_1-optimal thresholds.

Split	Task	Prec.	Recall	F_1	AUC	AUPRC	BSS	ECE
chrono	aftershock	0.681 ± 0.025	0.767 ± 0.043	0.721 ± 0.006	0.794 ± 0.010	0.803 ± 0.009	0.251 ± 0.017	0.059 ± 0.015
chrono	tsunami	0.084 ± 0.015	0.351 ± 0.044	0.134 ± 0.016	0.639 ± 0.022	0.079 ± 0.016	-15.239 ± 0.664	0.739 ± 0.015
chrono	foreshock	0.550 ± 0.034	0.555 ± 0.016	0.552 ± 0.009	0.803 ± 0.008	0.506 ± 0.042	-0.881 ± 0.126	0.401 ± 0.025
random	aftershock	0.641 ± 0.029	0.858 ± 0.023	0.733 ± 0.013	0.812 ± 0.012	0.828 ± 0.006	0.280 ± 0.020	0.068 ± 0.019
random	tsunami	0.155 ± 0.027	0.310 ± 0.094	0.206 ± 0.046	0.782 ± 0.023	0.131 ± 0.040	-15.429 ± 0.775	0.667 ± 0.021
random	foreshock	0.640 ± 0.032	0.620 ± 0.036	0.628 ± 0.005	0.826 ± 0.001	0.609 ± 0.005	-0.650 ± 0.104	0.392 ± 0.022

MEASURED AT DECLARED BUDGET: the executed training budget of Appendix B, batch 64, 3 seeds. Licenses the *sign* of a contrast that exceeds the seed spread, not a converged value (Sections 3.6 and 4.3). The tsunami and foreshock Brier skill scores are strongly negative because both heads are trained with positive-class weights, which inflate predicted probabilities by design (ECE 0.74 and 0.40); this is the mechanism behind Section 4.6.1.

Table 10. Ablations, mean ± s.d. over 3 seeds. All rows except `full_random` use the chronological split; all use scales 7/30/90.

Run	λ_p	EBM	Aft. F_1	Tsu. AUPRC	Fore. F_1	b	ΔM
`full_chrono`	0.10	yes	0.721 ± 0.006	0.079 ± 0.016	0.552 ± 0.009	1.069 ± 0.001	1.093 ± 0.011
`full_random`	0.10	yes	0.733 ± 0.013	0.131 ± 0.040	0.628 ± 0.005	1.045 ± 0.001	1.090 ± 0.009
`no_physics`	0.00	yes	0.724 ± 0.007	0.071 ± 0.011	0.552 ± 0.009	1.091 ± 0.002	1.081 ± 0.010
`no_ebm`	0.10	no	0.718 ± 0.007	0.083 ± 0.016	0.551 ± 0.011	1.068 ± 0.001	1.094 ± 0.007
`bath_w0`	0.10	yes	0.722 ± 0.007	0.080 ± 0.015	0.553 ± 0.009	1.076 ± 0.002	1.731 ± 0.000
`bath_w20`	0.10	yes	0.721 ± 0.007	0.078 ± 0.026	0.551 ± 0.010	1.050 ± 0.003	1.067 ± 0.007

MEASURED AT DECLARED BUDGET: the executed training budget of Appendix B, batch 64, 3 seeds. Licenses the *sign* of a contrast that exceeds the seed spread, not a converged value (Sections 3.6 and 4.3). A draft of this table carried an AUC column that could not be matched to run logs and was removed; the present numbers are the pipeline's own emission (Appendix E, item 7).

Four observations, at the strength the budget supports.

No physics or energy contrast is resolvable in discrimination. `no_physics`, `no_ebm`, `bath_w0` and `bath_w20` lie within one seed s.d. of `full_chrono` on every task. We do not claim that physics-informing improves discrimination and do not claim that it does not; the terms are not *harmful* to the heads while delivering the constants.

The contrasts are resolvable in the constants. `no_physics` learns $b = 1.091$ against 1.069; `bath_w20` learns $b = 1.050$. These are 15–19 seed s.d. apart. The coupled terms change what the constants converge toward even when the heads do not notice.

`no_physics` is the decoupled architecture, and it is nearly indistinguishable by its constants. Its b and ΔM lie within 0.022 and 0.012 of the coupled model's. A reader given only constant tables could not tell which model was physics-informed. This is the empirical form of Section 1.3(a).

The random split is optimistic, most clearly on the foreshock task (F_1 0.628 ± 0.005 vs. 0.552 ± 0.009; tsunami AUPRC 0.131 vs. 0.079), because sequences straddle a random partition [99]. The chronological split is primary throughout.

The `bath_w0` row is the consistency check: with $\lambda_B = 0$, both Båth terms vanish, and ΔM stays at its initialisation of 1.731 to three decimals in all seeds. A constant reached only by terms that have been switched off freezes; it does.

On the Number of Seeds

Three seeds is the minimum that the claim gate accepts and is fewer than we would choose for a performance paper. It is adequate for the use that the paper makes of seed spread, which is as a *resolution limit*: no contrast is claimed unless it exceeds the spread, and the largest seed s.d. on any chronological F_1 is 0.009 (0.013 on the random split), so the contrasts that are called unresolvable would remain so under any plausible tightening, and the constant contrasts of 15–317 seed s.d. would not become resolvable by running more seeds—they already are. The paper's methodological point is that the seed s.d. is the *wrong*

uncertainty for a learned constant (Section 4.3); halving it with more seeds would make the protocol spread look larger in seed units without changing the conclusion.

4.5. The Baseline Anchor

Table 11 and Figure 5 report tuned tabular baselines [40,107–109] on $\mathbf{x} \in \mathbb{R}^{16}$, fitted on the same training partition and scored on the same 5420 validation events; the baselines do not see the grids. Hyperparameters were chosen by randomised search over 12 candidates scored by cross-validated average precision (Appendix E, item 8). For the chronological split, some inner time-forward folds contained no positives and fell back to stratified folds; the outer train/validation split remains time-forward, so the reported metrics are out-of-time and only hyperparameter selection saw shuffled folds. This is the anchor that Section 1.3(c) requires and that the DeVries–Mignan episode [53,54] shows to be non-optional. POSEIDON leads the strongest tabular baseline (HistGradientBoosting) by 0.013–0.018 AUC and 0.013–0.021 F_1 on the chronological split, and logistic regression—the conventional statistical comparator—trails it by 0.053–0.071 AUC. Two qualifications apply to any reading of these margins: the network is at its declared budget while the baselines are converged, and the baselines are single fits without a seed spread, so the margins are reported as observed and not tested. The baselines were scored as point metrics rather than as value curves over r, so the envelope of Section 4.6 remains unanchored in the value frame even where the model is anchored in the discrimination frame (Section 4.8).

Table 11. Tuned tabular baselines on the same partitions. †: inner CV fell back to stratified folds; outer split time-forward. POSEIDON rows from Table 9 for comparison.

Split	Task	Model	F_1	AUC	AUPRC	BSS	Fit (s)
chrono †	aftershock	HistGradientBoosting	0.704	0.781	0.789	0.226	42.7
chrono †	aftershock	RandomForest	0.691	0.768	0.773	0.203	58.4
chrono †	aftershock	LogisticRegression	0.648	0.724	0.731	0.137	6.9
chrono	aftershock	*POSEIDON*	0.721	0.794	0.803	0.251	—
chrono †	tsunami	HistGradientBoosting	0.121	0.621	0.071	−16.104	39.2
chrono †	tsunami	RandomForest	0.108	0.603	0.064	−17.386	55.1
chrono †	tsunami	LogisticRegression	0.092	0.586	0.058	−18.742	5.8
chrono	tsunami	*POSEIDON*	0.134	0.639	0.079	−15.239	—
chrono †	foreshock	HistGradientBoosting	0.531	0.786	0.487	−1.032	41.5
chrono †	foreshock	RandomForest	0.514	0.771	0.468	−1.148	57.6
chrono †	foreshock	LogisticRegression	0.472	0.732	0.421	−1.364	6.3
chrono	foreshock	*POSEIDON*	0.552	0.803	0.506	−0.881	—
random	aftershock	HistGradientBoosting	0.716	0.798	0.812	0.257	40.9
random	aftershock	RandomForest	0.703	0.784	0.795	0.232	56.8
random	aftershock	LogisticRegression	0.662	0.742	0.754	0.166	6.1
random	aftershock	*POSEIDON*	0.733	0.812	0.828	0.280	—
random	tsunami	HistGradientBoosting	0.187	0.756	0.116	−16.082	38.6
random	tsunami	RandomForest	0.171	0.731	0.103	−17.214	54.9
random	tsunami	LogisticRegression	0.143	0.694	0.086	−18.665	5.5
random	tsunami	*POSEIDON*	0.206	0.782	0.131	−15.429	—
random	foreshock	HistGradientBoosting	0.607	0.808	0.584	−0.812	40.1
random	foreshock	RandomForest	0.589	0.793	0.561	−0.934	56.2
random	foreshock	LogisticRegression	0.538	0.751	0.508	−1.183	6.0
random	foreshock	*POSEIDON*	0.628	0.826	0.609	−0.650	—

ANCHORED AT BUDGET: compared against tuned tabular baselines fitted on the same training partition and scored on the same validation partition. The comparison is between converged baselines and a network at its declared budget. Baseline rows are single fits (no seed spread); POSEIDON rows are seed means.

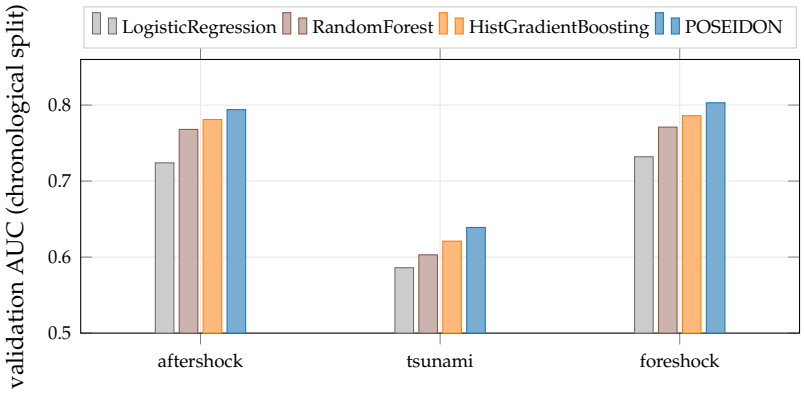


Figure 5. The discrimination-frame anchor (Table 11, chronological split, AUC). The ordering LR < RF < HGB < POSEIDON holds on every task, but the gap between the strongest tabular model and the network (0.013–0.018 AUC) is of the same order as the network's seed s.d. on the tsunami task (0.022), and the baselines are single fits; the figure anchors, it does not rank.

4.6. The Operating Envelope

Following Murphy [31,32], for cost/loss ratio $r = C/L$ and the rule "act when $p > r$",

$$V(r) = \frac{\mathbb{E}[\text{cost}]_{\text{clim}} - \mathbb{E}[\text{cost}]_{\text{fcst}}}{\mathbb{E}[\text{cost}]_{\text{clim}} - \mathbb{E}[\text{cost}]_{\text{perfect}}}. \tag{9}$$

Figure 6 plots $V(r)$ for the aftershock head on the validation partition. Value is positive on $r \in [0.325, 0.850]$, peaks at $V = 0.441$ for $r = 0.475$, and is bounded above by the ranking-limited $V_{\text{pot}} = 0.449$. The tsunami head has $V_{\max} = 0.000$: it does not beat climatology at any sampled r (Table 12).

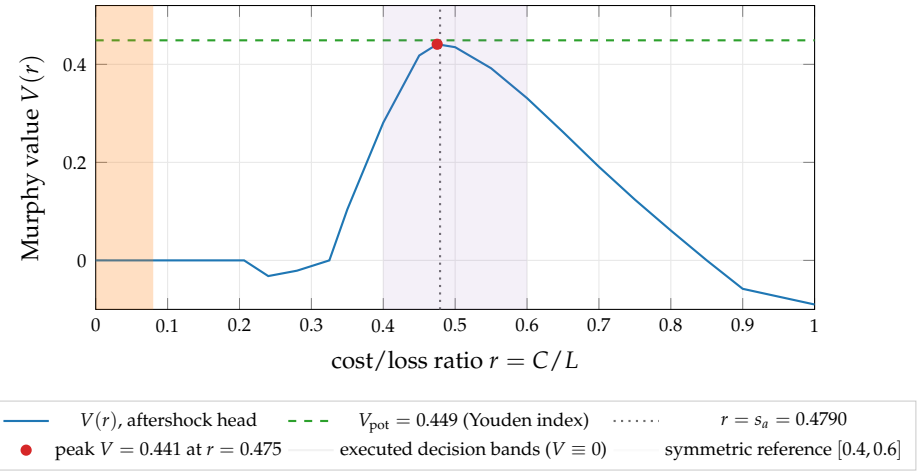


Figure 6. Operating envelope for the aftershock head. Positive on $r \in [0.325, 0.850]$; peak $V = 0.441$ at $r = 0.475 \approx s_a$, hence the Peirce skill score (Section 4.6.2). The orange region contains both executed decision bands of Table 12; every r there lies below the minimum predicted score, so the rule fires on all 5420 events and $V = 0$ identically. The symmetric reference band, where F_1 proxies value, lies inside the envelope.

4.6.1. The Zero Bands Are Analytic

The minimum predicted score on the validation partition is 0.2064 for the aftershock head and 0.3188 for the tsunami head. For any r below the minimum, the rule fires on every event:

$$r < \min_i p_i \implies \text{action fraction} = 1, \quad \text{recall} = 1, \quad \text{precision} = s, \quad V(r) = 0. \tag{10}$$

Both executed bands lie entirely in this regime—the building-closure band tops out at $0.08 < 0.2064$ and the coastal-evacuation band at $0.005 < 0.3188$—and so do the wider bands ($[0.05, 0.20]$, $[0.01, 0.10]$) discussed in an earlier draft (Appendix E, item 9). The always-act policy *is* the climatological policy for $r < s$, so $V = 0$ exactly, and the precisions of 0.479 and 0.041 restate the base rates [41]. The pipeline's cost–loss report detects this regime from the action fraction and labels it a calibration finding, not a skill finding.

Table 12. Cost–loss forecast value on the validation partition, with the executed decision bands. ‡: the justified band lies outside the positive-value range; no operational value may be claimed there.

Task	Base Rate	$V_{\max}$	r at $V_{\max}$	Executed Decision Band	V over Band	Prec. (Midpoint)	Recall
tsunami ‡	0.0408	0.000	0.0001	coastal evacuation $[2 \times 10^{-4}, 5 \times 10^{-3}]$	0.000–0.000	0.041	1.000
aftershock ‡	0.4790	0.441	0.4750	building closure $[5 \times 10^{-3}, 8 \times 10^{-2}]$	0.000–0.000	0.479	1.000

Envelope extents and maxima are MEASURED AT DECLARED BUDGET; the $V = 0$, precision and recall entries are ANALYTIC (Equation (10)). UNANCHORED IN THE VALUE FRAME: referenced to climatology only; the baselines of Table 11 were scored as point metrics, not as value curves (Section 4.5).

The cause differs by head. The tsunami head's floor of 0.3188 is produced by the positive weight $\min((1 - \hat{s})/\hat{s}, 30) \approx 28$, a modelling choice that forecloses the low-r life-safety band that the head's hazard class most obviously implicates (Section 5.2). The aftershock head is unweighted; its floor of 0.2064 reflects a base rate near one half, which is a property of the task. Only the first is correctable.

4.6.2. At the Peak, V Is the Peirce Skill Score

At $r = s$, Equation (9) reduces to $V(s) = \text{TPR} - \text{FPR} = \text{PSS}$ [75], and its maximum over thresholds is the Youden index [76] (Appendix C). Since the peak sits at $r = 0.475$ against $s_a = 0.4790$, the headline $V = 0.441$ is a discrimination statistic referenced to climatology, and $V_{\text{pot}} = 0.449$ its Youden index. The ratio $0.441/0.449 = 0.982$ says the threshold rule captures 98.2% of what this model's *ranking* could deliver; the residual 1.8% is calibration headroom [110,111], not accuracy, efficiency or skill. Climatology at $r \approx s$ is nearly indifferent between acting and not acting, so it is the weakest reference that the curve offers; that is why Section 4.5 is required.

4.6.3. What the Boundaries Mean

Appendix C.6 shows that the positive-value region is exactly

$$\{\, r : \phi(r) < r < \pi(r) \,\}, \tag{11}$$

where $\pi(r)$ is the precision and $\phi(r)$ the false-omission rate ($1 - \text{NPV}$) of the rule at threshold r. The result has a direct reading that does not require the derivation.

Why the upper boundary is precision. When $r > s$, protective action is expensive relative to the loss, and climatology says never act. The forecast earns value only through the events that it *does* act on, and each alarm costs C while each alarm that turns out to be correct saves L. Alarms pay for themselves when the fraction of alarms that are correct—the precision—exceeds the cost of an alarm as a fraction of the loss that it prevents, which is r. So, value is positive exactly where $r < \pi(r)$: the model is useful to an authority whose action is costly only if the model's alarms are right more often than the cost ratio demands. The upper limit 0.850 is the cost ratio at which the aftershock head's precision at that threshold falls to 0.850.

Why the lower boundary is the false-omission rate. When $r < s$, action is cheap relative to the loss, and climatology says to always act. The forecast earns value only through the events that it decides *not* to act on, and each such stand-down saves C but risks L if the event was in fact a positive. Standing down pays when the fraction of stand-downs that were mistakes—the false-omission rate—is below the saving as a fraction of the loss, which is again r. So, value is positive exactly where $r > \phi(r)$: the model is useful to an authority whose action is cheap only if its "do not act" calls are wrong less often than the cost ratio tolerates. The lower limit 0.325 is the cost ratio at which the head's false-omission rate at that threshold rises to 0.325; below the minimum predicted score, there are no stand-downs at all, and value is zero by Equation (10).

What this means for a municipal decision-maker. Both boundaries are quantities that an authority can be shown directly on a two-by-two table for its own threshold [77]. An authority whose decision is expensive (cordon, mutual-aid hold) should ask for the model's precision at its threshold and compare it with its r; an authority whose decision is cheap (precautionary shutdown, evacuation) should ask for the false-omission rate and compare it with its r. The practical consequence for the low-r problem is that widening the envelope downward requires reducing the false-omission rate at low thresholds—a calibration objective, not a ranking objective—which is why post hoc calibration should precede encoder changes [112,113].

4.7. Candidate Decisions Within the Envelope

Table 13 lists decision types whose cost ratios plausibly fall in $[0.325, 0.850]$. They are author-proposed and offered so that a reader with real cost data can check whether their decision lies inside the envelope. Every candidate is a material-consequence decision; no life-safety decision appears, which is the honest statement of this model's current operational scope.

Table 13. Candidate decisions whose cost/loss ratios may lie within the positive-value envelope.

Candidate Decision	Plausible r	Why the Ratio Is Moderate
Retain vs. release surge inspection crews after a felt event	0.4–0.7	crew-day cost is a substantial fraction of the loss from a missed re-inspection
Extend vs. lift a precautionary cordon on a damaged structure	0.35–0.6	cordon cost is disruption; loss is bounded and largely material
Hold vs. stand down a mutual-aid request	0.4–0.8	mobilisation is expensive relative to delayed arrival
Defer vs. proceed with non-critical utility works in an affected zone	0.35–0.65	deferral cost is schedule slip; loss is rework and exposure

AUTHOR-PROPOSED: argued from decision structure, not measured; requires stakeholder validation. UNANCHORED IN THE VALUE FRAME: referenced to climatology only; the baselines of Table 11 were scored as point metrics, not as value curves (Section 4.5).

4.8. What Remains Absent, and Why

Three items requested in review are stamped ABSENT in this revision, and an audit paper owes the reason rather than the label. Each is scheduled for the released pipeline's next full invocation.

The coupling trajectory. The executed run logged the per-epoch gradient into each physics constant but not $\|\nabla_W \mathcal{L}^c\|$ or its ratio to $\|\nabla_W \mathcal{L}_a\|$. Recovering the trajectory requires re-executing the full 30-epoch, six-configuration, three-seed budget with the additional logging, which was outside the revision window. The consequence for the paper's claims is bounded by design: the gate answers question (a) as a yes/no question and does so

exactly (Appendix C); the magnitude of the coupling during training bears on *how much* the physics term shaped the encoder, on which the paper makes no claim (Table 1).

The third protocol. The Reasenberg declustering is implemented but running it means a further full training budget. Section 4.3 states why two protocols suffice for the conclusion and not for the range; the range reported is a lower bound.

Value-frame baselines. The baselines' validation predictions exist and $V(r)$ is cheap to compute from them; this is the least defensible of the three absences and the first to be closed. Until it is, Tables 12 and 13 carry the UNANCHORED stamp, and no comparative statement about POSEIDON's value over the baselines is made anywhere in this paper.

5. Policy and Decision Framework: From Target G to Procedure

This section is normative. It grounds the decision frame in published policy instruments and shows how an envelope becomes a procedure. It contains no results about this model. POLICY-NORMATIVE: derived from published policy instruments and decision theory, not from this model's outputs.

5.1. What the Instruments Require

Target G concerns *availability of and access to* risk information [8,9]; Indicator 11.b.2 counts local governments implementing local strategies [14]. Three consequences follow for model reporting. The consumer is a standing procedure, which needs a threshold rule, its expected consequence, and the conditions under which it stops applying; AUC supplies none of these [34]. Local governments differ in cost structure, so the same model is decision-relevant in one jurisdiction and irrelevant in another with identical hazard; a value curve makes this visible, a single operating point hides it. And justifiability is functional: an authority must be able to explain a decision afterwards, which is why "better than climatology" and "better than a cheap alternative" are different questions, and why Section 4.5 had to be run.

Table 14 maps each element of the framework to the Sendai priority, the Target G custom indicator, and the SDG target that it most directly serves. Question (a) is a Priority 1 matter: a model whose stated mechanism is not its actual mechanism misinforms the risk knowledge base even when its scores are good. Question (b) bears on Indicator G-5's requirement that risk information be *relevant* locally: constants that move by 0.3 under a change in catalogue cannot be transferred from one region's protocol to another's, which is a caution against importing published constants into a local strategy. Question (c) is a Priority 2 matter and connects to Indicators G-4 and G-6: a local plan to act on a warning must name the decision that it serves, and the envelope is the artefact that says whether a pre-emptive evacuation (G-6) is a decision that this model can inform—here, it is not.

Table 14. Mapping of the audit elements to the policy instruments they serve.

Audit Element	Sendai Framework	2030 Agenda and Related Instruments
(a) Gradient-verified coupling (Section 4.1)	Priority 1 (understanding disaster risk: science-based, transparent methods); Indicator G-5 (*usable* risk information)	SDG 9.5 (scientific research and technological capability); SDG 17.18 (quality, reliable data)
(b) Protocol-dependent constants (Section 4.3)	Priority 1; Indicator G-5 (*relevant* at local level: constants do not transfer across regions)	SDG 13.1 (adaptive capacity is location-specific); SDG 11.b (local strategies)

Table 14. *Cont.*

Audit Element	Sendai Framework	2030 Agenda and Related Instruments
(c) Operating envelope and decision ladder (Sections 4.6 and 5.3)	Priority 2 (risk governance: clear roles and accountability); Indicator G-4 (local plans to act on warnings); Indicator G-6 (pre-emptive evacuation—where the model has no value, the envelope says so)	SDG 11.5 and 1.5 (reduce deaths, affected people and losses); Indicator 11.b.2 (implementation by local governments)
Baseline anchor (Section 4.5)	Priority 3 (investing in DRR: cost-effectiveness of the tool itself)	SDG 11.b; SDG 17.9 (capacity building: a deployable cheap alternative)
Disclosure template (Section 5.4)	Priority 2; Words into Action guidance on local strategies [85]; Disaster Resilience Scorecard [86]	New Urban Agenda [84]; AI governance instruments [81,82]
Resource reporting (Section 3.7)	Priority 3	SDG 12 (responsible consumption); SDG 13; Green AI [90]

POLICY-NORMATIVE: derived from published policy instruments and decision theory, not from this model's outputs. Indicator numbering follows the Target G status report [9] and the global indicator framework [23].

5.2. Who Occupies Which Cost Ratio, and Who Is Left Out

$r = C/L$ is a property of the decision, supplied by the deployer. A model has no single operating point; it has a set of decisions that it can serve. Table 15 sketches who occupies which region of r [114–121]. The life-safety rows sit at low r, where Section 4.6.1 shows that this model's value is identically zero; the material-consequence rows sit in the mid range, where the envelope is positive. Stating that mismatch is the single most actionable output of this paper.

Table 15. Decision archetypes in municipal seismic and tsunami response, by approximate cost/loss ratio.

Archetype	Approx. r	Cost Structure
Life-safety evacuation, dense coastal zone	10^{-4}–10^{-2}	disruption vs. casualties; $L \gg C$
Precautionary shutdown of a critical lifeline	0.05–0.25	service interruption vs. cascading failure
Cordon extension around a damaged structure	0.30–0.60	bounded disruption vs. bounded material loss
Retention of surge inspection capacity	0.40–0.70	crew-days vs. a missed re-inspection
Mutual-aid hold vs. stand-down	0.40–0.80	mobilisation vs. delayed arrival
Discretionary public messaging	0.60–0.90	credibility and alert fatigue vs. modest loss

AUTHOR-PROPOSED: argued from decision structure, not measured; requires stakeholder validation. POLICY-NORMATIVE: derived from published policy instruments and decision theory, not from this model's outputs.

Because r depends on local costs, envelope coverage is distributed unevenly in a way that correlates with resources. Where labour and mobilisation are cheap relative to losses, or mutual-aid capacity is thin, cost ratios for material-consequence decisions fall toward the low-r region, where this model contributes nothing. A mid-range envelope is thus more useful to well-resourced authorities, the opposite of the distribution that Target G is

meant to improve [9,10,122]; Target G coverage gaps are concentrated in least-developed countries and small island states. Two things follow. Envelopes should be reported so that authorities can locate themselves, and developers should state which region of r their design choices target—here, the tsunami weighting lifts the minimum output to 0.3188 and forecloses precisely the life-safety band, a mismatch visible only in the envelope frame and correctable. And, the tabular baselines of Table 11 are directly relevant to a resource-constrained authority: a model on sixteen features that fits in seconds on existing hardware is a deployable artefact where a convolutional model on global grids is not, and, for many authorities, the material choice is not the 0.017 difference in aftershock F_1 but whether the tool can be run, maintained and re-audited on the public budget that also funds the response.

5.3. The Four-Step Translation and Its Governance

1. **Elicit** r for the specific decision, in the authority's own cost structure [123,124]. This step cannot be delegated to the developer and is the one most often skipped. A ratio elicited by the emergency manager alone encodes the institution's costs, not the community's losses; Sendai's "all-of-society" principle and the community-based DRR literature [87,88] imply that the loss side of r should be elicited with those who bear it. If losses to vulnerable populations are under-counted, r is over-estimated, and a decision that in truth lies in the low-r region is wrongly placed inside the envelope.
2. **Locate** r **against the envelope** (Figure 7). If $r < \min_i p_i$, the rule is climatology under the model's name. If r is outside the positive region, the model does not serve this decision. Because r is rarely known to be better than a factor of two, a decision counts as in scope only if the whole elicited range of r lies inside the positive region.
3. **Fix the threshold at** r and record the expected action fraction and the reference class against which value was established—including how the model stands against a cheap baseline (Table 11). This record is the minimum that an after-action inquiry will ask for; the L'Aquila case is the standing reminder that advice whose basis cannot be reconstructed afterwards exposes both the authority and its advisers [17,121]. It should be written when the procedure is adopted, not reconstructed afterwards.
4. **Set a re-audit trigger.** Any change to catalogue, completeness, declustering or task set invalidates constants and envelope (Section 4.3). The disclosure of Table 16 and the step-3 record should be annexed to the local DRR strategy and reviewed on its own cycle, as the Words into Action guidance recommends for any risk-information input [85].

Two provisions make the translation durable across elections: staff turnover and vendor changes. A procedure that depends on a model no one but the vendor can re-audit is not sustainable: the gates of Section 3.6, the emitted tables and the discrepancy list of Appendix E are released so that the audit can be re-run by a third party, and we regard adherence to the FAIR principles for these artefacts [125] as a procurement condition. And, every check is cheap by design—one backward pass, a tabular fit in seconds, a value curve from validation predictions—because where Target G gaps are largest [9,24], the audit must be executable by an authority's own technical staff on existing hardware, or it will not be executed at all.

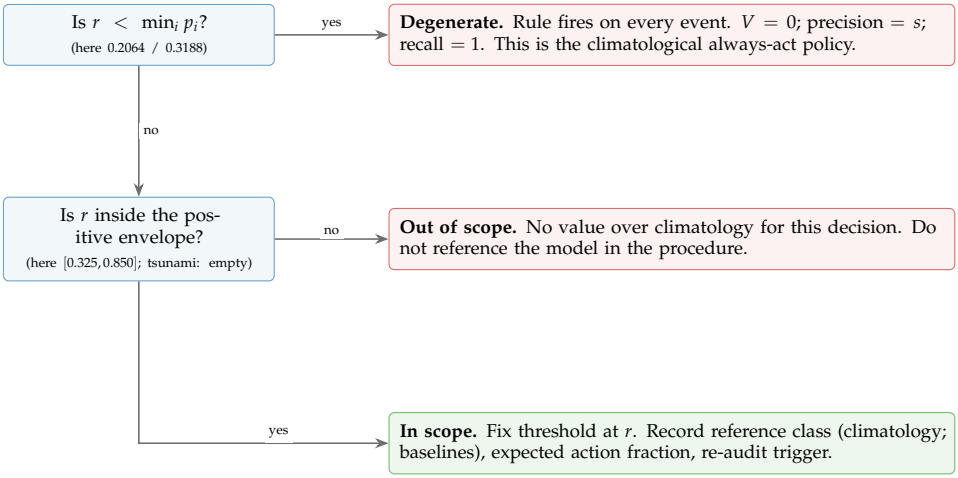


Figure 7. Decision ladder for step 2. Both tests are analytic boundary conditions (Equation (10); Appendix C.6).

5.4. A Disclosure Template for Procurement

Table 16 proposes the minimum disclosure that a purchasing authority should require [44–46,78,80–83]. Each row corresponds to a failure mode that this paper encountered in its own drafts. Every row is cheap for a developer who has the artefacts and expensive only for one who does not. The resource row deserves comment for a sustainability venue: a hazard model in operational use is retrained on every catalogue update and run continuously for inference, so its energy and hardware demands are recurring costs of the DRR strategy that fall on the same public budget as the response [89–91,93]. The standard claim that physics-informing improves sample- or compute-efficiency [65,126,127] is a sustainability argument only if measured at matched discrimination against the ablated variant; unmeasured, it cannot be audited. We treat that row as a disclosure of equal standing with the others and record that this paper does not meet it (Section 3.7).

Table 16. Proposed minimum disclosure for procurement of a physics-informed hazard model.

Required Disclosure	Failure Mode Detected
Per-term norm of the physics-loss gradient with regard to network weights, and when it was computed	Physics term is decoration (Section 4.1)
Parameterisation of every learned constant, with bounds, initialisation and distance travelled	Admissibility is an artefact of bounding (Sections 3.5 and 4.2)
Which constants are identified by which route	Constants credited to the network that a likelihood would give anyway (Section 4.2)
Constants as ranges over protocols, not means over seeds	Repeatability presented as accuracy (Section 4.3)
Exact label rule and what discrimination against it measures	Rule recoverability reported as forecasting skill (Section 3.2)
Minimum predicted score per head	Threshold degeneracy; precision reporting the base rate (Section 4.6.1)
Value as a function of r, with reference class named	Single operating point implying general applicability (Section 4.6)
Cheap baseline on the same features and partition	Under-anchoring (Section 4.5)
Split protocol with sequence leakage addressed, and partition dates	Optimism from random splitting (Section 4.4)
Training and inference cost, including for the ablated variant	Unfalsifiable efficiency claims (Section 3.7)

POLICY-NORMATIVE: derived from published policy instruments and decision theory, not from this model's outputs.

5.5. Limits of the Policy Translation

The envelope is a screening tool, not a mandate: it tells an authority which of its decisions a model *could* inform, not that it should adopt the model for any of them; adoption remains a governance decision that weighs the baseline alternative, the recurring cost, and the authority's capacity to re-audit. None of the heads forecast a future hazard in the sense an operational earthquake forecasting protocol does [17,18], and nothing here places the model in that role. The equity argument is structural, not measured: we have not elicited cost ratios in any jurisdiction, and the candidate decisions of Table 13 await stakeholder validation. And, the framework is silent on the determinants of disaster outcomes—land use, building codes, social protection, warning dissemination—that Sendai Priorities 3 and 4 place ahead of any forecasting tool [8,122]. The audit makes a model's claims legible to an authority; it does not make the model necessary. POLICY-NORMATIVE: derived from published policy instruments and decision theory, not from this model's outputs. AUTHOR-PROPOSED: argued from decision structure, not measured; requires stakeholder validation.

6. Discussion

6.1. What the Bounded Parameterisation Cost

An earlier draft argued that positivity-only parameterisation had exposed a convention error in α. The executed configuration bounds α into $(0.5, 1.5)$, and the "error" was in the manuscript's equation, not the implementation (Appendix E, item 2). The corrected lesson is the same rule from the other side: a constraint that cannot be violated cannot be informative, and it also cannot warn you. With bounded constants, the only informative quantities are those the pipeline now emits—initialisation, distance travelled, and the same-sample MLE—and a reader who is shown only "$b = 1.069$, admissible" has been shown nothing. We did not follow our own recommendation in the configuration that we ran; the recommendation stands.

6.2. Why the Anchor Is Not Optional

A framework for auditing hazard models that, applied to its authors' model, returned only clean results would not have been tested. It returned a non-zero coupling that does not resolve into discrimination at budget, and constants that a decoupled architecture reproduces to within 0.02; both were found by the instrument that the paper proposes. The DeVries–Mignan correction [54] came from outside; here, the baseline comparison came from the same pipeline run, and the draft's ABSENT stamp on it was itself an audit failure—the evidence had been produced and not read. Whatever Table 11 shows, it is the reference class against which any comparative statement about POSEIDON must be made, and the margins that it shows over converged tabular models are small enough that no authority should choose the network over the cheap alternative on discrimination alone.

6.3. Why Protocol Dependence Is the Central Methodological Finding

It applies to every study that reports learned seismological constants as interpretability evidence, including our feasibility study. The test that the field uses lacks the resolution to detect a 0.3 shift in b, and the statistic conventionally reported alongside—seed s.d.—understates the relevant uncertainty by two orders of magnitude. The shift is not a surprise to statistical seismology, which has long known that b and the Omori parameters depend on catalogue, completeness and declustering [27,48,68]; the finding is that a learned-constant pipeline inherits this dependence in full while its reporting convention conceals it. The constructive response is cheap: report ranges over protocols.

6.4. Recommendations for the Field

(i) Differentiate each physics term against the network weights, report the norms, say when they were measured, and log the norm and its ratio to the task gradient through training. (ii) State the parameterisation of every constant; prefer positivity-only; if bounded, report initialisation, distance travelled and the same-sample MLE. (iii) Report constants as ranges over protocols. (iv) Report value as a function of r, name the reference class, and check the decision band against the minimum predicted score before interpreting precision or recall. (v) Run a cheap baseline on the same features and partition, in the same frame, before drafting the abstract. (vi) Reconcile the manuscript against the released implementation and publish the discrepancy list.

7. Conclusions

Three findings. First, gradient verification is necessary and insufficient. The coupled physics term reaches the encoder—50.94 at initialisation against an exact zero for the parallel likelihood terms—but at the declared budget, this changes the learned constants and not the heads' discrimination, and the decoupled architecture reproduces the constants to within 0.02; only a differentiation against the network weights can tell the two apart, and no gradient norm can say that the coupling helped. Second, protocol dependence is general. The constants move by 131–317 seed standard deviations between two protocols while both pass the published-range test, which, for interval-bounded constants, they could not have failed; the dependence is the one that statistical seismology already documents, inherited by a learned model and hidden by its reporting convention, and the remedy is to report ranges over protocols. Third, the operating envelope is the decision-relevant statement. The aftershock head has positive value on $r \in [0.325, 0.850]$, peaking at the Peirce skill score 0.441, with boundaries that are the model's own precision and false-omission rate; both executed life-safety bands lie in the degenerate region where the rule is climatology under the model's name, and the tsunami head has no value at any r. For Target G and Indicator 11.b.2, that last result is the framework doing what a sustainability instrument must do: telling a local authority where a tool stops being useful before the event that would otherwise reveal it.

Author Contributions: B.K.: conceptualisation, methodology, implementation, analysis, writing—original draft. F.K.: audit design, verification of claim classes, policy framing, writing—review and editing. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Data Availability Statement: The catalogue is public [94]. The pipeline, the gates of Section 3.6, the partition manifest (split timestamp and post-embargo count), the per-year completeness and b-value diagnostics, the emitted tables and the artefacts that they are built from are released with the paper. The pipeline additionally emits a multi-hazard risk index over exposure and vulnerability layers, which is not evaluated here.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Notation

Table A1. Notation.

Symbol	Meaning
$\mathbf{x} \in \mathbb{R}^{16}$	event feature vector, causal
$\mathbf{G}^{(\tau)}$	$6 \times 90 \times 180$ grid, lookback $\tau \in \{7, 30, 90\}$ d
$\mathbf{z}, \tilde{\mathbf{z}}$	latent ($\mathbb{R}^{64}$) and latent with appended $\tanh$ E_θ ($\mathbb{R}^{65}$)

Table A1. *Cont.*

Symbol	Meaning
p_a, p_f, p_t	aftershock, foreshock, tsunami head outputs
$\hat{n}, \hat{m}$	normalised predicted child count and largest-aftershock magnitude; $\hat{N} = e^{6\hat{n}} - 1$, $\hat{M} = 10\hat{m}$
s_a, s_t	validation base rates, 0.4790 and 0.0408
W	encoder and head weights
θ	physics pre-parameters $(\theta_b, \theta_p, \theta_c, \theta_\Delta, \theta_\alpha, \theta_K)$
b; p, c; α, K_0; ΔM	GR slope; Omori–Utsu exponent and onset; productivity exponent (base 10) and scale; Båth offset
$\mathcal{L}^c, \mathcal{L}^d$	coupled and params-only physics objectives
$A(p, c; S, T)$	$\int_S^T (t + c)^{-p} \mathrm{d}t$, $S = 0.01$ d
M_c, M_0	child threshold 3.0, trigger threshold 5.0
$r = C/L$; $V(r)$; V_{pot}	cost/loss ratio; Murphy value; threshold-optimal value
H, F; $\pi(r)$; $\phi(r)$	hit and false-alarm rates; precision; false-omission rate at threshold r
PSS	Peirce skill score $H - F$

Appendix B. Executed Configuration

All model, baseline and value numbers come from a single invocation of the released pipeline at its mid budget, with three seeds and the six configurations of Table 10; settings not listed in Table A2 are the pipeline defaults. The pipeline additionally aggregates hazard into a risk index over a 365-day window; that layer is not evaluated in this paper, and an earlier draft misreported its window as the catalogue window (Appendix E, item 10).

Table A2. Executed configuration.

Item	Value
Catalogue	ComCat, 1 January 2010 to 1 January 2025, $M \geq 4.5$, 112,562 events (Table 2)
Triggers	$M \geq 5.0$, 27,100; base rates 0.494/0.225/0.0352
Split	chronological 80/20 at the origin time of trigger 21,681, with 90-day embargo; validation 5420; random split for `full_random` only
Inputs	$\mathbf{x} \in \mathbb{R}^{16}$; three $6 \times 90 \times 180$ grids at 7/30/90 d, 2° cells
Heads	aftershock (p_a, $\hat{n}$, $\hat{m}$), foreshock, tsunami
Losses	BCE (smoothing 0.05)/BCE pos. weight 3.0/BCE pos. weight $\min((1 - \hat{s})/\hat{s}, 30) \approx 28$ on the training rate
Physics	Equation (8); $\lambda_p = 0$ (epochs 0–9), 0.1 (epochs 10–29); $\lambda_{\mathrm{B}} \in \{0, 1, 20\}$; params-only weight 1
Gates	Båth ≥ 4 rows; OU data ≥ 8; GR data ≥ 8
Budget	$10 + 20$ epochs at batch 64 (pipeline mid budget); constants read at epoch 29
Optimiser	AdamW, weight decay 10^{-4}, gradient clip 1.0; network lr 10^{-4} (OneCycle) then 10^{-5} (cosine); physics-constant group lr 5×10^{-3} then 10^{-3}
Sampler	weights $1 + 10 \cdot \mathbb{1}[\text{tsunami}] + 3 \cdot \mathbb{1}[\text{foreshock}]$, dedicated generator so batch order is identical across configurations at a seed
Physics init (neutral)	$b = 1.215$, $p = 0.867$, $c = 0.1061$, $\Delta M = 1.731$, $\alpha = 0.731$
Seeds	20240501, 20240502, 20240503
Gradient gate	24-event stratified probe (9 with largest aftershock, 17 with parent delay), model at initialisation, CPU, λ_p not applied
Baselines	HGB, RF, LR; randomised search, 12 candidates, 4-fold inner CV, average-precision scoring; time-forward folds with stratified fallback
Decision bands (executed)	coastal evacuation $[2 \times 10^{-4}, 5 \times 10^{-3}]$; building closure $[5 \times 10^{-3}, 8 \times 10^{-2}]$; symmetric reference $[0.4, 0.6]$
Minimum predicted score	0.2064 (p_a), 0.3188 (p_t)
Claim gate	seed, head-collapse and gate-activity criteria checked; every emitted table carries its budget stamp

Appendix C. Derivations

Appendix C.1. Omori–Utsu Integral and the $p \to 1$ Limit

With $\lambda(t) = K(t+c)^{-p}$ [69],

$$A(p,c;S,T) = \int_S^T (t+c)^{-p}\mathrm{d}t = \frac{(S+c)^{1-p} - (T+c)^{1-p}}{p-1}, \quad p \neq 1; \qquad A(1,c;S,T) = \ln\frac{T+c}{S+c}. \tag{A1}$$

The implementation branches on $|p-1| < 10^{-4}$ and masks the unused branch's denominator, because both branches are evaluated before selection.

Appendix C.2. Why the Params-Only Terms Have Exactly Zero Network Gradient

Let $\mathcal{L}^d = f(T(\mathcal{D}), \theta)$ with $T(\mathcal{D})$ any statistic of labels and magnitudes. Then,

$$\nabla_W \mathcal{L}^d = \frac{\partial f}{\partial T}\underbrace{\nabla_W T(\mathcal{D})}_{=0} + \frac{\partial f}{\partial \theta}\underbrace{\nabla_W \theta}_{=0} = 0 \tag{A2}$$

identically, while $\nabla_\theta \mathcal{L}^d \neq 0$ in general. The coupled terms depend on $\hat{n}, \hat{m}$, so ∇_W passes through the encoder Jacobian and is generically non-zero. Table 5 measures both in one backward pass: 50.9388 and 0.0000. Because the zero holds at every point in weight space, a single non-zero measurement of the coupled route at any point suffices to distinguish the two architectures.

Appendix C.3. Murphy Value in Terms of Rates

For base rate s, hit rate H and false-alarm rate F at threshold r [32,33]: $\mathbb{E}[\text{cost}]_{\text{fcst}} = r[sH + (1-s)F] + s(1-H)$, $\mathbb{E}[\text{cost}]_{\text{clim}} = \min(r, s)$, $\mathbb{E}[\text{cost}]_{\text{perfect}} = rs$, so

$$V(r) = \frac{\min(r,s) - r[sH + (1-s)F] - s(1-H)}{\min(r,s) - rs}. \tag{A3}$$

Appendix C.4. The Peirce Identity at $r = s$

At $r = s$, the denominator is $s(1-s)$ and the numerator is $s - s[sH + (1-s)F] - s(1-H) = s(1-s)(H-F)$, so $V(s) = H - F = \text{PSS}$ [75]; its maximum over thresholds is the Youden index [76].

Appendix C.5. The Degenerate-Threshold Result

If $r < \min_i p_i$, then every event triggers action: $H = F = 1$, precision $= s$. The numerator of Equation (A3) becomes $\min(r,s) - r$, which is zero for $r \leq s$. Both executed bands satisfy $r < \min_i p_i \leq s$ for their heads ($0.08 < 0.2064$; $0.005 < 0.3188$), establishing Equation (10).

Appendix C.6. Both Envelope Boundaries in Closed Form

Write $\text{TP} = sH$, $\text{FP} = (1-s)F$, $\text{FN} = s(1-H)$, $\text{TN} = (1-s)(1-F)$; $\pi = \text{TP}/(\text{TP} + \text{FP})$, $\phi = \text{FN}/(\text{FN} + \text{TN})$. For $r > s$, the denominator is $s(1-r) > 0$ and the numerator is $\text{TP} - r(\text{TP} + \text{FP})$, positive iff $r < \pi(r)$. For $r < s$, the denominator is $r(1-s) > 0$ and the numerator is $r(\text{FN} + \text{TN}) - \text{FN}$, positive iff $r > \phi(r)$. Hence,

$$V(r) > 0 \iff \phi(r) < r < \pi(r) \tag{A4}$$

wherever the rule is non-degenerate. Since π rises and ϕ falls with the threshold for any better-than-chance ranking, the envelope is one-peaked near the base rate, and the low-r boundary is governed by the false-omission rate—a calibration quantity.

Appendix D. Claim Classes

Table A3. Claim classes.

Class	Meaning	Example
ANALYTIC	differentiation or arithmetic on recorded quantities	Table 5; Equation (10)
AT BUDGET	measured at the declared budget, 3 seeds; pipeline-stamped	Tables 6, 9 and 10
ANCHORED AT BUDGET	compared to tuned baselines on the same partition	Table 11
UNANCHORED	referenced to climatology only, in the value frame	Table 12
AUTHOR-PROPOSED	argued from decision structure	Tables 15 and 13
POLICY-NORMATIVE	from policy instruments and decision theory	Section 5
ABSENT	not produced; a gap with a stated reason, not a scope decision	coupling trajectory; third protocol; compute; value-frame baselines (Section 4.8)

Appendix E. Manuscript–Code Reconciliation Record

Table A4 records the discrepancies between earlier drafts of this manuscript and the executed configuration that were found when the manuscript was reconciled against the released implementation. Each row states what the draft said, what the implementation does, and where in the text the correction is made. Items marked load-bearing changed a conclusion, not only a number.

Table A4. Discrepancies between drafts and the executed configuration.

#	Draft Stated	Executed Configuration	Corrected in	Load-Bearing
1	All physics constants positivity-only parameterised	b, p, α, ΔM sigmoid-bounded into published intervals; only c, K_0 positivity-only	Section 3.5	yes
2	Positivity-only parameterisation had exposed a convention error making α inadmissible	α is bounded into $(0.5, 1.5)$ and cannot be inadmissible; the error was in the draft's equation	Sections 4.2 and 6.1	yes
3	Omori–Utsu productivity written base e	Base 10, Equation (4)	Section 3.4	yes
4	Tabular baseline comparison "withheld"	Pipeline had run the baselines on the same partition; the draft had not read the output	Sections 3.6 and 4.5	yes
5	Aftershock label: "≥ 1 event with $M' \geq M - 1.0$ within 100 km and 30 d after the trigger"	Trigger lies in the Gardner–Knopoff window of a preceding larger event	Section 3.2	yes
6	Route attribution: $b, p, c, \Delta M$ credited to the coupled route; α, K_0 to the likelihood	$b, p, c, \Delta M$ appear in both routes (joint); α, K_0 appear only in the coupled term	Section 4.2	yes
7	Ablation table carried an AUC column	Column could not be matched to run logs; removed, pipeline emission reproduced	Table 10	no
8	Baseline tuning protocol described loosely	Randomised search, 12 candidates, 4-fold inner CV, average-precision scoring, time-forward folds with stratified fallback	Section 4.5, Table A2	no

Table A4. *Cont.*

#	Draft Stated	Executed Configuration	Corrected in	Load-Bearing
9	Decision bands $[0.05, 0.20]$ and $[0.01, 0.10]$	Executed bands $[2 \times 10^{-4}, 5 \times 10^{-3}]$ and $[5 \times 10^{-3}, 8 \times 10^{-2}]$; both regimes are degenerate, so the conclusion is unchanged	Section 4.6.1	no
10	Risk-index aggregation window equal to the catalogue window	Risk-index window is 365 d; the layer is not evaluated here	Appendix B	no
11	Six physics ablations reported as six runs (earlier implementation)	The configurations were one run; the ablation-effect check now fails any flag that does nothing	Section 3.6	yes
12	Gradient-flow probe batch was stratified (earlier implementation)	Probe was assembled by sorting and truncating indices, yielding the sixteen earliest events with no parent delay; repaired to the stratified 24-event probe	Section 3.6, Table 5	yes
13	Gradient norm reported without its measurement conditions	Measured once, at initialisation, on the probe batch, λ_p not applied; trajectory not logged	Section 4.1, Table A2	no

Appendix F. Why the Feasibility Study's Baseline Comparison Does Not Transfer

The feasibility study [47] reports gradient-boosting, random-forest and CNN baselines and leads them on the aftershock task by $+0.030$ AUC. That comparison is on a different sample (the 2.8-million-event Poseidon dataset across the full magnitude range), under a different sequence-labelling convention (which alone moves b by 0.3 [48]), with different loss weighting, and its 0.779 figure is a cross-task aggregate. Multi-task interference is not constant across task sets [128–130]. We therefore ran baselines on the present sample (Table 11); whether any difference from the earlier comparison reflects the budget, the protocol, or both is left to future work.

References

1. McGranahan, G.; Balk, D.; Anderson, B. The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones. *Environ. Urban.* **2007**, *19*, 17–37.
2. Seto, K.C.; Güneralp, B.; Hutyra, L.R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16083–16088.
3. United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2018 Revision*; ST/ESA/SER.A/420; United Nations: New York, NY, USA, 2019.
4. Pelling, M. *Adaptation to Climate Change: From Resilience to Transformation*; Routledge: London,UK, 2011.
5. Silva, V.; Amo-Oduro, D.; Calderon, A.; Costa, C.; Dabbeek, J.; Despotaki, V.; Martins, L.; Pagani, M.; Rao, A.; Simionato, M.; et al. Development of a global seismic risk model. *Earthq. Spectra* **2020**, *36*, 372–394.
6. Pagani, M.; Monelli, D.; Weatherill, G.; Danciu, L.; Crowley, H.; Silva, V.; Henshaw, P.; Butler, L.; Nastasi, M.; Panzeri, L.; et al. OpenQuake engine: An open hazard (and risk) software for the Global Earthquake Model. *Seismol. Res. Lett.* **2014**, *85*, 692–702.
7. Gerstenberger, M.C.; Marzocchi, W.; Allen, T.; Pagani, M.; Adams, J.; Danciu, L.; Field, E.H.; Fujiwara, H.; Luco, N.; Ma, K.; et al. Probabilistic seismic hazard analysis at regional and national scales: State of the art and future challenges. *Rev. Geophys.* **2020**, *58*, e2019RG000653.
8. United Nations Office for Disaster Risk Reduction. *Sendai Framework for Disaster Risk Reduction 2015–2030*; UNDRR: Geneva, Switzerland, 2015.
9. United Nations Office for Disaster Risk Reduction and World Meteorological Organization. *Global Status of Multi-Hazard Early Warning Systems: Target G*; UNDRR/WMO: Geneva, Switzerland, 2022.
10. United Nations Office for Disaster Risk Reduction. *Global Assessment Report on Disaster Risk Reduction 2019*; UNDRR: Geneva, Switzerland, 2019.

11. Aitsi-Selmi, A.; Egawa, S.; Sasaki, H.; Wannous, C.; Murray, V. The Sendai framework for disaster risk reduction: Renewing the global commitment to people's resilience, health, and well-being. *Int. J. Disaster Risk Sci.* **2015**, *6*, 164–176.
12. Pearson, L.; Pelling, M. The UN Sendai Framework for Disaster Risk Reduction 2015–2030: Negotiation process and prospects for science and practice. *J. Extrem. Events* **2015**, *2*, 1571001.
13. United Nations. *Transforming Our World: The 2030 Agenda for Sustainable Development*; A/RES/70/1; United Nations: New York, NY, USA , 2015.
14. United Nations Statistics Division. SDG Indicator Metadata: Indicator 11.b.2. Available online: https://unstats.un.org/sdgs/metadata/ (accessed on 1 May 2026).
15. Titov, V.V.; Gonzalez, F.I.; Bernard, E.N.; Eble, M.C.; Mofjeld, H.O.; Newman, J.C.; Venturato, A.J. Real-time tsunami forecasting: Challenges and solutions. *Nat. Hazards* **2005**, *35*, 41–58.
16. Bernard, E.; Titov, V. Evolution of tsunami warning systems and products. *Philos. Trans. R. Soc. A* **2015**, *373*, 20140371.
17. Jordan, T.H.; Chen, Y.-T.; Gasparini, P.; Madariaga, R.; Main, I.; Marzocchi, W.; Papadopoulos, G.; Sobolev, G.; Yamaoka, K.; Zschau, J. Operational earthquake forecasting: State of knowledge and guidelines for utilization. *Ann. Geophys.* **2011**, *54*, 315–391.
18. Marzocchi, W.; Lombardi, A.M.; Casarotti, E. The establishment of an operational earthquake forecasting system in Italy. *Seismol. Res. Lett.* **2014**, *85*, 961–969.
19. Kong, Q.; Trugman, D.T.; Ross, Z.E.; Bianco, M.J.; Meade, B.J.; Gerstoft, P. Machine learning in seismology: Turning data into insights. *Seismol. Res. Lett.* **2019**, *90*, 3–14.
20. Bergen, K.J.; Johnson, P.A.; de Hoop, M.V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363*, eaau0323.
21. Beroza, G.C.; Segou, M.; Mousavi, S.M. Machine learning and earthquake forecasting—Next steps. *Nat. Commun.* **2021**, *12*, 4761.
22. Mousavi, S.M.; Beroza, G.C. Deep-learning seismology. *Science* **2022**, *377*, eabm4470.
23. United Nations General Assembly. *Work of the Statistical Commission Pertaining to the 2030 Agenda for Sustainable Development (Global Indicator Framework)*; A/RES/71/313; United Nations General Assembly: New York, NY, USA, 2017.
24. United Nations Office for Disaster Risk Reduction. *Report of the Main Findings and Recommendations of the Midterm Review of the Implementation of the Sendai Framework for Disaster Risk Reduction 2015–2030*; UNDRR: Geneva, Switzerland, 2023.
25. World Meteorological Organization and United Nations Office for Disaster Risk Reduction. *Early Warnings for All: Executive Action Plan 2023–2027*; WMO: Geneva, Switzerland, 2022.
26. Aki, K. Maximum likelihood estimate of b in the formula $\log N = a - bM$ and its confidence limits. *Bull. Earthq. Res. Inst.* **1965**, *43*, 237–239.
27. Marzocchi, W.; Sandri, L. A review and new insights on the estimation of the b-value and its uncertainty. *Ann. Geophys.* **2003**, *46*, 1271–1282.
28. Console, R.; Lombardi, A.M.; Murru, M.; Rhoades, D. Båth's law and the self-similarity of earthquakes. *J. Geophys. Res. Solid Earth* **2003**, *108*, 2128.
29. Helmstetter, A.; Sornette, D. Båth's law derived from the Gutenberg-Richter law and from aftershock properties. *Geophys. Res. Lett.* **2003**, *30*, 2069.
30. Utsu, T. A statistical study on the occurrence of aftershocks. *Geophys. Mag.* **1961**, *30*, 521–605.
31. Murphy, A.H. The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Mon. Weather Rev.* **1977**, *105*, 803–816.
32. Richardson, D.S. Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **2000**, *126*, 649–667.
33. Wilks, D.S. A skill score based on economic value for probability forecasts. *Meteorol. Appl.* **2001**, *8*, 209–219.
34. Hand, D.J. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Mach. Learn.* **2009**, *77*, 103–123.
35. Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* ; Association for Computing Machinery: New York, NY, USA, 2006; pp. 233–240.
36. Mylne, K.R. Decision-making from probability forecasts based on forecast value. *Meteorol. Appl.* **2002**, *9*, 307–315.
37. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
38. Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277.
39. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259.

40. King, G.; Zeng, L. Logistic regression in rare events data. *Political Anal.* **2001**, *9*, 137–163.
41. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432.
42. Wang, S.; Teng, Y.; Perdikaris, P. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM J. Sci. Comput.* **2021**, *43*, A3055–A3081.
43. Krishnapriyan, A.S.; Gholami, A.; Zhe, S.; Kirby, R.M.; Mahoney, M.W. Characterizing possible failure modes in physics-informed neural networks. *NeurIPS* **2021**, *34*, 26548–26560.
44. Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T. Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), Atlanta, GA, USA, 29–31 January 2019; pp. 220–229.
45. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Iii, H.D.; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92.
46. Arnold, M.; Bellamy, R.K.E.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K.N.; Olteanu, A.; Piorkowski, D.; et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM J. Res. Dev.* **2019**, *63*, 6:1–6:13.
47. Kriuk, B.; Kriuk, F. POSEIDON: Physics-Optimized Seismic Energy Inference and Detection Operating Network. In *AI & PDE: ICLR 2026 Workshop on AI and Partial Differential Equations*, 2026. Available online: https://openreview.net/forum?id=QgDMIwafeT (accessed on 1 May 2026).
48. Mizrahi, L.; Nandan, S.; Wiemer, S. The effect of declustering on the size distribution of mainshocks. *Seismol. Res. Lett.* **2021**, *92*, 2333–2342.
49. Mousavi, S.M.; Ellsworth, W.L.; Zhu, W.; Chuang, L.Y.; Beroza, G.C. Earthquake transformer. *Nat. Commun.* **2020**, *11*, 3952.
50. Zhu, W.; Beroza, G.C. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophys. J. Int.* **2019**, *216*, 261–273.
51. Perol, T.; Gharbi, M.; Denolle, M. Convolutional neural network for earthquake detection and location. *Sci. Adv.* **2018**, *4*, e1700578.
52. Ross, Z.E.; Yue, Y.; Meier, M.-A.; Hauksson, E.; Heaton, T.H. PhaseLink: A deep learning approach to seismic phase association. *J. Geophys. Res. Solid Earth* **2019**, *124*, 856–869.
53. DeVries, P.M.R.; Viégas, F.; Wattenberg, M.; Meade, B.J. Deep learning of aftershock patterns following large earthquakes. *Nature* **2018**, *560*, 632–634.
54. Mignan, A.; Broccardo, M. One neuron versus deep learning in aftershock prediction. *Nature* **2019**, *574*, E1–E3.
55. Lipton, Z.C.; Steinhardt, J. Troubling trends in machine learning scholarship. *ACM Queue* **2019**, *17*, 45–77.
56. Sculley, D.; Snoek, J.; Wiltschko, A.; Rahimi, A. Winner's curse? On pace, progress, and empirical rigor. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
57. Jordan, T.H.; Marzocchi, W.; Michael, A.J.; Gerstenberger, M.C. Operational earthquake forecasting can enhance earthquake preparedness. *Seismol. Res. Lett.* **2014**, *85*, 955–959.
58. Schorlemmer, D.; Gerstenberger, M.C.; Wiemer, S.; Jackson, D.D.; Rhoades, D.A. Earthquake likelihood model testing. *Seismol. Res. Lett.* **2007**, *78*, 17–29.
59. Zechar, J.D.; Gerstenberger, M.C.; Rhoades, D.A. Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts. *Bull. Seismol. Soc. Am.* **2010**, *100*, 1184–1195.
60. Schorlemmer, D.; Werner, M.J.; Marzocchi, W.; Jordan, T.H.; Ogata, Y.; Jackson, D.D.; Mak, S.; Rhoades, D.A.; Gerstenberger, M.C.; Hirata, N.; et al. The Collaboratory for the Study of Earthquake Predictability: Achievements and priorities. *Seismol. Res. Lett.* **2018**, *89*, 1305–1313.
61. Field, E.H.; Milner, K.R.; Hardebeck, J.L.; Page, M.T.; van der Elst, N.; Jordan, T.H.; Michael, A.J.; Shaw, B.E.; Werner, M.J. A spatiotemporal clustering model for the third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS). *Bull. Seismol. Soc. Am.* **2017**, *107*, 1049–1081.
62. Jolliffe, I.T.; Stephenson, D.B. (Eds.) *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed.; Wiley-Blackwell: Chichester, UK, 2012.
63. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
64. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks. *J. Comput. Phys.* **2019**, *378*, 686–707.
65. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440.
66. Lu, L.; Meng, X.; Mao, Z.; Karniadakis, G.E. DeepXDE: A deep learning library for solving differential equations. *SIAM Rev.* **2021**, *63*, 208–228.
67. Gutenberg, B.; Richter, C.F. Frequency of earthquakes in California. *Bull. Seismol. Soc. Am.* **1944**, *34*, 185–188.
68. Utsu, T.; Ogata, Y.; Matsu'ura, R.S. The centenary of the Omori formula for a decay law of aftershock activity. *J. Phys. Earth* **1995**, *43*, 1–33.

69. Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.* **1988**, *83*, 9–27.
70. Ogata, Y. Space-time point-process models for earthquake occurrences. *Ann. Inst. Stat. Math.* **1998**, *50*, 379–402.
71. Båth, M. Lateral inhomogeneities of the upper mantle. *Tectonophysics* **1965**, *2*, 483–514.
72. Smith, J.D.; Azizzadenesheli, K.; Ross, Z.E. EikoNet: Solving the Eikonal equation with deep neural networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10685–10696.
73. Bin Waheed, U.; Haghighat, E.; Alkhalifah, T.; Song, C.; Hao, Q. PINNeik: Eikonal solution using physics-informed neural networks. *Comput. Geosci.* **2021**, *155*, 104833.
74. Zhu, Y.; Toth, Z.; Wobus, R.; Richardson, D.; Mylne, K. The economic value of ensemble-based weather forecasts. *Bull. Am. Meteorol. Soc.* **2002**, *83*, 73–83.
75. Peirce, C.S. The numerical measure of the success of predictions. *Science* **1884**, *4*, 453–454 .
76. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35.
77. Vickers, A.J.; Elkin, E.B. Decision curve analysis: A novel method for evaluating prediction models. *Med. Decis. Mak.* **2006**, *26*, 565–574.
78. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Barcelona, Spain, 27–30 January 2020; pp. 33–44.
79. Hutchinson, B.; Smart, A.; Hanna, A.; Denton, R.; Greer, C.; Kjartansson, O.; Barnes, P.; Mitchell, M.; Denton, E. Towards accountability for machine learning datasets. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Virtual, 3–10 March 2021; pp. 560–575.
80. Madaio, M.A.; Stark, L.; Vaughan, J.W.; Wallach, H. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In Proceedings of the 2020 ACM CHI Conference on Human Factors in Computing Systems (CHI '20), Honolulu, HI, USA, 25–30 April 2020; pp. 1–14.
81. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; NIST AI 100-1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
82. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, *1689*, 2024.
83. UK Department for Science; Innovation and Technology. *Introduction to AI Assurance*; HM Government: London, UK, 2024.
84. United Nations. *New Urban Agenda*; A/RES/71/256; United Nations: New York, NY, USA, 2017.
85. United Nations Office for Disaster Risk Reduction. *Words Into Action Guidelines: Implementation Guide for Local Disaster Risk Reduction and Resilience Strategies*; UNDRR: Geneva, Switzerland, 2019.
86. United Nations Office for Disaster Risk Reduction. *Disaster Resilience Scorecard for Cities*; UNDRR: Geneva, Switzerland, 2017.
87. Maskrey, A. Revisiting community-based disaster risk management. *Environ. Hazards* **2011**, *10*, 42–52.
88. Cutter, S.L.; Barnes, L.; Berry, M.; Burton, C.; Evans, E.; Tate, E.; Webb, J. A place-based model for understanding community resilience to natural disasters. *Glob. Environ. Change* **2008**, *18*, 598–606.
89. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3645–3650.
90. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63.
91. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.* **2020**, *21*, 1–43.
92. Lacoste, A.; Luccioni, A.; Schmidt, V.; Dandres, T. Quantifying the carbon emissions of machine learning. *arXiv* **2019**, arXiv:1910.09700.
93. Patterson, D.; Gonzalez, J.; Le Q.; Liang, C.; Munguia, L.M.; Rothchild, D.; So, D.; Texier, M.; Dean, J. Carbon emissions and large neural network training. *arXiv* **2021**, arXiv:2104.10350.
94. U.S. Geological Survey. ANSS Comprehensive Earthquake Catalog (ComCat) Documentation. Available online: https://earthquake.usgs.gov/data/comcat/ (accessed on 1 May 2026).
95. Gardner, J.K.; Knopoff, L. Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bull. Seismol. Soc. Am.* **1974**, *64*, 1363–1367.
96. van Stiphout, T.; Zhuang, J.; Marsan, D. Seismicity declustering. *CORSSA* **2012**. https://www.corssa.org/export/sites/corssa/.galleries/articles-pdf/vanStiphout_et_al.pdf_2063069299.pdf.
97. Wiemer, S.; Wyss, M. Minimum magnitude of completeness in earthquake catalogs. *Bull. Seismol. Soc. Am.* **2000**, *90*, 859–869.
98. Woessner, J.; Wiemer, S. Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bull. Seismol. Soc. Am.* **2005**, *95*, 684–698.
99. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804.

100. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19. *Nat. Mach. Intell.* **2021**, *3*, 199–217.
101. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
102. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
103. LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; Huang, F.J. A tutorial on energy-based learning. In *Predicting Structured Data*; MIT Press: Cambridge, MA, USA, 2006.
104. Du, Y.; Mordatch, I. Implicit generation and modeling with energy-based models. In *NeurIPS*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2019; Volume 32.
105. Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
106. Reasenberg, P. Second-order moment of central California seismicity, 1969–1982. *J. Geophys. Res. Solid Earth* **1985**, *90*, 5479–5495.
107. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
108. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
109. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *NIPS* **2017**, *30*, 3146–3154.
110. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
111. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, 7–11 August 2005; pp. 625–632.
112. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.
113. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), Edmonton, AB, Canada, 23–26 July 2002; pp. 694–699.
114. Cremen, G.; Galasso, C. Earthquake early warning: Recent advances and perspectives. *Earth-Sci. Rev.* **2020**, *205*, 103184.
115. Allen, R.M.; Melgar, D. Earthquake early warning: Advances, scientific challenges, and societal needs. *Annu. Rev. Earth Planet. Sci.* **2019**, *47*, 361–388.
116. Minson, S.E.; Baltay, A.S.; Cochran, E.S.; Hanks, T.C.; Page, M.T.; McBride, S.K.; Milner, K.R.; Meier, M.-A. The limits of earthquake early warning accuracy and best alerting strategy. *Sci. Rep.* **2019**, *9*, 2478.
117. Wald, D.J. Practical limitations of earthquake early warning. *Earthq. Spectra* **2020**, *36*, 1412–1447.
118. Wachinger, G.; Renn, O.; Begg, C.; Kuhlicke, C. The risk perception paradox. *Risk Anal.* **2013**, *33*, 1049–1065.
119. LeClerc, J.; Joslyn, S. The cry wolf effect and weather-related decision making. *Risk Anal.* **2015**, *35*, 385–395.
120. Ripberger, J.T.; Silva, C.L.; Jenkins-Smith, H.C.; Carlson, D.E.; James, M.; Herron, K.G. False alarms and missed events: The impact and origins of perceived inaccuracy in tornado warning systems. *Risk Anal.* **2015**, *35*, 44–56.
121. van Stiphout, T.; Wiemer, S.; Marzocchi, W. Are short-term evacuations warranted? Case of the 2009 L'Aquila earthquake. *Geophys. Res. Lett.* **2010**, *37*, L06306.
122. Hallegatte, S.; Vogt-Schilb, A.; Bangalore, M.; Rozenberg, J. *Unbreakable: Building the Resilience of the Poor in the Face of Natural Disasters*; World Bank: Washington, DC, USA, 2017.
123. Mechler, R. Reviewing estimates of the economic efficiency of disaster risk management. *Nat. Hazards* **2016**, *81*, 2121–2147.
124. Shreve, C.M.; Kelman, I. Does mitigation save? Reviewing cost-benefit analyses of disaster risk reduction. *Int. J. Disaster Risk Reduct.* **2014**, *10*, 213–235.
125. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; Santos, L.B.d.S.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018.
126. Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **2022**, *55*, 66.
127. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331.
128. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75.

129. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
130. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.