https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

PISAS v1.0: A Physics-Informed Neural Surrogate for Earthquake Productivity and Aftershock-Sequence Modeling from Open Seismic Catalogs

Antonino D'Alessandro
Istituto Nazionale di Geofisica e Vulcanologia, Osservatorio Nazionale Terremoti, Rome, Italy
antonino.dalessandro@ingv.it
https://orcid.org/0000-0002-0074-3125

Abstract

Rapid estimates of aftershock productivity and temporal decay are valuable immediately after a large earthquake, when sequence-specific observations are still sparse and catalogue completeness may be degraded. Here I present PISAS v1.0, a physics-informed neural surrogate that predicts the quality-controlled number of aftershocks expected during the first 30 d and the Omori–Utsu decay exponent using only information available at mainshock origin time. The model uses mainshock magnitude, hypocentral depth, and three multiscale measures of pre-mainshock seismicity derived from the open U.S. Geological Survey Comprehensive Earthquake Catalog. An analytical decoder converts the neural outputs into a complete mean aftershock-rate history while enforcing positive productivity, bounded decay exponents, monotonic temporal decay, and exact agreement between the predicted event count and the integral of the reconstructed rate. The primary dataset contains 175 independent global sequences from 2000–2025, with model development based on rolling-origin validation and final evaluation on an untouched 2020–2025 temporal test. PISAS reproduced a moderate and transferable component of 30 d productivity variability, with coefficients of determination of 0.454 in out-of-fold validation and 0.534 on the temporal test. Relative to a magnitude-only linear model, productivity root-mean-square error improved by 3.9 % out of fold, 5.0 % on the temporal test, and 4.3 % under leave-one-region-out validation. Prediction of the decay exponent was substantially weaker, with near-zero or negative explained variance in out-of-fold and regional evaluation. Mainshock magnitude was the dominant productivity predictor, whereas hypocentral depth provided the most stable, but limited, information about temporal decay. Compared with an otherwise equivalent unconstrained neural network, PISAS matched or improved predictive accuracy while eliminating all violations of rate normalization and output admissibility. PISAS v1.0 therefore provides a reproducible, physically consistent initial estimate of aftershock productivity and mean temporal evolution, while clearly delimiting the weaker predictability of sequence-specific decay.

Introduction

Large earthquakes are commonly followed by elevated seismic activity that can persist for days to years and substantially modify the short-term probability of damaging ground motion. Rapid characterization of this evolving hazard is therefore an important component of operational earthquake forecasting and post-event decision support (Jordan et al., 2011; Hardebeck et al., 2024). Two quantities are especially relevant for describing the first-order evolution of an aftershock sequence: its productivity, which controls the expected number of events above a specified magnitude threshold, and its temporal decay, commonly represented by the modified Omori–Utsu law. The latter expresses the mean aftershock rate as a power-law function of elapsed time and remains one of the most widely used empirical descriptions of post-mainshock seismicity (Utsu et al., 1995).

Statistical aftershock forecasting has traditionally relied on empirical scaling laws and point-process models. In epidemic-type aftershock-sequence models, the conditional seismicity rate is represented as the superposition of background activity and cascades triggered by previous earthquakes, with productivity increasing with parent-event magnitude and temporal decay governed by an Omori-type kernel (Ogata, 1998). Such models provide a rigorous stochastic representation of clustered seismicity and underpin several operational forecasting frameworks. They are particularly effective once enough post-mainshock observations have accumulated to estimate sequence-specific parameters, and Bayesian or likelihood-based updating can extract useful information from the early part of an ongoing sequence (Omi et al., 2014). However, the first hours following a large earthquake are also those in which catalogue completeness is most strongly degraded and sequence-specific estimates are least stable.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

Global aftershock forecasting is further complicated by substantial intersequence variability. Mainshock magnitude provides the dominant first-order control on the expected number of aftershocks, but earthquakes of comparable magnitude can differ in productivity by orders of magnitude (Helmstetter et al., 2005; Marsan and Helmstetter, 2017; Dascher-Cousineau et al., 2020). Regional tectonic setting, seismogenic volume, focal depth, rupture geometry, stress drop, magnitude calibration, and time-dependent catalogue incompleteness can all contribute to departures from a universal magnitude–productivity relation (Wetzler et al., 2016). A complementary global analysis based on fixed 30 d association windows found that mainshock magnitude and tectonic setting explain the dominant component of productivity variability, whereas depth and distance from plate boundaries exert weaker secondary effects (D'Alessandro, 2026b). Global operational models have consequently benefited from accounting for tectonic region, detection changes and intersequence variability rather than transferring a single parameter set among all earthquakes (Page et al., 2016).

The predictability of the temporal decay exponent is less clear. The exponent p varies among fitted sequences, but its estimation is sensitive to finite event counts, the adopted lower time cutoff, missing early events, magnitude completeness and the treatment of secondary triggering. Consequently, part of the observed variation may reflect physical differences among sequences, while another part may arise from catalogue and estimation uncertainty. This makes p a substantially more difficult target than cumulative productivity and motivates treating productivity and decay as distinct, rather than assuming that both are equally predictable from mainshock properties.

Machine-learning methods offer a flexible way to represent nonlinear relationships among source properties, background seismicity and subsequent earthquake occurrence. Neural networks have, for example, been used to predict spatial aftershock patterns from mainshock stress-change information and to formulate flexible neural point-process models of evolving seismicity (DeVries et al., 2018; Dascher-Cousineau et al., 2023; Stockman et al., 2023). These studies demonstrate that data-driven models can identify predictive structure not fully represented by conventional low-dimensional relationships. At the same time, comparisons with simpler statistical models have shown that apparent gains from complex architectures must be evaluated against strong baselines and under genuinely independent validation (Mignan and Broccardo, 2019).

A further limitation of unconstrained multi-output learning is that individually plausible predictions need not jointly define a physically or statistically coherent process. If productivity, the Omori–Utsu exponent and the rate normalization are predicted independently, the integral of the resulting rate curve may disagree with the predicted event count. Negative rates, inadmissible exponents or non-decaying rate histories can likewise arise unless they are prevented explicitly. These inconsistencies are particularly undesirable when the model is intended to provide a compact surrogate for an analytical seismicity law rather than merely predict unrelated scalar targets.

Physics-guided machine learning provides one route for combining the flexibility of neural networks with known scientific structure. Instead of asking the training data to rediscover every admissibility condition, constraints can be incorporated directly into the architecture or analytical output transformation. Such approaches can reduce the feasible solution space, improve physical consistency and limit implausible extrapolation, especially when training data are scarce (Karpatne et al., 2017; Daw et al., 2020; Karniadakis et al., 2021). In the present context, the modified Omori–Utsu relation supplies a natural analytical decoder: a neural model can predict independent sequence descriptors, while positivity, parameter bounds, monotonic decay and integral normalization are imposed exactly rather than through soft penalties.

What remains insufficiently developed is a reproducible global framework that combines open seismic catalogues, mainshock-time predictors, strict temporal and regional validation, neural nonlinearities and exact closure of the reconstructed aftershock-rate history. Existing approaches generally emphasize full point-process forecasting, regional sequence classification, spatial aftershock occurrence or statistical updating after the sequence has begun. Recent extensions of the ETAS framework have begun to incorporate mainshock source information to improve the initial forecasting of aftershock activity (Asayesh et al., 2025). A complementary problem is to estimate an immediate, physically admissible prior for productivity and decay before informative aftershock observations are available. Such a surrogate would not replace epidemic-type models or operational probabilistic forecasts; rather, it could provide a globally trained initial condition that is subsequently updated as the sequence develops.

Here I introduce PISAS v1.0, a Physics-Informed Surrogate for Aftershock Sequences constructed from the open U.S. Geological Survey Comprehensive Earthquake Catalog. PISAS maps five quantities available at mainshock origin time—magnitude, hypocentral depth and three multiscale measures of pre-mainshock seismicity—to two independent targets: the logarithm of the quality-controlled 30 d aftershock count and the

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

Omori–Utsu exponent p. An analytical decoder then reconstructs the productivity coefficient and mean temporal rate while enforcing positive productivity, bounded p, positive monotonic decay and exact agreement between the 30 d rate integral and the predicted count.

The study addresses four questions. First, can a compact nonlinear surrogate improve global aftershock-productivity prediction beyond a magnitude-only scaling relation? Second, is the temporal decay exponent predictably related to the same mainshock-time information? Third, does embedding the Omori–Utsu closure directly in the architecture improve reliability without degrading statistical accuracy? Fourth, do the learned relationships generalize to later earthquakes and geographically withheld regions, and can their predictor dependence be interpreted consistently across global and local diagnostics? I address these questions using rolling-origin cross-validation, an untouched 2020–2025 temporal test, leave-one-region-out experiments, physical-admissibility tests, predictor ablations and ensemble-aware interpretation.

The remainder of the manuscript is organized as follows. Section 2 describes catalogue processing and sequence construction. Section 3 presents the PISAS architecture and analytical decoder, and Sect. 4 defines training, model selection and evaluation. Sections 5–8 report predictive performance, physical consistency, temporal and regional transfer, and model interpretation, respectively. Section 9 discusses the scientific scope, limitations and potential operational role of the model, and Sect. 10 summarizes the principal conclusions.

2 Data and sequence construction

I constructed the sequence-level dataset from the Advanced National Seismic System Comprehensive Earthquake Catalog (ComCat), maintained by the U.S. Geological Survey Earthquake Hazards Program and accessed through the FDSN Event Web Service (U.S. Geological Survey, 2017). ComCat combines source parameters and associated products contributed by multiple seismic networks, including origin times, hypocentral locations, preferred magnitudes, magnitude types, and focal-mechanism information when available. I queried the global catalogue for events classified as earthquakes with reported magnitude $M \geq 4.0$ between 1 January 1999 and 30 January 2026. The principal analysis interval was restricted to mainshocks occurring from 1 January 2000 to 31 December 2025. The preceding year provided the complete 365 d history required to calculate pre-mainshock seismicity rates and to identify dependent candidates near the beginning of the analysis interval, whereas the final 30 d buffer provided complete post-mainshock exposure for events occurring near the end of 2025.

The initial query returned 368 077 event records. Quality control required a valid event identifier, origin time, latitude, longitude, focal depth, and magnitude. Coordinates outside their physically admissible ranges were removed, longitudes were normalized to $[-180°180°)$, and duplicated event identifiers were resolved by retaining the most recently updated ComCat solution. One invalid record was removed, leaving a quality-controlled catalogue of 368 076 earthquakes. Because the preferred ComCat solution can originate from different contributing networks and magnitude scales, magnitude type was preserved as a diagnostic attribute, and the preferred magnitude was not assumed to constitute a perfectly homogeneous physical magnitude scale.

Figure 1 summarizes the spatial coverage, annual occurrence rates, frequency–magnitude distribution, and temporal detectability of the catalogue. Annual completeness magnitude M_c was estimated using the 90 % goodness-of-fit criterion, which compares the observed frequency–magnitude distribution with the Gutenberg–Richter distribution fitted above a trial cutoff (Wiemer and Wyss, 2000). The annual M_c estimates were computed from the quality-controlled catalogue before any event was classified as a dependent mainshock or assigned to an aftershock sequence. This ordering preserves M_c as a diagnostic of catalogue detectability rather than of the subsequently constructed sequence sample, because declustering and other selective event-removal procedures can reshape the lower-magnitude portion of the frequency–magnitude distribution and thereby alter catalogue-based completeness estimates (Guastella et al., 2026). The resulting annual estimates are generally lower than the $M = 4.5$ threshold adopted for aftershock extraction, although M_c approaches or slightly exceeds this threshold during part of 2009–2012. I retained this interval to preserve the temporal extent of the analysis but considered the locally reduced detectability when interpreting sequence productivity and temporal generalization.

Candidate mainshocks were required to have $M \geq 6.5$, focal depth $z \leq 70$ km, and origin time within the 2000–2025 analysis interval. Following the general logic of magnitude-dependent window declustering and adaptive space–time interaction zones (Gardner and Knopoff, 1974; Reasenberg, 1985), I defined study-specific spatial and temporal association windows as

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

$$R_{\mathrm{eff}}(M) = \min[500, 10^{-1.471+0.5M}]\ \mathrm{km}, \qquad (1)$$

and

$$T_{\mathrm{eff}}(M) = \min\{365, \max[30, 10^{-5.141+M}]\}\ \mathrm{d}. \qquad (2)$$

These expressions were fixed a priori in the PISAS sequence-building pipeline and should not be interpreted as the original Gardner–Knopoff or Reasenberg parametrizations. They provide smoothly magnitude-dependent windows while limiting the maximum search domain to 500 km and 365 d.
Candidates were processed chronologically. A candidate was classified as dependent when an earlier independent candidate had equal or larger magnitude and the later candidate occurred within the earlier event's R_{eff} and T_{eff} windows. When more than one potential parent was available, I assigned the candidate to the parent minimizing the normalized space–time distance

$$\eta = \left[\left(\frac{r}{R_{\mathrm{eff}}}\right)^2 + \left(\frac{\Delta t}{T_{\mathrm{eff}}}\right)^2\right]^{1/2}, \qquad (3)$$

where r is the epicentral great-circle distance and $\Delta t$is the elapsed time from the potential parent. Mainshocks in the 1999 warm-up interval participated in the dependence test so that candidates occurring near the start of 2000 were not incorrectly classified as independent. Among 913 candidates in the principal analysis interval, 707 were classified as independent and 206 as dependent.
Aftershocks were selected from quality-controlled catalogue events with $M \geq 4.5$. An event was initially eligible for an independent mainshock when it occurred after that mainshock, had magnitude lower than the mainshock magnitude, and fell within the associated spatial and temporal windows. I additionally searched for the first subsequent earthquake larger than the mainshock within $R_{\mathrm{eff}}$and the available T_{eff}interval. When such an event occurred, the association window was truncated at its origin time. Independent sequences followed by a larger earthquake during the first 30 d were excluded because their target interval could not be interpreted as an uninterrupted aftershock sequence of the original mainshock.
Magnitude-dependent association windows can overlap, particularly in tectonically active regions and during periods of elevated seismicity. I therefore assigned each eligible event uniquely to the independent mainshock that minimized Eq. (3). Independent mainshocks in the warm-up interval were included in this competition to prevent preferential assignment to sequences beginning after 1 January 2000. The procedure yielded 37 259 uniquely assigned aftershocks. The complete sequence-building workflow, including candidate screening, dependence classification, competitive event assignment, truncation at subsequent larger events, and sequence-level filtering, is summarized in Fig. 2a.
For each retained sequence, I defined the quality-controlled 30 d productivity as

$$N_{30}^{qc} = \sum_i \mathbb{I}\,(0.1, \leq t_i \leq 30\ \mathrm{d}\ M_i \geq 4.5), \qquad (4)$$

where t_i is the elapsed time from the mainshock and $\mathbb{I}(\cdot)$ is the indicator function. The sum includes only uniquely assigned aftershocks. Events occurring during the first 0.1 d were excluded because catalogue completeness is strongly rate dependent immediately after large earthquakes. Overlapping mainshock and aftershock coda waves can obscure smaller events, bias observed early-time rates, and distort estimates of the Omori–Utsu parameters (Hainzl, 2016; de Arcangelis et al., 2018). The first independent surrogate target was therefore

$$y_N = \log_{10} N_{30}^{qc}. \qquad (5)$$

The fixed lower cutoff does not eliminate every possible detectability difference among regions and sequences, but it reduces the most severe and least consistently observable part of the post-mainshock interval.
I represented the temporal decay of each sequence using the modified Omori–Utsu law

$$\lambda(t) = K(t+c)^{-p},\ c = 0.01\ \mathrm{d}, \qquad (6)$$

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

over the interval $0.1 \leq t \leq 30$ d. The modified Omori–Utsu law provides a standard representation of aftershock-rate decay, although its short-time parameter $c$can be strongly affected by early catalogue incompleteness (Utsu et al., 1995). I therefore fixed $c$rather than attempting to infer it independently from sequences whose earliest 0.1 d were excluded.

Following the non-stationary Poisson-process formulation of Ogata (1983), the log-likelihood for a sequence containing N_{fit} events at times t_i was

$$\ell(K, p) = \sum_{i=1}^{N_{\text{fit}}} \log[K(t_i + c)^{-p}] - K \int_{0.1}^{30} (t + c)^{-p} \, \mathrm{d}t. \qquad (7)$$

For each trial value of p, the productivity coefficient was profiled analytically as

$$\hat{K}(p) = \frac{N_{\text{fit}}}{\int_{0.1}^{30} (t + c)^{-p} \, \mathrm{d}t}, \qquad (8)$$

and p was estimated by maximizing the profiled likelihood subject to $0.2 \leq p \leq 2.5$. Under the common magnitude and time limits used here, $N_{\text{fit}} = N_{30}^{qc}$. Consequently, K is fully determined by N_{30}^{qc} and p and was retained as a derived diagnostic rather than treated as a third independent learning target. The second independent surrogate target was the maximum-likelihood estimate $\hat{p}$. Sequence-specific 95 % uncertainty intervals for $p$and K were obtained from 300 non-parametric bootstrap resamples of the fitted aftershock times. The likelihood formulation follows the direct point-process treatment introduced by Ogata (1983), whereas the broader empirical interpretation of K, c, and p follows Utsu et al. (1995).

A sequence entered the primary dataset when it had complete 30 d catalogue exposure, was not followed by a larger event during that interval, produced a converged Omori–Utsu fit, yielded an estimate of $p$more than 0.01 from either optimization bound, and contained $N_{\text{fit}} \geq 20$ aftershocks. These criteria produced 175 primary sequences. A broader sensitivity dataset was generated using the same physical and quality-control criteria but reducing the minimum information requirement to $N_{\text{fit}} \geq 10$. It contains 276 sequences, comprising the complete primary dataset and 101 additional lower-information sequences. Twenty-six otherwise independent sequences were excluded because a subsequent larger earthquake occurred during their 30 d target interval. Figure 2b illustrates the fitting procedure for a representative primary sequence, while Fig. 2c–d shows the productivity–magnitude relationship, the distribution of p, and the magnitude-type composition of the retained events.

Finally, I calculated pre-mainshock background-seismicity rates using only earthquakes with $M \geq 4.5$ that occurred before the candidate mainshock. For a temporal window W and epicentral radius R, I defined

$$\mu_{W,R} = \frac{1}{W} \sum_{i} \mathbb{I}(-W \leq t_i < 0, r_i \leq R, M_i \geq 4.5). \qquad (9)$$

The primary predictor set retained three paired scales: $\mu_{30,100}$, $\mu_{90,200}$, and $\mu_{365,300}$, where the first subscript denotes days and the second kilometres. The inclusion of pre-mainshock seismicity rates is motivated by evidence that aftershocks preferentially occur in areas that were already seismically active before the mainshock (Page and van der Elst, 2022). Together with mainshock magnitude and focal depth, these quantities are available at the mainshock origin time and therefore introduce no post-event information. Association-window dimensions, aftershock counts, fitted parameters, overlap diagnostics, magnitude-type composition, subsequent-event metadata, and bootstrap uncertainties were retained for traceability and quality assessment but excluded from the primary predictor vector.

3 PISAS model formulation

PISAS was formulated as a sequence-level neural surrogate that maps information available at the mainshock origin time to two independent descriptors of the subsequent aftershock sequence: the quality-controlled 30 d

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

productivity and the Omori–Utsu decay exponent. The model does not ingest aftershock observations, elapsed post-mainshock time, or any other variable measured after the mainshock. Its predictor vector is

$$x = \left[M, z, \log\left(1+\mu_{30,100}\right) \log\left(1+\mu_{90,200}\right) \log\left(1+\mu_{365,300}\right)\right]^{\mathrm{T}}, \qquad (10)$$

where M and z are the mainshock magnitude and hypocentral depth, respectively, and the three $\mu_{W,R}$ terms are the pre-mainshock background rates defined in Eq. (9). The logarithmic transformation reduces the strong positive skew of the rate variables while retaining zero-rate cases. Each predictor was standardized as

$$\tilde{x}_j = \frac{x_j - \bar{x}_j}{s_j}, \qquad (11)$$

where $\bar{x}_j$ and s_j are the mean and standard deviation estimated exclusively from the training portion of the relevant cross-validation fold or from the complete development dataset for final-model training. Applying the transformation separately within each training partition prevented information from the validation or temporal-test intervals from entering the input scaling.

The model combines a data-driven shared representation with an analytical decoder that enforces the adopted sequence-rate formulation. This design follows the general principle of physics-informed and physics-guided machine learning, in which prior scientific knowledge restricts the set of admissible model outputs rather than leaving all output relationships to be inferred from data alone (Raissi et al., 2019; Daw et al., 2020). PISAS is not a physics-informed neural network in the narrow sense of a solver constrained by partial-differential-equation residuals. Instead, it is a supervised neural surrogate whose output architecture hard-codes the modified Omori–Utsu relation, the exact productivity closure, positivity, bounded decay exponents, and monotonic temporal decay. The complete architecture is summarized in Fig. 3.

The standardized predictor vector is passed through a shared multilayer perceptron. For the deployed PISAS v1.0 configuration, selected using the temporal cross-validation procedure described in Sect. 4, the hidden representation is

$$h_1 = \mathrm{GELU}(W_1\tilde{x} + b_1),\ h_1 \in \mathbb{R}^{32}, \qquad (12)$$

$$\tilde{h}_1 = \mathrm{Dropout}(h_1; q),\ q = 0.10, \qquad (13)$$

and

$$h_2 = \mathrm{GELU}\left(W_2\tilde{h}_1 + b_2\right),\ h_2 \in \mathbb{R}^{16}. \qquad (14)$$

The Gaussian Error Linear Unit (GELU) was used as a smooth nonlinear activation, and dropout was applied only during training to limit co-adaptation in the relatively small sequence-level dataset (Hendrycks and Gimpel, 2016; Srivastava et al., 2014). Dropout is disabled during deterministic inference. The two hidden layers and output heads contain 754 trainable parameters in total.

Two scalar heads act on the shared latent representation. The productivity head is linear,

$$\hat{y}_N = w_N^{\mathrm{T}} h_2 + b_N, \qquad (15)$$

and predicts the logarithmic target defined in Eq. (5). The corresponding 30 d productivity is recovered as

$$\hat{N}_{30}^{qc} = 10^{\hat{y}_N}, \qquad (16)$$

which guarantees strictly positive productivity without clipping. The decay head first produces an unconstrained latent value,

$$a_p = w_p^{\mathrm{T}} h_2 + b_p, \qquad (17)$$

which is transformed through a scaled logistic function,

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

$$\hat{p} = p_{\min} + (p_{\max} - p_{\min})\sigma(a_p),\ p_{\min} = 0.2,\ p_{\max} = 2.5, \quad (18)$$

where $\sigma(a) = [1 + \exp(-a)]^{-1}$. Equation (18) imposes the same admissible interval used during sequence fitting and prevents the neural surrogate from generating negative or implausibly large decay exponents. The productivity and decay heads share the hidden representation but remain separate output mappings; consequently, the model can learn common dependence on the mainshock and background state without imposing a direct empirical regression between $\hat{N}_{30}^{qc}$ and $\hat{p}$.

The productivity coefficient K is not predicted by an additional neural head. Instead, it is reconstructed analytically from the two independent outputs. For the fixed fitting interval $t_0 = 0.1$ d to $t_1 = 30$ d and $c = 0.01$ d, I define

$$I(p) = \int_{t_0}^{t_1} (t + c)^{-p}\, \mathrm{d}t = \begin{cases} \dfrac{(t_1 + c)^{1-p} - (t_0 + c)^{1-p}}{1 - p}, & p \neq 1, \\ \log\left(\dfrac{t_1 + c}{t_0 + c}\right), & p = 1. \end{cases} \quad (19)$$

The predicted productivity coefficient is then

$$\hat{K} = \frac{\hat{N}_{30}^{qc}}{I(\hat{p})}, \quad (20)$$

and the complete predicted aftershock-rate curve is

$$\hat{\lambda}(t) = \hat{K}(t + c)^{-\hat{p}},\ t_0 \leq t \leq t_1. \quad (21)$$

This decoder retains the empirical modified Omori–Utsu representation introduced in Sect. 2 but makes its normalization an exact deterministic consequence of the learned productivity and decay exponent (Utsu et al., 1995).

The construction in Eqs. (19)–(21) imposes exact sequence-level closure,

$$\int_{t_0}^{t_1} \hat{\lambda}(t)\, \mathrm{d}t = \hat{K} I(\hat{p}) = \hat{N}_{30}^{qc}. \quad (22)$$

Because $\hat{N}_{30}^{qc} > 0$, $I(\hat{p}) > 0$, and $\hat{p} > 0$, it follows that

$$\hat{K} > 0,\ \hat{\lambda}(t) > 0, \quad (23)$$

And

$$\frac{\mathrm{d}\hat{\lambda}}{\mathrm{d}t} = -\hat{p}\hat{K}(t + c)^{-\hat{p}-1} < 0. \quad (24)$$

Thus, every PISAS prediction is positive, monotonically decreasing over the target interval, bounded in p, and exactly normalized to the independently predicted 30 d productivity. These properties are architectural constraints rather than soft penalties: they hold for every parameter configuration and do not depend on the availability of labelled examples at a particular input location. This distinction is central to the model design because an unconstrained network can attain comparable prediction errors while producing output combinations that do not define a self-consistent aftershock-rate curve.

The neural component therefore learns only the nonlinear map

$$\mathcal{F}_\theta : x \mapsto (\hat{y}_N, \hat{p}), \quad (25)$$

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

where θ denotes the 754 trainable weights and biases. The analytical decoder subsequently maps these two outputs to $\widehat{K}$ and $\hat{\lambda}(t)$ without introducing additional trainable parameters. This separation reduces output redundancy, preserves the statistical definition of the two sequence-level targets, and ensures that the predicted rate history remains consistent with the same temporal integral used to construct the observed productivity.

4 Training and evaluation framework

All model development, hyperparameter selection, and primary performance evaluation were conducted using the 175-sequence primary dataset defined in Sect. 2. The 276-sequence sensitivity dataset was excluded from model selection and was used only for the robustness analyses described in Sect. 6. I partitioned the primary dataset chronologically into a development interval containing 143 mainshocks from 2000–2019 and an untouched temporal-test interval containing 32 mainshocks from 2020–2025. The test sequences were not used to estimate feature transformations, select the neural architecture, tune optimization parameters, determine the number of training epochs, or compare candidate configurations. This strict temporal separation was adopted to approximate prospective application to mainshocks occurring after the period represented in model development.

Within the development interval, I used an expanding-window rolling-origin design, which preserves temporal ordering and evaluates transfer from earlier to later periods rather than mixing sequences randomly across time (Tashman, 2000). The four folds were defined as follows: training on 2000–2004 and validating on 2005–2008; training on 2000–2008 and validating on 2009–2012; training on 2000–2012 and validating on 2013–2016; and training on 2000–2016 and validating on 2017–2019. The corresponding training and validation sample sizes were 29 and 28, 57 and 37, 94 and 28, and 122 and 21 sequences, respectively. The four non-overlapping validation blocks therefore provided out-of-fold predictions for 114 development sequences. The earliest 29 sequences served only as training observations because no preceding data were available from which to construct a temporally valid model for them.

All feature transformations described in Sect. 3 were fitted independently within each fold. In particular, the means and standard deviations used to standardize the predictors were calculated from the corresponding training block and subsequently applied without modification to its validation block. The same procedure was used for the linear benchmarks. This prevented future validation-period information from entering either the neural or statistical models through preprocessing.

The two output targets have different error structures. The logarithmic productivity target is directly determined by an integer aftershock count, whereas the fitted decay exponent has sequence-dependent uncertainty that increases for sparse or weakly constrained sequences. I therefore optimized a composite robust objective based on the Huber loss (Huber, 1964),

$$\rho_\delta(r) = \begin{cases} \frac{1}{2} r^2, & |\, r \,| \leq \delta, \\ \delta\left(|\, r \,| - \frac{\delta}{2}\right), & |\, r \,| > \delta, \end{cases} \quad (26)$$

where r denotes a prediction residual. Relative to squared error, this loss reduces the influence of unusually large residuals while retaining quadratic behaviour near the optimum.

For sequence i, I approximated the uncertainty of the fitted p value from its bootstrap confidence interval as

$$\sigma_{p,i} = \max\left[\frac{p_{i,\mathrm{high}} - p_{i,\mathrm{low}}}{2 z_{0.975}}, 0.03\right], \; z_{0.975} = 1.95996. \quad (27)$$

The preliminary inverse-variance weight was $u_i = \sigma_{p,i}^{-2}$. To prevent a few apparently precise sequences from dominating the decay-target loss, these weights were clipped at their 5th and 95th percentiles within the evaluated partition and normalized to unit mean,

$$w_i = \frac{\mathrm{clip}(u_i, Q_{0.05}, Q_{0.95})}{n^{-1} \sum_{j=1}^{n} \mathrm{clip}(u_j, Q_{0.05}, Q_{0.95})}. \quad (28)$$

The training objective was then

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \rho_{0.20}\left(\hat{y}_{N,i} - y_{N,i}\right) + \lambda_p \frac{\sum_{i=1}^{n} w_i \, \rho_{0.15}(\hat{p}_i - p_i)}{\sum_{i=1}^{n} w_i}, \ \lambda_p = 1. \quad (29)$$

Thus, productivity and decay contributed equally at the loss-component level, while individual $p$observations were weighted according to their estimated precision. The validation objective had the same form, with uncertainty weights computed from the validation sequences. These weights affected training and model selection but were not used in the unweighted predictive metrics reported in Sect. 5.

I evaluated eight neural configurations formed by the Cartesian product of first-layer widths $h_1 \in \{16, 32\}$, second-layer widths $h_2 \in \{8, 16\}$, and dropout fractions $q \in \{0, 0.10\}$. The learning rate was fixed at 10^{-3}. Each configuration was trained on all four rolling-origin folds using three fixed random seeds, 20260712, 20260713, and 20260714, giving 12 independent fits per configuration and 96 cross-validation fits in total. Model weights were initialized using the Xavier uniform scheme (Glorot and Bengio, 2010).

Optimization used Adam W, which combines the adaptive moment updates of Adam with decoupled weight decay (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). The batch size was 32 sequences, and the weight-decay coefficient was 10^{-4}. Training was allowed to continue for at most 500 epochs, with a minimum of 30 epochs. Early stopping was triggered when the validation composite loss failed to improve by at least $10^{-5}$for 60 consecutive epochs. The Euclidean gradient norm was clipped at 5.0 as a safeguard against unstable updates, although no clipping event occurred in the selected cross-validation or final-model runs. All seed-level calculations used deterministic PyTorch operations where available.

Candidate configurations were ranked by their mean minimum validation loss across the 12 fold–seed runs. Ties were resolved first by the standard deviation of validation loss and then by the number of trainable parameters. Figure 4a shows the resulting hyperparameter ranking. The selected configuration used $h_1 = 32$, $h_2 = 16$, $q = 0.10$, and a learning rate of 10^{-3}, corresponding to the 754-parameter architecture described in Sect. 3. Its mean minimum validation loss was 0.04296, with a standard deviation of 0.00666 across the 12 runs.

The selected-model training trajectories are shown in Fig. 4b. The best epoch ranged from 113 to 500 across folds and seeds, with a median of 249.5 epochs (Fig. 4c). I therefore retrained the selected architecture on all 143 development sequences for a fixed 250 epochs. Fixing the final duration from the cross-validation median avoided monitoring the untouched test set and provided a common training budget across the final ensemble members. Five final models were trained using seeds 20260721–20260725. The largest gradient norm observed during the selected cross-validation runs was 0.411, substantially below the clipping threshold, and the optimization remained numerically stable throughout training (Fig. 4d). The maximum numerical discrepancy in the analytical closure relation of Eq. (22) was 5.68×10^{-14}in cross-validation and 2.27×10^{-13}in the final models.

For each validation or test sequence, I averaged the predictions across the relevant seed ensemble. Out-of-fold estimates were means over the three models trained for the corresponding temporal fold, whereas final temporal-test estimates were means over the five models trained on the complete development set. The associated seed-wise standard deviations were retained as measures of optimization and initialization sensitivity. They should not be interpreted as calibrated predictive confidence intervals because they represent only between-seed model dispersion and do not include catalogue, target-estimation, or irreducible sequence variability.

I compared PISAS against two transparent regression benchmarks. The first was an unregularized magnitude-only linear model. The second was a ridge-regression model using the complete five-predictor vector and a fixed penalty $\alpha = 1$ (Hoerl and Kennard, 1970). For a standardized design matrix X, ridge coefficients were obtained from

$$\hat{\beta}_\alpha = \left[\|y - X\beta\|_2^2 + \alpha\|\beta\|_2^2\right], \quad (30)$$

with the intercept excluded from penalization. Both benchmarks were fitted independently for y_N and p within every temporal fold and were refitted on the full development interval before temporal-test evaluation. Linear predictions of p were restricted to the same interval $[0.2, 2.5]$imposed on PISAS.

Predictive performance was quantified using the mean absolute error, root-mean-square error, and signed bias,

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | r_i |, \ \ \mathrm{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n} r_i^2\right)^{1/2}, \ B = \frac{1}{n}\sum_{i=1}^{n} r_i, \qquad (31)$$

where $r_i = \hat{y}_i - y_i$. I additionally calculated the coefficient of determination,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad (32)$$

and the Pearson correlation coefficient between observed and predicted values. Negative R^2 values were retained rather than truncated because they indicate performance inferior to prediction by the evaluation-set mean. Relative improvement over the magnitude-only benchmark was expressed as RMSE skill,

$$S_{\mathrm{RMSE}} = 1 - \frac{\mathrm{RMSE}_{\mathrm{model}}}{\mathrm{RMSE}_{M\text{-only}}}. \qquad (33)$$

Positive values of S_{RMSE} indicate improvement over the magnitude-only relation, whereas negative values indicate degradation. All metrics were calculated from the seed-ensemble mean predictions, separately for the 114 rolling-origin out-of-fold sequences and the 32 untouched temporal-test sequences.

5 Predictive performance

Figure 5 summarizes the predictive performance of PISAS under rolling-origin evaluation and on the untouched 2020–2025 temporal test. All reported quantities were calculated from the ensemble-mean predictions defined in Sect. 4. The out-of-fold results comprise the 114 sequences belonging to the four non-overlapping temporal validation blocks, whereas the final test results comprise 32 sequences excluded from all stages of model development.

For the logarithmic productivity target, the pooled out-of-fold predictions reproduced a substantial fraction of the observed between-sequence variability (Fig. 5a). PISAS attained a mean absolute error of 0.220, a root-mean-square error of 0.282, and a coefficient of determination of $R^2 = 0.454$. The Pearson correlation between observed and predicted values was $r = 0.699$. The mean signed residual was -0.048, indicating a modest overall tendency to underpredict productivity. This bias was small relative to the total dispersion of $\log_{10} N_{30}^{qc}$, although the predictions showed shrinkage towards the central part of the observed distribution. In particular, the most productive sequences were generally underestimated, whereas some of the least productive sequences were overestimated. This behaviour indicates that the model captured the dominant productivity scaling but only partly reproduced the extreme sequence-to-sequence departures from it.

Prediction of the Omori–Utsu exponent was substantially weaker (Fig. 5b). The pooled out-of-fold mean absolute error was 0.170 and the root-mean-square error was 0.214. The signed bias was effectively zero, $B = 2.8 \times 10^{-5}$, but the coefficient of determination was negative, $R^2 = -0.069$, and the Pearson correlation was only $r = 0.151$. The near-zero bias therefore reflects calibration of the pooled mean rather than successful discrimination among individual decay exponents. Predicted values occupied a narrower range than the observations, and the model reproduced little of the sequence-to-sequence variability in p. The five mainshock-time predictors consequently contain considerably less stable information about temporal decay than about cumulative 30 d productivity.

Performance on the untouched temporal test was consistent with the cross-validation assessment (Fig. 5c). For $\log_{10} N_{30}^{qc}$, PISAS achieved a mean absolute error of 0.250, a root-mean-square error of 0.310, $R^2 = 0.534$, and $r = 0.745$. The signed bias was -0.066, again indicating moderate underprediction on average. Although the absolute errors were slightly larger than in the pooled out-of-fold sample, the explained variance increased because the 2020–2025 sequences spanned a broader productivity range. The persistence of positive $R^2$and strong correlation under strict temporal separation indicates that the learned productivity relation was not confined to the development period.

For the decay exponent, temporal-test performance remained limited but was somewhat better than the pooled out-of-fold result. The mean absolute error was 0.161, the root-mean-square error was 0.201, and the bias was

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

+0.020. The coefficient of determination increased to $R^2 = 0.105$, with a Pearson correlation of $r = 0.367$. These values indicate weak but non-zero discrimination among the test sequences. The improvement relative to out-of-fold evaluation should nevertheless be interpreted cautiously because the test set contains only 32 sequences and the predicted $p$values still span a substantially narrower range than the observations.

The benchmark comparison in Fig. 5d clarifies the contribution of nonlinear multi-predictor learning. For out-of-fold productivity, the magnitude-only linear model yielded an RMSE of 0.294, the regularized five-predictor linear model also yielded 0.294, and PISAS yielded 0.282. The corresponding PISAS skill relative to the magnitude-only benchmark was +3.9 %. The regularized linear model did not improve on magnitude alone, with a skill of −0.2 %. Thus, focal depth and the three background-rate predictors provided measurable additional information for productivity only when incorporated through the nonlinear shared representation used by PISAS.

On the temporal test, the magnitude-only productivity RMSE was 0.327, the regularized multi-predictor linear RMSE was 0.317, and the PISAS RMSE was 0.310. Relative to the magnitude-only model, these values correspond to skills of +2.9 %for the regularized linear model and +5.0 % for PISAS. PISAS also produced the lowest test mean absolute error, 0.250 compared with 0.266 for the magnitude-only relation and 0.257 for the regularized linear benchmark. The nonlinear surrogate therefore provided a modest but reproducible reduction in predictive error beyond the first-order magnitude–productivity scaling documented in regional and global aftershock studies (Helmstetter et al., 2005; Dascher-Cousineau et al., 2020).

The benchmark results for $p$were less consistent. In pooled out-of-fold evaluation, the magnitude-only model achieved an RMSE of 0.209, compared with 0.214 for both the regularized multi-predictor model and PISAS. The resulting PISAS skill was −2.1 %, confirming that the neural model did not outperform the simplest benchmark for this target during rolling-origin validation. On the temporal test, the magnitude-only, regularized linear, and PISAS RMSE values were 0.213, 0.203, and 0.201, respectively. PISAS therefore attained a positive test skill of +5.7 %, marginally exceeding the +4.8 %improvement of the regularized linear model. This gain was small in absolute terms and does not alter the principal result that $p$remains only weakly predictable from the available mainshock-time information.

Overall, the predictive evaluation supports two distinct conclusions. First, PISAS extracts reproducible information about 30 d aftershock productivity beyond a magnitude-only scaling relation, with positive skill in both rolling-origin validation and the untouched temporal test. Second, the same predictors provide only limited information about the Omori–Utsu decay exponent. The near-zero or negative out-of-fold R^2, narrow predicted range, and small benchmark improvements indicate that much of the observed variability in p is unresolved by the adopted predictor set. Part of this variability may also reflect finite sequence sizes, short-term and magnitude-dependent catalogue incompleteness, and the sensitivity of clustering-model parameters to catalogue truncation (Hainzl, 2016; Harte, 2016). The physical-consistency, ablation, and sensitivity analyses presented in Sect. 6 evaluate whether these contrasting outcomes depend on the constrained architecture, predictor composition, or minimum sequence-information threshold.

6 Physical consistency and robustness

Figure 6 evaluates whether the physics-informed decoder improves the reliability of the reconstructed aftershock-rate histories, whether the five predictors contribute independently to model skill, and how performance changes when the minimum sequence-information requirement is relaxed. I compared PISAS with an unconstrained neural network that used the same five predictors, shared hidden architecture, temporal folds, random seeds, and optimization protocol. Unlike PISAS, however, the comparator used three independent linear output heads for $\log_{10} N_{30}^{qc}$, p, and $\ln K$. Its decay exponent was not restricted to the admissible interval, and the independently predicted $K$was not required to satisfy the integral closure relation in Eq. (22). This comparison therefore isolates the effect of the output formulation rather than differences in predictor information or evaluation design.

The constrained architecture provided equal or better predictive accuracy than the unconstrained network (Fig. 6a). For productivity, the pooled rolling-origin RMSE decreased from 0.306 for the unconstrained model to 0.282 for PISAS, corresponding to a reduction of approximately 7.6 %. The associated coefficient of determination increased from $R^2 = 0.360$to $R^2 = 0.454$. On the untouched temporal test, the productivity RMSE decreased from 0.317 to 0.310 and R^2increased from 0.512 to 0.534. The improvement was therefore larger during temporally ordered cross-validation than on the final test, but the constrained model remained the more accurate productivity predictor under both protocols.

For p, PISAS also improved the pooled out-of-fold result: the RMSE decreased from 0.225 to 0.214, and R^2 increased from -0.186 to -0.069. On the temporal test, the two formulations were effectively indistinguishable, with RMSE values of 0.2013 for PISAS and 0.2013 for the unconstrained network and $R^2 \approx 0.105$ for both. Thus, imposing the bounded decay head and analytical normalization did not degrade the prediction of either target. Instead, the constrained architecture modestly improved productivity and out-of-fold p performance while simultaneously removing physically inconsistent output combinations.

Physical-admissibility diagnostics demonstrate the principal distinction between the two formulations (Fig. 6b). A closure violation was identified when both the absolute and relative discrepancies between the independently predicted productivity and the integral of the reconstructed rate exceeded 10^{-10}. PISAS produced no closure violations, no values of p outside $[0.2, 2.5]$, no non-positive rates, and no monotonicity violations in either evaluation setting. Across all seed-level predictions, the maximum relative closure error was 1.56×10^{-16} out of fold and 1.65×10^{-16} on the temporal test, consistent with floating-point precision.

By contrast, every unconstrained sequence exhibited a closure violation under both evaluation protocols. The median relative closure discrepancy was 0.238 out of fold and 0.184 on the temporal test, while the corresponding maximum seed-level discrepancies reached 2.32 and 0.534. One of the 114 ensemble-mean out-of-fold predictions, equivalent to 0.88 %, also fell outside the prescribed range for p; no out-of-range value occurred on the temporal test. The independently predicted K values remained positive and the predicted p values were generally positive, so explicit monotonicity violations were not common. Nevertheless, the lack of normalization meant that the rate integral disagreed with the model's own productivity prediction for every sequence. The unconstrained network could therefore produce individually plausible values of $\log_{10} N_{30}^{qc}$, p, and $\ln K$ that did not jointly define a self-consistent aftershock sequence. This result illustrates the practical advantage of enforcing scientific relationships architecturally rather than treating them as optional post-processing conditions (Daw et al., 2020).

The leave-one-predictor-out experiments show that mainshock magnitude contains the dominant information for productivity (Fig. 6c). Removing M increased the pooled out-of-fold productivity RMSE by 45.0 % relative to the full model and reduced R^2 from 0.454 to -0.146. None of the remaining single-feature removals produced a comparable effect. Excluding $\mu_{365,300}$, $\mu_{90,200}$, and $\mu_{30,100}$ increased RMSE by 2.24 %, 1.34 %, and 0.37 %, respectively. Removing focal depth slightly reduced RMSE by 1.16 %, yielding $R^2 = 0.467$. The latter result does not establish that depth is universally uninformative; rather, it indicates that depth did not provide a stable incremental contribution after the other four predictors were retained in the rolling-origin training sets. The modest effects of removing individual background-rate variables are also compatible with partial redundancy among the three nested temporal and spatial scales.

Feature-ablation effects were much smaller and less stable for the decay exponent (Fig. 6d). Removing depth produced the largest deterioration, increasing the out-of-fold RMSE by 2.60 % and reducing R^2 from -0.069 to -0.126. Excluding $\mu_{365,300}$, $\mu_{30,100}$, or $\mu_{90,200}$ increased RMSE by only 0.58 %, 0.46 %, and 0.08 %, respectively. Removal of mainshock magnitude slightly improved RMSE by 0.98 %. The fold- and seed-level dispersion was substantially larger than these mean changes, and all ablated variants retained negative pooled out-of-fold R^2. Consequently, no individual predictor supplied a strong and consistently transferable control on p, although depth showed the largest average contribution. This result is consistent with the weak predictive skill reported in Sect. 5 and with the global importance diagnostics presented later in Sect. 8.

The sensitivity-only analysis examined the 101 additional sequences with $10 \leq N_{\text{fit}} < 20$, none of which contributed to training or model selection (Fig. 6e). Productivity performance degraded from RMSE values of 0.282 in primary out-of-fold evaluation and 0.310 on the primary temporal test to 0.458 for the complete sensitivity-only subset. The degradation was strongest for the sparsest sequences: RMSE was 0.525 for $N_{\text{fit}} = 10$–12, 0.449 for $N_{\text{fit}} = 13$–15, and 0.367 for $N_{\text{fit}} = 16$–19. The corresponding mean biases were $+0.512$, $+0.415$, and $+0.337$ in logarithmic productivity, showing systematic overprediction that weakened as the sequence count approached the primary threshold.

Prediction of p also deteriorated for the sensitivity-only sequences. Its RMSE increased to 0.278 for the complete subset, compared with 0.214 in primary out-of-fold evaluation and 0.201 on the primary temporal test. Within the three information classes, p RMSE decreased from 0.322 for $N_{\text{fit}} = 10$–12 to 0.254 for $N_{\text{fit}} = 13$–15 and 0.235 for $N_{\text{fit}} = 16$–19. The monotonic improvement with increasing event count supports the use of $N_{\text{fit}} \geq 20$ for the primary analysis. The sensitivity-only experiment combines two effects: reduced statistical information for estimating the decay law and a target-support shift created by selecting sequences whose observed counts fall below the primary threshold. Its poorer performance should therefore not be interpreted

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

solely as neural-model failure. It also reflects the known sensitivity of aftershock parameters to finite sample size, catalogue truncation, and rate-dependent detectability (Ogata, 1983; Hainzl, 2016; Harte, 2016).

The rate reconstructions in Fig. 6f illustrate how these structural differences propagate to complete aftershock histories. The three examples represent the 10th, 50th, and 90th percentiles of predicted temporal-test productivity: us7000pwkn with $M = 7.4$, $\widehat{N}_{30} = 36.9$, and $\hat{p} = 1.02$; us6000rfwz with $M = 7.4$, $\widehat{N}_{30} = 44.8$, and $\hat{p} = 0.68$; and us6000rtdt with $M = 7.6$, $\widehat{N}_{30} = 123.9$, and $\hat{p} = 0.81$. The PISAS curves are positive and monotonically decreasing, and their integrals over 0.1–30 d equal the corresponding predicted productivities by construction. Between-seed variability, represented by the interquartile envelopes, changes both the normalization and the decay shape but never violates the analytical closure. The unconstrained median curves can appear qualitatively similar, yet their independently estimated normalizations do not generally integrate to the productivity predicted by their separate output head.

Taken together, the analyses in Fig. 6 show that the physical decoder is not obtained at the expense of statistical performance. PISAS matches or improves the corresponding unconstrained neural network while guaranteeing admissible and internally consistent rate histories. Magnitude remains essential for productivity, whereas the additional predictors provide smaller, partly redundant corrections. The decay exponent is less robustly associated with any individual input and becomes increasingly uncertain in low-information sequences. These results justify retaining the constrained two-target formulation and the $N_{\mathrm{fit}} \geq 20$ primary threshold for the subsequent temporal and regional generalization analyses.

7 Temporal and regional generalization

Figure 7 evaluates whether PISAS retains predictive skill when the evaluation sequences are separated from the training data in time or geographic space. Temporal transfer was assessed using the four rolling-origin validation intervals and the untouched 2020–2025 test interval defined in Sect. 4. Regional transfer was examined independently through spatially blocked validation restricted to the 143 development sequences. These two experiments test different forms of generalization: temporal evaluation preserves the evolving catalogue and network context, whereas regional evaluation requires prediction for geographic domains that contribute no sequences to model fitting.

Productivity performance varied among temporal blocks but showed no systematic deterioration towards the end of the catalogue (Fig. 7a). PISAS yielded RMSE values of 0.276, 0.281, 0.274, 0.304, and 0.310 for the 2005–2008, 2009–2012, 2013–2016, 2017–2019, and 2020–2025 intervals, respectively. Relative to the magnitude-only benchmark, the corresponding RMSE skills were +4.1 %, +9.9 %, −10.2 %, +6.0 %, and +5.0 %. PISAS therefore outperformed the magnitude-only relation in four of the five chronological evaluation intervals. The isolated negative result for 2013–2016 indicates that the additional predictors do not improve every finite temporal sample, but it is not followed by a persistent loss of skill in the subsequent blocks.

The signed productivity residuals from the 114 rolling-origin validation sequences and 32 temporal-test sequences exhibited no significant monotonic relationship with mainshock year. The Spearman coefficient was $\rho = 0.068$ with $p = 0.416$. Absolute residual magnitude was also unrelated to time, with $\rho = 0.078$ and $p = 0.351$. Thus, neither systematic bias nor overall error amplitude increased detectably over the 2005–2025 evaluation period. The small positive linear slope of the signed residuals, 0.0042log units yr^{-1}, was dominated by intersequence dispersion and did not constitute evidence of temporal model drift.

Temporal performance for the decay exponent remained weaker and more variable (Fig. 7b). The PISAS RMSE values for $p$were 0.210, 0.176, 0.242, 0.239, and 0.201 across the same five intervals. RMSE skill relative to the magnitude-only model was −2.4 %, +0.3 %, +0.1 %, −7.7 %, and +5.7 %, respectively. Except for the 2017–2019 deterioration and the modest improvement on the temporal test, these differences were close to zero. The signed residuals showed no temporal trend ($\rho = 0.006$, $p = 0.943$), and neither did their absolute values ($\rho = 0.088$, $p = 0.294$). The limited prediction of $p$therefore persisted throughout the catalogue rather than emerging from progressive temporal degradation.

To construct geographically separated evaluation blocks, I converted the epicentral coordinates of each development mainshock to a three-dimensional unit vector,

$$s_i = \begin{bmatrix} \cos\phi_i \cos\lambda_i \\ \cos\phi_i \sin\lambda_i \\ \sin\phi_i \end{bmatrix}, \qquad (34)$$

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

where ϕ_i and λ_i are latitude and longitude in radians. I then applied spherical k-means with $k = 5$, 100 initializations, and fixed seed 20260730. Each sequence was assigned to the unit-norm centroid with the largest cosine similarity,

$$g_i = s_i^{\mathrm{T}} c_k, \; \|c_k\|_2 = 1. \qquad (35)$$

Spherical clustering avoids the discontinuity at the antimeridian and treats geographic positions as directional data on the unit sphere (Banerjee et al., 2005). The clustering was fitted exclusively to the 2000–2019 development coordinates; temporal-test sequences were assigned afterwards to the nearest frozen centroid and were displayed only to document geographic coverage. Spatially blocked evaluation provides a more demanding estimate of geographic transfer than random cross-validation because spatially neighbouring observations are prevented from appearing simultaneously in training and evaluation sets (Roberts et al., 2017). The resulting blocks were labelled descriptively from their centroid locations as South America, Eurasia–Indian Ocean, Maritime Southeast Asia, Northwest Pacific, and Southwest Pacific (Fig. 7c). They contained 20, 18, 35, 22, and 48 development sequences, respectively. The corresponding temporal-test counts were 3, 3, 8, 6, and 12. These labels identify geographically coherent data partitions rather than homogeneous tectonic regimes: individual blocks may include several plate boundaries, focal-mechanism populations, and network configurations. Their purpose is therefore predictive validation, not tectonic classification.

For each leave-one-region-out experiment, all sequences assigned to one development block were withheld, and the selected PISAS architecture was trained on the remaining four blocks. Predictor standardization and target-related weights were recomputed using only the retained training sequences. Each experiment used the fixed 250-epoch training budget and the three cross-validation seeds specified in Sect. 4, producing 15 regional neural checkpoints. The magnitude-only and regularized five-predictor linear benchmarks were refitted under the same block exclusions. This design ensures that every regional prediction is produced without training examples from the evaluated geographic block.

Productivity transferred moderately across the spatial partitions (Fig. 7d). When predictions from the five held-out regions were pooled, PISAS achieved an MAE of 0.227, an RMSE of 0.280, $R^2 = 0.447$, and $r = 0.677$. The magnitude-only model yielded an RMSE of 0.293 and $R^2 = 0.396$, whereas the regularized multi-predictor linear model yielded an RMSE of 0.282 and $R^2 = 0.441$. PISAS therefore attained 4.3 %RMSE skill relative to magnitude alone and slightly exceeded the 3.8 %skill of the regularized linear model. Its pooled bias was -0.025, indicating that geographic exclusion did not introduce a large systematic productivity offset.

The regional productivity results were heterogeneous. PISAS improved on the magnitude-only benchmark by 17.7 %in the Northwest Pacific, 1.3 % in Maritime Southeast Asia, and 0.9 %in the Southwest Pacific. It performed slightly worse than magnitude alone in South America and the Eurasia–Indian Ocean block, with skills of -2.1 % and -3.7 %, respectively. Regional PISAS RMSE ranged from 0.266 in Maritime Southeast Asia to 0.320 in the Northwest Pacific, where the magnitude-only RMSE was substantially larger at 0.389. The positive pooled skill is therefore not produced uniformly across all geographic domains; it reflects strong improvement in one block and smaller or neutral changes elsewhere. The broad bootstrap intervals associated with individual blocks also indicate that differences of only a few percentage points should not be interpreted as definitive regional superiority.

Regional transfer of $p$remained weak (Fig. 7e). Across all 143 leave-one-region-out predictions, PISAS yielded an MAE of 0.169, an RMSE of 0.216, $R^2 = -0.068$, and $r = 0.140$. The magnitude-only and regularized linear models produced RMSE values of 0.217 and 0.217, respectively. The resulting PISAS skill was only $+0.6$ %, while the regularized linear skill was $+0.2$ %. Thus, the five-predictor nonlinear representation provided essentially no pooled geographic-transfer advantage for the decay exponent.

Block-specific $p$performance varied in both direction and magnitude. PISAS improved on the magnitude-only benchmark in South America ($+16.6$ %), Maritime Southeast Asia ($+7.8$ %), and the Northwest Pacific ($+2.6$ %), but deteriorated in Eurasia–Indian Ocean (-15.0 %) and the Southwest Pacific (-6.7 %). Positive regional R^2 values were obtained only for South America ($R^2 = 0.233$) and, marginally, the Northwest Pacific ($R^2 = 0.043$); the remaining blocks had negative values. The opposing regional outcomes, together with the small sample sizes and overlapping bootstrap intervals, show that the apparent depth and background-rate controls on $p$are not spatially uniform.

The temporal and regional tests lead to a consistent distinction between the two surrogate targets. Productivity retains moderate predictive structure when models are transferred to later years or withheld geographic blocks,

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

and its pooled performance remains better than magnitude alone. The gain is modest and region dependent, but it is reproduced under both forms of distribution shift. By contrast, the decay exponent shows neither progressive temporal drift nor robust spatial transfer; its errors remain broadly stable in time, yet the model rarely explains substantial intersequence variance. These results support the use of PISAS as a constrained surrogate for first-order aftershock productivity and rate reconstruction, while indicating that predictions of p should be treated as strongly regularized estimates rather than precise sequence-specific forecasts.

8 Model interpretability

Figure 8 examines how the five input predictors influence the two PISAS outputs at global and sequence-specific scales. All interpretability analyses were performed after the architecture, training procedure, and temporal-test predictions had been frozen. None of these results contributed to hyperparameter selection or model fitting. I combined repeated permutation importance, accumulated local effects, checkpoint-level rank stability, and integrated gradients because these diagnostics address complementary questions: whether predictive accuracy depends on a predictor, how the fitted response changes across its observed range, whether its importance is stable across independently trained models, and how it contributes to individual predictions.
Global predictor dependence was quantified using repeated permutation importance. For predictor j, model checkpoint m, and permutation repeat b, I calculated

$$\Delta\mathrm{RMSE}_{j,m}^{(b)} = \mathrm{RMSE}\left[y, f_m\left(X_{\pi_j}^{(b)}\right)\right] - \mathrm{RMSE}[y, f_m(X)]\,, \qquad (36)$$

where $X_{\pi_j}^{(b)}$ denotes the evaluation matrix after randomly permuting only predictor j. Positive values indicate that disrupting the association between the predictor and the fitted model degrades prediction. This model-agnostic extension of permutation importance follows the general principle introduced for random forests by Breiman (2001). I used 200 independent permutations for each predictor and checkpoint. Rolling-origin importance was evaluated separately for the 12 fold–seed checkpoints on their corresponding validation blocks, whereas temporal-test importance was evaluated for the five final models. The reported intervals were obtained from 5000 bootstrap resamples of the checkpoint-level estimates.
Mainshock magnitude was the dominant predictor of $\log_{10} N_{30}^{qc}$ under both evaluation protocols (Fig. 8a). Permuting M increased RMSE by 0.207 out of fold and 0.199 on the temporal test, corresponding to relative increases of 70.8 %and 63.4 %, respectively. The 95 % intervals excluded zero, and magnitude had positive importance for every checkpoint. The remaining predictors produced substantially smaller changes. In out-of-fold evaluation, $\mu_{365,300}$, $\mu_{90,200}$, and $\mu_{30,100}$ increased RMSE by 0.053, 0.032, and 0.026, respectively, whereas depth produced an increase of only 0.002. On the temporal test, $\mu_{30,100}$ and $\mu_{90,200}$were the strongest secondary predictors, with increases of 0.040 and 0.034; depth contributed 0.011 and $\mu_{365,300}$ only 0.007. The reordering of the three background-rate scales indicates that their individual importance is sensitive to the evaluated time interval. Because these predictors are temporally and spatially nested, their scores should be interpreted collectively as multiscale background-state information rather than as independent physical controls.
Permutation importance for $p$was weaker by approximately an order of magnitude (Fig. 8b). Hypocentral depth ranked first under both protocols, increasing RMSE by 0.014 out of fold and 0.034 on the temporal test. Its importance was positive for every checkpoint, with mean ranks of 1.17 and 1.00, respectively. All other out-of-fold effects were close to zero: magnitude increased RMSE by 0.002, while the three background-rate effects ranged from approximately zero to 0.001 and had intervals that included zero. On the temporal test, $\mu_{30,100}$, magnitude, and $\mu_{365,300}$ produced increases of 0.016, 0.012, and 0.008, respectively, whereas $\mu_{90,200}$ remained negligible. These results support the interpretation that depth supplies the most stable input dependence for the decay exponent, but the absolute predictive information available for $p$remains limited.
I used first-order accumulated local effects to examine the signed nonlinear response of the fitted model while reducing extrapolation outside the observed multivariate predictor distribution. For a fitted response f, the population form of the first-order effect for predictor x_j is

$$\mathrm{ALE}_j(x) = \int_{x_{j,0}}^{x} \mathrm{E}\left[\frac{\partial f(X)}{\partial X_j}\middle| X_j = z\right] \mathrm{d}z - C_j, \qquad (37)$$

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

where C_j centres the effect to zero over the empirical distribution. In practice, I estimated local finite differences within 10 empirical intervals, using the complete set of 175 primary predictor vectors and each of the five final-model checkpoints. Curves in Fig. 8c–d show the checkpoint mean, and the envelopes represent seed-wise 95 % intervals. Accumulated local effects are preferable to ordinary partial-dependence curves when predictors are dependent because they avoid evaluating the model at many unrealistic combinations of predictor values (Apley and Zhu, 2020).

The accumulated effect of mainshock magnitude on productivity was strongly positive and nonlinear (Fig. 8c). Relative to the centred model response, the magnitude effect increased from approximately -0.40 at the lower end of the primary sample to about $+1.01$near its upper end. The response steepened for the largest mainshocks, consistent with the dominant permutation importance and the ablation results in Sect. 6. The effect of $\mu_{90,200}$ was also generally positive, increasing from approximately $-0.09$to $+0.39$, although uncertainty widened markedly in the sparsely populated upper tail. In contrast, the accumulated effect of $\mu_{30,100}$ was weakly negative over much of its observed range, declining to approximately -0.11. This negative signed response is not inconsistent with its positive permutation importance: permutation importance measures the magnitude of the model's dependence on a predictor, whereas accumulated local effects describe the direction and shape of that dependence conditional on the observed predictor distribution.

For p, hypocentral depth produced the clearest accumulated response (Fig. 8d). The effect decreased from approximately $+0.10$ for shallow mainshocks to -0.25 for the deepest events represented in the primary dataset. Thus, within the fitted model, greater depth generally shifted the predicted decay exponent towards lower values and hence slower temporal decay. The magnitude response was much smaller and non-monotonic, increasing slightly across the central magnitude range before declining for the largest events. The response to $\mu_{30,100}$was likewise weak and non-monotonic, with broad uncertainty at high rates. These curves characterize the behaviour learned by PISAS and should not be interpreted as direct causal effects, particularly where the predictor support is sparse or where the multiscale background rates are mutually dependent.

Checkpoint-level rankings further distinguish robust dependencies from unstable secondary effects (Fig. 8e). Magnitude had a mean productivity rank of 1.17 out of fold and 1.00 on the temporal test and increased RMSE for 100 % of checkpoints under both protocols. Depth had the corresponding mean rank of 1.17 and 1.00 for p, again with positive importance for every checkpoint. Rankings among the background-rate predictors varied more strongly. For productivity, $\mu_{365,300}$ ranked second out of fold but fifth on the temporal test, whereas $\mu_{30,100}$ moved from a mean rank of 3.50 to 2.40. For p, several secondary predictors increased RMSE in only 58–75 % of out-of-fold checkpoints, and $\mu_{90,200}$ had positive test importance for only 40 % of the five final models. The stable first-ranked predictors are therefore more defensible than the precise ordering of the weaker effects.

Local explanations were obtained using integrated gradients (Sundararajan et al., 2017). For input x, reference input x', and output f, the attribution to predictor j is

$$\mathrm{IG}_j(x) = \left(x_j - x'_j\right) \int_0^1 \frac{\partial f[x' + \alpha(x - x')]}{\partial x_j} \, \mathrm{d}\alpha. \qquad (38)$$

The reference was the development-sample predictor mean, equivalent to the zero vector in standardized coordinates for each final checkpoint. I evaluated Eq. (38) using 256 trapezoidal integration steps and averaged the attributions across the five final-model seeds. The integrated-gradient completeness property requires the sum of the predictor contributions to reproduce the difference between the prediction and its reference prediction. The maximum numerical completeness error was 4.98×10^{-6}, and the frozen Phase 2A predictions were reproduced to within 4.44×10^{-16}.

Figure 8f presents local explanations for three temporal-test sequences selected near the 10th, 50th, and 90th percentiles of ensemble-mean predicted productivity. The low-productivity sequence, us7000pwkn, had $M = 7.4$, $\widehat{N}_{30} = 36.9$, and $\hat{p} = 1.02$. Its productivity prediction combined small positive contributions from magnitude and $\mu_{30,100}$with negative contributions from depth, $\mu_{90,200}$, and particularly $\mu_{365,300}$. The median sequence, us6000rfwz, had $M = 7.4$, $\widehat{N}_{30} = 44.0$, and $\hat{p} = 0.68$; its largest productivity contribution was the negative depth attribution of -0.197, partly offset by positive contributions from magnitude and the long-term background rate. The high-productivity sequence, us6000rtdt, had $M = 7.6$, $\widehat{N}_{30} = 122.8$, and $\hat{p} = 0.81$. Its

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

elevated prediction was driven principally by magnitude (+0.199), $\mu_{90,200}$(+0.098), and $\mu_{365,300}$(+0.106), while depth contributed negatively.

The local p explanations were dominated by depth for the median- and high-productivity examples. Its contributions were +0.069, −0.292, and −0.155 for the low-, median-, and high-productivity sequences, respectively. Contributions from magnitude and the three background-rate scales were generally smaller, and several seed-wise 95 % intervals included zero, as denoted by daggers in Fig. 8f. The local results therefore reinforce the global diagnostics: magnitude provides the most consistent control on predicted productivity, while depth supplies the strongest input dependence for the decay exponent. At the same time, the changing signs and magnitudes of the background-rate attributions show that PISAS uses these variables conditionally rather than as fixed additive corrections.

Taken together, the interpretation analyses show that the neural surrogate has not replaced the established magnitude–productivity relation with an opaque alternative. Instead, it retains magnitude as the dominant global predictor and uses focal depth and multiscale background seismicity to introduce smaller, nonlinear, and sequence-dependent adjustments. The decay exponent relies most consistently on depth, but its weaker permutation effects, unstable secondary rankings, broad accumulated-effect uncertainty, and small local contributions are consistent with the limited predictive skill documented in Sects. 5–7. The interpretability results therefore support the physical plausibility of the learned productivity mapping while reinforcing the need for caution when assigning sequence-specific meaning to predicted values of p.

9 Discussion

PISAS v1.0 was designed to test whether a compact set of variables available at mainshock origin time can support physically admissible predictions of aftershock productivity and temporal decay at the global scale. The results reveal a clear asymmetry between these two objectives. The model captured a reproducible component of the variability in 30 d productivity, whereas its ability to discriminate among sequence-specific Omori–Utsu exponents remained limited. This distinction persisted across rolling-origin validation, the untouched 2020–2025 temporal test, leave-one-region-out evaluation, feature-ablation experiments, and post hoc interpretation. The contrasting outcomes are therefore unlikely to be artifacts of a single partition or metric and instead reflect differences in the information content of the adopted predictors and targets.

The productivity results confirm that mainshock magnitude supplies the dominant first-order control. Removing M increased the out-of-fold RMSE by 45.0 %, while permutation of M increased RMSE by 0.207 out of fold and 0.199 on the temporal test. These effects were much larger than those associated with depth or any individual background-rate predictor. This behaviour is consistent with the established magnitude dependence of earthquake triggering and aftershock productivity (Helmstetter et al., 2005). The dominant magnitude dependence and the weaker secondary contribution of depth are also consistent with a complementary global analysis based on fixed association windows and explicit tectonic classification (D'Alessandro, 2026b). Unlike that descriptive framework, PISAS learns a nonlinear mainshock-time mapping and reconstructs an exactly normalized Omori–Utsu rate history. At the same time, sequences of similar magnitude can differ substantially in productivity because rupture properties, available brittle volume, stress state, and tectonic setting modulate the number of triggered earthquakes (Dascher-Cousineau et al., 2020). PISAS appears to recover a limited part of this secondary variability through its nonlinear use of depth and multiscale background seismicity, but magnitude remains the principal source of predictable signal.

The improvement over the magnitude-only benchmark was statistically consistent but modest. PISAS reduced productivity RMSE by 3.9 % in pooled rolling-origin validation, 5.0 % on the temporal test, and 4.3 % in pooled leave-one-region-out evaluation. These gains show that the additional predictors contain transferable information, but they also impose an important constraint on interpretation: PISAS should not be presented as replacing the magnitude–productivity relation. It provides a nonlinear correction to that relation. The regularized linear model produced little or no out-of-fold improvement, indicating that the added information was not well represented by a single global additive dependence. Nevertheless, the limited difference between PISAS and the linear benchmarks also shows that model complexity must remain commensurate with the small sequence-level sample.

The three background-rate predictors contributed less strongly and less consistently than magnitude. Their individual rankings changed between rolling-origin and temporal-test evaluation, and removing any one of them produced only a small degradation because the three variables describe nested temporal and spatial windows. Finite event counts also impose fundamental limits on the resolution of seismic-rate changes, so

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

weak or unstable rate-based signals do not necessarily imply the absence of underlying temporal variability (D'Alessandro, 2026a). Because the three PISAS background predictors are estimated from finite counts within fixed space–time windows, part of their variable importance across evaluation blocks may reflect sampling uncertainty in the predictors themselves rather than exclusively changes in their physical relevance. Their shared information cannot therefore be partitioned uniquely among $\mu_{30,100}$, $\mu_{90,200}$, and $\mu_{365,300}$. The positive permutation importance of a background rate also does not imply that its fitted effect must be positive everywhere. For example, $\mu_{30,100}$ had a weakly negative accumulated local effect across much of its observed range while still improving predictive accuracy when left unpermuted. Permutation importance quantifies dependence of the fitted prediction on a variable, whereas accumulated local effects describe the signed conditional response. ALE curves are specifically intended to reduce unrealistic extrapolation when predictors are correlated, but they do not transform associational model behaviour into causal relationships (Apley and Zhu, 2020).

The productivity residuals nevertheless retained substantial unexplained variability. Regional observations also indicate that aftershock productivity can be related to the degree of mechanical coupling along a subduction interface, illustrating that hypocentral depth and location may act as proxies for tectonic conditions not represented explicitly by the present predictor set (Hainzl et al., 2019). The model tended to contract predictions towards the centre of the observed distribution, underpredicting some highly productive sequences and overpredicting some low-productivity sequences. Several physically relevant controls were not represented in the five-variable predictor vector. These include rupture length and area, focal mechanism, stress drop, finite-fault geometry, regional seismogenic thickness, heat flow, plate-boundary type, and spatially resolved pre-mainshock seismicity. The global analysis necessarily trades local physical detail for predictor availability and reproducibility. The strong productivity dependence on the volume of brittle lithosphere inferred by Dascher-Cousineau et al. (2020), together with evidence that triggering potential can vary with depth, suggests that future versions could replace hypocentral depth with more physically explicit measures of the mainshock's position within the local seismogenic layer (Asayesh et al., 2023).

The weak prediction of $p$is an equally important result. Pooled out-of-fold $R^2$was negative, temporal-test R^2 was only 0.105, and leave-one-region-out R^2 was again negative. Improvements over the magnitude-only benchmark were small and changed sign among temporal and regional blocks. Hypocentral depth was the most stable predictor according to permutation importance and ablation, but its effect was insufficient to explain a large fraction of the observed intersequence variability. The depth dependence learned by PISAS is therefore better interpreted as a regularizing association than as evidence that depth directly controls the Omori–Utsu exponent. Depth may act as a proxy for temperature, rheology, tectonic environment, rupture dimensions, or catalogue performance, and the present analysis cannot separate these alternatives.

Part of the apparent unpredictability of p is also attributable to the target itself. In contrast to N_{30}^{qc}, which is a directly counted quantity under the adopted selection rule, $p$is estimated from finite and incomplete event times. Rate-dependent incompleteness immediately after large earthquakes can bias the apparent triggering function and the calibration of epidemic-type models, while magnitude truncation and missing events can affect fitted point-process parameters (Hainzl, 2016; Harte, 2016; Mizrahi et al., 2021). The fixed lower time limit of 0.1 d reduces the most severe short-term incompleteness but also removes the interval with the highest rate and potentially the greatest information about early decay. The $N_{\mathrm{fit}} \geq 20$ threshold improves estimator stability, as demonstrated by the progressive deterioration for sequences containing 10–19 events, but it cannot eliminate uncertainty arising from finite sample size or heterogeneous detection. Such transient post-mainshock increases in the detection threshold constitute a distinct short-term aftershock incompleteness problem and require explicit, reproducible treatment when estimating catalogue completeness and sequence parameters (Figlioli et al., 2024).

The restricted 0.1–30 d interval and fixed $c = 0.01$ d also define the meaning of the estimated p. PISAS does not predict a universal, observation-independent decay exponent; it predicts the exponent of a specific modified Omori–Utsu fit under fixed temporal, magnitude, and assignment rules. Alternative lower cutoffs, sequence durations, completeness corrections, or decay functions could produce different estimates. This conditional definition is necessary for reproducibility, but it limits comparison with studies that jointly estimate c, use dynamically varying completeness thresholds, or account explicitly for secondary triggering. The relatively narrow predicted range may therefore combine genuine regularization towards a global mean with uncertainty and heterogeneity in the fitted targets.

The physical decoder constitutes the principal methodological contribution of the model. PISAS did not improve accuracy by relaxing scientific constraints. Instead, the hard-coded normalization matched or

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

outperformed an otherwise comparable unconstrained network while guaranteeing positive productivity, bounded p, positive monotonically decreasing rates, and exact agreement between the 30 d rate integral and predicted productivity. The unconstrained model produced closure violations for every evaluated sequence, even when its individual outputs appeared plausible. This result illustrates why independently predicting correlated physical quantities can be problematic: statistical accuracy of each scalar output does not ensure consistency of the reconstructed process. Encoding the closure analytically reduces output redundancy and eliminates an entire class of inadmissible solutions. This strategy is consistent with theory-guided and physics-guided machine learning, in which scientific knowledge is integrated into model structure to improve consistency and restrict extrapolation (Karpatne et al., 2017; Daw et al., 2020).

The constraints imposed here are empirical rather than fundamental physical laws. The modified Omori–Utsu relation is a phenomenological representation of aftershock decay, and monotonic decay over 0.1–30 d does not describe every observed sequence. Large secondary earthquakes, delayed triggering, volcanic or fluid-driven activity, and complex multi-stage sequences can generate rate increases or departures from a single power law. PISAS addresses part of this problem by excluding sequences interrupted by a larger event during the target interval, but this selection also narrows the range of seismic behaviour represented in the primary dataset. The model should therefore be interpreted as a surrogate for isolated, quality-controlled aftershock sequences under the adopted construction rules, not as a universal model of clustered seismicity.

Temporal and regional validation provide evidence of transferability but do not constitute fully prospective validation, a distinction that is central to the development and evaluation of operational earthquake-forecasting systems (Mizrahi et al., 2024). The 2020–2025 test was chronologically isolated before training and model selection, yet both its catalogue content and data-processing environment were known retrospectively when the study was conducted. Similarly, the leave-one-region-out experiment tests spatial extrapolation among broad global blocks but does not guarantee performance in a new regional network with different magnitude calibration or completeness. Structured validation is more informative than random splitting when data are temporally and spatially dependent, because random partitions can place closely related observations in both training and evaluation sets (Roberts et al., 2017). The observed positive productivity skill under both temporal and regional blocking is therefore encouraging, but external validation on independently maintained regional catalogues remains necessary.

Regional performance heterogeneity further indicates that a single global mapping cannot capture every tectonic environment equally well. Productivity skill was strongly positive in the Northwest Pacific but neutral or negative in some other blocks. The broad geographic clusters contain several tectonic regimes and should not be interpreted as physically homogeneous regions. Spatially coherent variations in earthquake-sequence productivity have also been documented across California and Nevada and associated with differences in regional seismotectonic conditions (Trugman and Ben-Zion, 2023). Their purpose was to impose geographic separation, not to estimate region-specific model parameters. Future extensions could use hierarchical or mixture-of-experts formulations in which a global representation is supplemented by tectonic-domain effects. Such models would require careful regularization because the current sample contains only 175 primary sequences, and finer regional subdivision would rapidly reduce the number of training examples per regime.

The interpretation diagnostics support the internal plausibility of the fitted mapping but do not establish physical causation. Permutation importance measures the dependence of predictive error on a feature under a particular data distribution. ALE curves summarize the fitted conditional response over the empirical predictor support. Integrated gradients decompose an individual prediction relative to a chosen baseline and satisfy a completeness relation, but the attributions remain model- and baseline-dependent (Sundararajan et al., 2017). Their agreement in identifying magnitude for productivity and depth for $p$increases confidence that these dependencies are not artifacts of one interpretation method. Conversely, the changing rankings and signs of weaker background-rate effects indicate that those contributions should not be assigned a universal physical meaning.

PISAS is also distinct from an epidemic-type aftershock-sequence model. ETAS represents earthquake occurrence as a self-exciting point process in which every event can trigger subsequent events and can generate forecasts in time, space, and magnitude (Ogata, 1998). PISAS instead predicts two sequence-level summaries from information available at the mainshock origin time and reconstructs a deterministic mean temporal-rate curve. It does not model aftershock locations, magnitudes, generation structure, or stochastic event-to-event variability. PISAS also differs from neural temporal point-process models such as RECAST, which learn the conditional occurrence process and can generate stochastic earthquake sequences (Dascher-Cousineau et al., 2023). PISAS instead predicts sequence-level descriptors and reconstructs a deterministic conditional mean

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

rate constrained by an analytical decay law. Consequently, the present implementation is not a complete operational earthquake-forecasting system. Operational models generally require probabilistic occurrence distributions, explicit treatment of secondary triggering, and propagation of parameter and model uncertainty (Field et al., 2017).

The most appropriate near-term role for PISAS is therefore as a rapid initial-condition surrogate. Immediately after a large mainshock, it can provide a physically self-consistent prior estimate of 30 d productivity, decay, and mean rate before a sufficiently informative aftershock sequence has accumulated. This initial estimate could subsequently be updated using early observations within an ETAS, Bayesian, or data-assimilation framework. Methods that explicitly account for early catalogue incompleteness have shown that limited initial aftershock observations can already support the updating of point-process parameters and short-term event-rate forecasts (Omi et al., 2014; Hainzl et al., 2024). Combining such dynamic updating with the global prior learned by PISAS would preserve the advantage of immediate prediction while allowing the forecast to adapt to the realized sequence. This role is complementary to recent ETAS developments that use mainshock information to improve early sequence-specific forecasts, but PISAS differs by learning a global mainshock-time prior and imposing exact analytical normalization of the reconstructed rate (Asayesh et al., 2025).

Several developments are required before such use. First, predictive uncertainty must extend beyond between-seed dispersion to include target-estimation error, catalogue uncertainty, model uncertainty, and intrinsic count variability. Second, productivity should be represented probabilistically rather than only through its conditional mean. Third, external catalogues with homogeneous moment magnitudes and lower completeness thresholds should be used to test magnitude-scale sensitivity and regional transfer. Fourth, richer source and tectonic predictors should be evaluated without compromising availability at mainshock time. Finally, the sequence-construction rules, magnitude-dependent association windows, lower time cutoff, and target duration should be examined through explicit sensitivity experiments or alternative catalogue realizations.

Within these limits, PISAS demonstrates that a small neural surrogate can add reproducible information beyond a magnitude-only productivity law while retaining exact consistency with an analytical aftershock-decay model. Its main scientific result is not that aftershock evolution has become accurately predictable from five global predictors. Rather, the results delimit what is currently predictable: cumulative productivity contains a moderate transferable signal, whereas the sequence-specific decay exponent remains weakly constrained. The physics-informed architecture ensures that this limited statistical information is converted into an admissible rate history rather than an internally inconsistent collection of parameters. That combination of modest predictive improvement, explicit physical closure, structured validation, and transparent interpretation defines the appropriate scope of PISAS v1.0.

10 Conclusions

PISAS v1.0 provides a physics-informed neural surrogate for predicting two complementary descriptors of global aftershock sequences from information available at the mainshock origin time: the quality-controlled 30 d productivity, N_{30}^{qc}, and the Omori–Utsu decay exponent, p. The model combines a compact five-predictor neural representation with an analytical decoder that reconstructs the productivity coefficient and the complete mean aftershock-rate history. This formulation guarantees positive productivity, a bounded decay exponent, positive and monotonically decreasing rates, and exact agreement between the predicted productivity and the integral of the reconstructed rate over the target interval.

The global dataset was constructed from the U.S. Geological Survey ComCat catalogue for 2000–2025 using reproducible magnitude-dependent association windows, unique event assignment, exclusion of sequences interrupted by a larger earthquake, and explicit quality-control criteria. The primary dataset contains 175 independent sequences with at least 20 fitted aftershocks, while an additional 101 lower-information sequences were retained exclusively for sensitivity analysis. Model development used temporally ordered rolling-origin validation, and the final evaluation was performed on an untouched 2020–2025 test set.

The results establish a clear difference between productivity and decay predictability. PISAS reproduced a moderate and transferable component of the variability in $\log_{10} N_{30}^{qc}$, with $R^2 = 0.454$ in pooled rolling-origin validation and $R^2 = 0.534$ on the untouched temporal test. Relative to a magnitude-only linear model, productivity RMSE improved by 3.9 % out of fold, 5.0 % on the temporal test, and 4.3 % under leave-one-region-out validation. These gains are modest but consistent across independent forms of distribution shift, indicating that focal depth and multiscale pre-mainshock seismicity contain additional information beyond the dominant magnitude–productivity scaling.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

Prediction of the Omori–Utsu exponent remained substantially weaker. Out-of-fold R^2 was negative, temporal-test R^2 was only 0.105, and pooled regional-transfer R^2 was again negative. Hypocentral depth was the most stable predictor of p, but its contribution was insufficient to explain much of the observed intersequence variability. The weak performance likely reflects both limited physical information in the adopted predictor set and uncertainty in estimating $p$from finite, heterogeneous, and incompletely observed aftershock sequences. PISAS estimates of p should therefore be interpreted as regularized sequence-level predictions rather than precise forecasts of decay dynamics.

The physics-informed decoder improved reliability without reducing predictive skill. Compared with an otherwise equivalent unconstrained neural network, PISAS matched or improved performance while eliminating all violations of productivity closure and output admissibility. The unconstrained model generated rate histories whose integrals disagreed with its independently predicted productivity for every evaluated sequence. This comparison demonstrates that independently accurate scalar outputs do not necessarily define a coherent physical process and that analytical constraints can remove inadmissible solutions without requiring additional training data or penalty terms.

Interpretability analyses confirmed that mainshock magnitude is the dominant and most stable predictor of productivity. Background seismicity supplied smaller, nonlinear, and sequence-dependent corrections, whereas focal depth contributed most consistently to the prediction of p. The changing importance and attribution of the background-rate variables indicate that their role is conditional and partly redundant across the three nested space–time scales. These dependencies characterize the fitted surrogate and should not be interpreted as direct causal relationships.

PISAS v1.0 is not a complete operational earthquake-forecasting model. It does not represent aftershock locations, magnitudes, secondary triggering, stochastic event counts, or the full propagation of catalogue and model uncertainty. Its appropriate present role is as an immediate, physically consistent prior for the expected productivity and mean temporal decay of isolated aftershock sequences. Future developments should incorporate probabilistic count distributions, calibrated predictive uncertainty, richer rupture and tectonic descriptors, homogeneous regional catalogues, and sequential updating as early aftershock observations become available.

The principal contribution of PISAS is therefore not a claim of complete predictability. Rather, it defines a reproducible boundary between the aspects of global aftershock behaviour that can presently be inferred from a small set of mainshock-time predictors and those that remain poorly constrained. Cumulative 30 d productivity contains a moderate transferable signal, whereas sequence-specific temporal decay does not. By embedding this limited statistical information within an exactly normalized analytical rate model, PISAS converts neural predictions into physically admissible and interpretable aftershock-rate histories.

Code and data availability. The original earthquake data were obtained from the U.S. Geological Survey Advanced National Seismic System Comprehensive Earthquake Catalog (ComCat; U.S. Geological Survey, 2017; https://doi.org/10.5066/F7MS3QZH), which is publicly accessible through the USGS Earthquake Hazards Program. The catalogue query covered the period from 1 January 1999 to 30 January 2026 and included earthquakes with reported magnitude $M \geq 4.0$; the primary analyses were restricted to mainshocks occurring between 2000 and 2025. Because ComCat is continuously revised, the exact catalogue snapshot used in this study, together with the complete query parameters and retrieval metadata, is preserved in the study archive.

The frozen PISAS v1.0 reproducibility package is available from Zenodo at [10.5281/zenodo.21370942] (D'Alessandro, 2026c). The archive contains the complete source code for catalogue acquisition and quality control, annual completeness analysis, mainshock-dependence classification, aftershock-sequence construction, modified Omori–Utsu fitting, predictor generation, model training, benchmark evaluation, physics-ablation experiments, temporal and regional validation, model-interpretation analyses, and reproduction of Figs. 1–8. It also includes the quality-controlled catalogue snapshot, the primary and sensitivity sequence tables, the temporal and regional data partitions, configuration files, random seeds, fitted preprocessing parameters, trained neural-network checkpoints, numerical results underlying the figures, and the software-environment specifications required to reproduce the analyses.

The archived source code is released under the MIT License. Derived tabular data and figure-source data are released under the Creative Commons Attribution 4.0 International License. The original USGS catalogue records retain their applicable U.S. Geological Survey terms of use. The Zenodo release identified above is the version of record corresponding exactly to PISAS v1.0 and to the results reported in this manuscript.

Author contribution. AD conceived and designed the study, developed the PISAS v1.0 methodology and software, acquired and curated the data, performed the analyses, validated the results, produced the figures, interpreted the findings, and wrote and revised the manuscript.

Competing interests. The author declares that there is no conflict of interest.

Acknowledgements. The author acknowledges the U.S. Geological Survey Earthquake Hazards Program for maintaining the Advanced National Seismic System Comprehensive Earthquake Catalog and providing public access to the earthquake data used in this study.

Financial support. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Apley, D. W. and Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models, J. R. Stat. Soc. B, 82, 1059–1086, https://doi.org/10.1111/rssb.12377, 2020.

Asayesh, B. M., Hainzl, S., and Zöller, G.: Improved aftershock forecasts using mainshock information in the framework of the ETAS model, J. Geophys. Res.-Sol. Ea., 130, e2024JB030287, doi:10.1029/2024JB030287, 2025.

Asayesh, B. M., Hainzl, S., and Zöller, G.: Depth-dependent aftershock trigger potential revealed by 3D-ETAS modeling, J. Geophys. Res.-Sol. Ea., 128, e2023JB026377, https://doi.org/10.1029/2023JB026377, 2023.

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S.: Clustering on the unit hypersphere using von Mises–Fisher distributions, J. Mach. Learn. Res., 6, 1345–1382, available at: https://www.jmlr.org/papers/v6/banerjee05a.html (last access: 14 July 2026), 2005.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

D'Alessandro, A.: Detectability limits of seismic rate changes: statistical constraints from synthetic and global earthquake catalogs, Seismol. Res. Lett., https://doi.org/10.1785/0220260074, 2026a.

D'Alessandro, A.: Global controls on aftershock productivity: roles of magnitude and tectonic setting, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2026-877, 2026b.

D'Alessandro, A.: PISAS v1.0: source code, trained models, processed data, and figure reproduction package, Zenodo [code and data set], https://doi.org/10.5281/zenodo.21370942, 2026c.

Dascher-Cousineau, K., Brodsky, E. E., Lay, T., and Goebel, T. H. W.: What controls variations in aftershock productivity?, J. Geophys. Res.-Sol. Ea., 125, e2019JB018111, https://doi.org/10.1029/2019JB018111, 2020.

Dascher-Cousineau, K., Shchur, O., Brodsky, E. E., and Günnemann, S.: Using deep learning for flexible and scalable earthquake forecasting, Geophys. Res. Lett., 50, e2023GL103909, doi:10.1029/2023GL103909, 2023.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

Daw, A., Thomas, R. Q., Carey, C. C., Read, J. S., Appling, A. P., and Karpatne, A.: Physics-Guided Architecture (PGA) of neural networks for quantifying uncertainty in lake temperature modeling, in: Proceedings of the 2020 SIAM International Conference on Data Mining, 532–540, https://doi.org/10.1137/1.9781611976236.60, 2020.

de Arcangelis, L., Godano, C., and Lippiello, E.: The overlap of aftershock coda waves and short-term postseismic forecasting, J. Geophys. Res.-Sol. Ea., 123, 5661–5674, https://doi.org/10.1029/2018JB015518, 2018.

DeVries, P. M. R., Viégas, F., Wattenberg, M., and Meade, B. J.: Deep learning of aftershock patterns following large earthquakes, Nature, 560, 632–634, https://doi.org/10.1038/s41586-018-0438-y, 2018.

Figlioli, A., Vitale, G., Taroni, M., and D'Alessandro, A.: Tremors—A software app for the analysis of the completeness magnitude, Geosciences, 14, 149, https://doi.org/10.3390/geosciences14060149, 2024.

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., and Werner, M. J.: A spatiotemporal clustering model for the Third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): Toward an operational earthquake forecast, Bull. Seismol. Soc. Am., 107, 1049–1081, https://doi.org/10.1785/0120160173, 2017.

Gardner, J. K. and Knopoff, L.: Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?, Bull. Seismol. Soc. Am., 64, 1363–1367, https://doi.org/10.1785/BSSA0640051363, 1974.

Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, edited by: Teh, Y. W. and Titterington, M., Proc. Mach. Learn. Res., 9, 249–256, available at: https://proceedings.mlr.press/v9/glorot10a.html (last access: 14 July 2026), 2010.

Guastella, M., Figlioli, A., Martorana, R., and D'Alessandro, A.: Quantifying the impact of declustering techniques on completeness magnitude estimation, Bull. Seismol. Soc. Am., https://doi.org/10.1785/0120250144, 2026.

Hainzl, S.: Apparent triggering function of aftershocks resulting from rate-dependent incompleteness of earthquake catalogs, J. Geophys. Res.-Sol. Ea., 121, 6499–6509, https://doi.org/10.1002/2016JB013319, 2016.

Hainzl, S., Sippl, C., and Schurr, B.: Linear relationship between aftershock productivity and seismic coupling in the Northern Chile subduction zone, J. Geophys. Res.-Sol. Ea., 124, 8726–8738, doi:10.1029/2019JB017764, 2019

Hainzl, S., Kumazawa, T., and Ogata, Y.: Aftershock forecasts based on incomplete earthquake catalogues: ETASI model application to the 2023 SE Türkiye earthquake sequence, Geophys. J. Int., 236, 1609–1620, doi:10.1093/gji/ggae006, 2024.

Hardebeck, J. L., Llenos, A. L., Michael, A. J., Page, M. T., Schneider, M., and van der Elst, N. J.: Aftershock forecasting, Annu. Rev. Earth Pl. Sc., 52, 61–84, doi:10.1146/annurev-earth-040522-102129, 2024.

Harte, D. S.: Model parameter estimation bias induced by earthquake magnitude cut-off, Geophys. J. Int., 204, 1266–1287, https://doi.org/10.1093/gji/ggv524, 2016.

Helmstetter, A., Kagan, Y. Y., and Jackson, D. D.: Importance of small earthquakes for stress transfers and earthquake triggering, J. Geophys. Res.-Sol. Ea., 110, B05S08, https://doi.org/10.1029/2004JB003286, 2005.

Hendrycks, D. and Gimpel, K.: Gaussian error linear units (GELUs), arXiv, https://doi.org/10.48550/arXiv.1606.08415, 2016.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12, 55–67, https://doi.org/10.1080/00401706.1970.10488634, 1970.

Huber, P. J.: Robust estimation of a location parameter, Ann. Math. Stat., 35, 73–101, https://doi.org/10.1214/aoms/1177703732, 1964.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed machine learning, Nat. Rev. Phys., 3, 422–440, doi:10.1038/s42254-021-00314-5, 2021.

Jordan, T. H., Chen, Y.-T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., Papadopoulos, G., Sobolev, G., Yamaoka, K., and Zschau, J.: Operational earthquake forecasting: State of knowledge and guidelines for utilization, Ann. Geophys., 54, 315–391, https://doi.org/10.4401/ag-5350, 2011.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A. R., Shekhar, S., Samatova, N., and Kumar, V.: Theory-guided data science: A new paradigm for scientific discovery from data, IEEE T. Knowl. Data En., 29, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168, 2017.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, San Diego, USA, 7–9 May 2015, arXiv [preprint], https://doi.org/10.48550/arXiv.1412.6980, 2015.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, New Orleans, USA, 6–9 May 2019, arXiv [preprint], https://doi.org/10.48550/arXiv.1711.05101, 2019.

Marsan, D. and Helmstetter, A.: How variable is the number of triggered aftershocks?, J. Geophys. Res.-Sol. Ea., 122, 5544–5560, doi:10.1002/2016JB013807, 2017.

Mignan, A. and Broccardo, M.: One neuron versus deep learning in aftershock prediction, Nature, 574, E1–E3, https://doi.org/10.1038/s41586-019-1582-8, 2019.

Mizrahi, L., Nandan, S., and Wiemer, S.: Embracing data incompleteness for better earthquake forecasting, J. Geophys. Res.-Sol. Ea., 126, e2021JB022379, doi:10.1029/2021JB022379, 2021.

Mizrahi, L., Dallo, I., van der Elst, N. J., Christophersen, A., Spassiani, I., Werner, M. J., Iturrieta, P. C., Bayona, J., Iervolino, I., Schneider, M., Page, M. T., Zhuang, J., Herrmann, M., Michael, A. J., Falcone, G., Marzocchi, W., Rhoades, D., Gerstenberger, M., Gulia, L., Schorlemmer, D., Becker, J., Han, M., Kuratle, L., Marti, M., and Wiemer, S.: Developing, testing, and communicating earthquake forecasts: Current practices and future directions, Rev. Geophys., 62, e2023RG000823, doi:10.1029/2023RG000823, 2024.

Ogata, Y.: Estimation of the parameters in the modified Omori formula for aftershock frequencies by the maximum likelihood procedure, J. Phys. Earth, 31, 115–124, https://doi.org/10.4294/jpe1952.31.115, 1983.

Ogata, Y.: Space-time point-process models for earthquake occurrences, Ann. Inst. Stat. Math., 50, 379–402, https://doi.org/10.1023/A:1003403601725, 1998.

Omi, T., Ogata, Y., Hirata, Y., and Aihara, K.: Estimating the ETAS model from an early aftershock sequence, Geophys. Res. Lett., 41, 850–857, https://doi.org/10.1002/2013GL058958, 2014.

Page, M. T., van der Elst, N., Hardebeck, J. L., Felzer, K. R., and Michael, A. J.: Three ingredients for improved global aftershock forecasts: Tectonic region, time-dependent catalog incompleteness, and intersequence variability, Bull. Seismol. Soc. Am., 106, 2290–2301, https://doi.org/10.1785/0120160073, 2016.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

Page, M. T. and van der Elst, N. J.: Aftershocks preferentially occur in previously active areas, Seism. Rec., 2, 100–106, doi:10.1785/0320220005, 2022.

Raissi, M., Perdikaris, P., and Karniadakis, G. E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys., 378, 686–707, https://doi.org/10.1016/j.jcp.2018.10.045, 2019.

Reasenberg, P.: Second-order moment of central California seismicity, 1969–1982, J. Geophys. Res., 90, 5479–5495, https://doi.org/10.1029/JB090iB07p05479, 1985.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, available at: https://www.jmlr.org/papers/v15/srivastava14a.html, 2014.

Stockman, S., Lawson, D. J., and Werner, M. J.: Forecasting the 2016–2017 Central Apennines earthquake sequence with a neural point process, Earth's Future, 11, e2023EF003777, https://doi.org/10.1029/2023EF003777, 2023.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning, edited by: Precup, D. and Teh, Y. W., Proc. Mach. Learn. Res., 70, 3319–3328, available at: https://proceedings.mlr.press/v70/sundararajan17a.html, 2017.

Tashman, L. J.: Out-of-sample tests of forecasting accuracy: An analysis and review, Int. J. Forecasting, 16, 437–450, https://doi.org/10.1016/S0169-2070(00)00065-0, 2000.

Trugman, D. T. and Ben-Zion, Y.: Coherent spatial variations in the productivity of earthquake sequences in California and Nevada, Seism. Rec., 3, 322–331, doi:10.1785/0320230039, 2023.

U.S. Geological Survey, Earthquake Hazards Program: Advanced National Seismic System (ANSS) Comprehensive Catalog of Earthquake Events and Products, U.S. Geological Survey [data set], https://doi.org/10.5066/F7MS3QZH, 2017.

Utsu, T., Ogata, Y., and Matsu'ura, R. S.: The centenary of the Omori formula for a decay law of aftershock activity, J. Phys. Earth, 43, 1–33, https://doi.org/10.4294/jpe1952.43.1, 1995.

Wetzler, N., Brodsky, E. E., and Lay, T.: Regional and stress drop effects on aftershock productivity of large megathrust earthquakes, Geophys. Res. Lett., 43, 12012–12020, doi:10.1002/2016GL071104, 2016.

Wiemer, S. and Wyss, M.: Minimum magnitude of completeness in earthquake catalogs: Examples from Alaska, the western United States, and Japan, Bull. Seismol. Soc. Am., 90, 859–869, https://doi.org/10.1785/0119990114, 2000.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

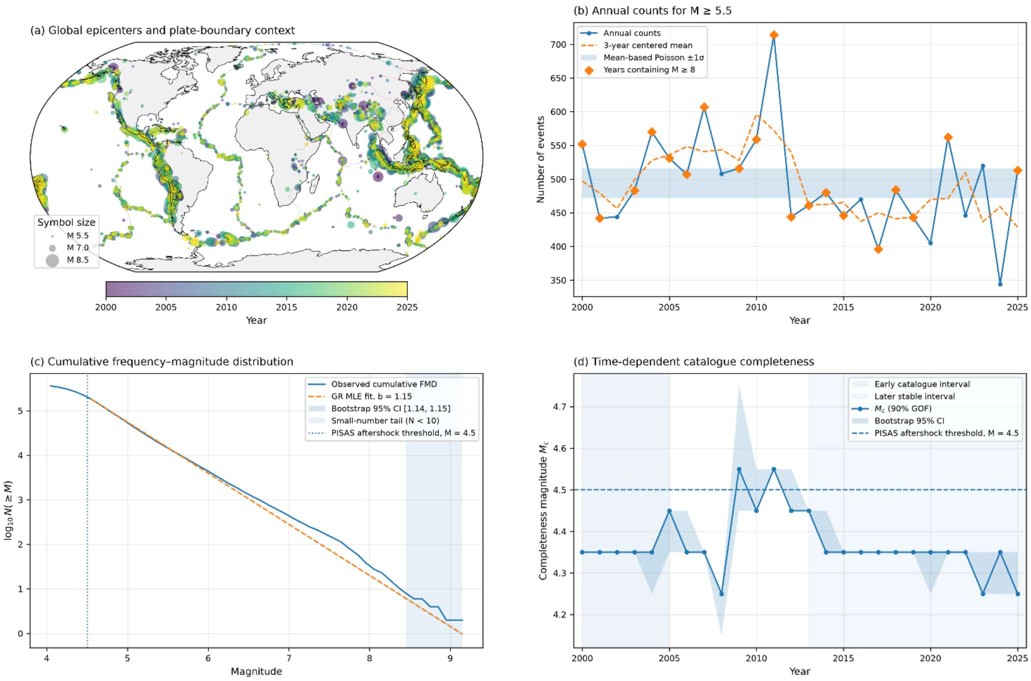

Figure 1. *Global coverage and detectability of the ComCat catalogue used to construct PISAS v1.1. The catalogue was downloaded for* $M \geq 4.0$ *from 1 January 1999 to 30 January 2026 to provide the pre-analysis warm-up and post-analysis buffer required by the sequence-building procedure, whereas all diagnostics shown here refer to the main analysis interval from 1 January 2000 to 31 December 2025. (a) Global distribution of the 12.847 earthquakes with* $M \geq 5.5$*, plotted in Robinson projection and superimposed on the PB2002 plate boundaries. Symbols are coloured by occurrence year and scaled by magnitude; events with hypocentral depth of at least 70 km are outlined. (b) Annual numbers of earthquakes with* $M \geq 5.5$*, together with the centred 3-year running mean and a mean-based Poisson* $\pm 1\sigma$ *reference interval. Diamonds identify years containing at least one event with* $M \geq 8.0$*. (c) Cumulative frequency–magnitude distribution of the 358 673 catalogue events with* $M \geq 4.0$*. The Gutenberg–Richter relation is fitted by maximum likelihood above the PISAS aftershock threshold* $M = 4.5$*, yielding* $b = 1.15$ *with a 95% bootstrap interval of 1.14–1.15. The vertical dotted line marks the lower magnitude threshold used for aftershock extraction, whereas the shaded high-magnitude interval denotes the cumulative tail containing fewer than 10 events. (d) Annual completeness magnitude* M_c*, estimated using the 90% goodness-of-fit criterion, with 95% bootstrap uncertainty. The horizontal dashed line marks the PISAS aftershock threshold* $M = 4.5$*. Annual* M_c *is generally below this threshold, although values approach or slightly exceed it during part of the 2009–2012 interval, indicating locally reduced detectability that is considered in the interpretation of early aftershock productivity. The background shading distinguishes the early catalogue interval from the later visually stable detectability regime.*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

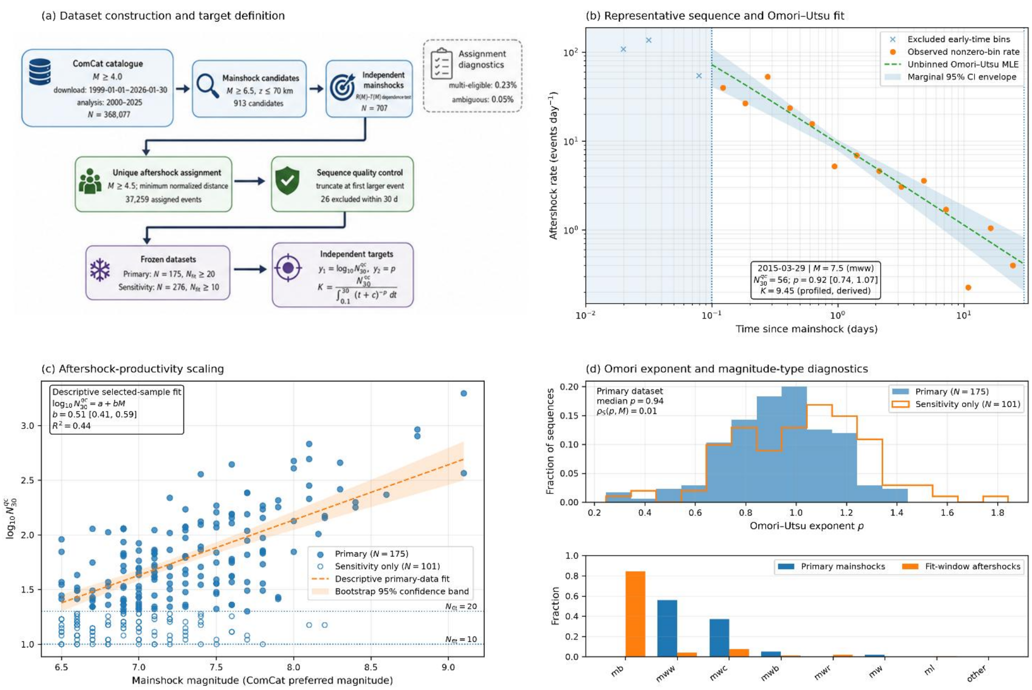

Figure 2. *Construction and characterization of the PISAS v1.1 aftershock-sequence datasets. (a) The USGS Comprehensive Earthquake Catalog (ComCat) was downloaded for $M \geq 4.0$ from 1 January 1999 to 30 January 2026, providing a 365 d pre-analysis warm-up and a 30 d post-analysis buffer, whereas mainshocks were analysed from 1 January 2000 to 31 December 2025. Among 913 candidate mainshocks with $M \geq 6.5$ and depth $z \leq 70$ km, 707 were classified as independent using magnitude-dependent spatial and temporal windows, $R(M)$ and $T(M)$. A total of 37,259 aftershocks with $M \geq 4.5$ were uniquely assigned by minimizing the normalized space–time distance, and associations were truncated at the first subsequent larger event; 26 sequences affected by a larger earthquake within 30 d were excluded. The resulting primary and sensitivity datasets contain 175 sequences with $N_{fit} \geq 20$ and 276 sequences with $N_{fit} \geq 10$, respectively. The independent surrogate targets are $log_{10}N_{30}^{qc}$ and the Omori–Utsu exponent p, whereas $K$is obtained analytically from N_{30}^{qc} and p. (b) Log-binned aftershock rates for a deterministically selected representative primary sequence ($M = 7.5$, mww, $N_{30}^{qc} = 56$) are compared with the unbinned Omori–Utsu maximum-likelihood fit, yielding $p = 0.92$ with a 95% bootstrap interval of 0.74–1.07. The shaded region is an envelope constructed from the marginal bootstrap intervals of $p$and the profiled K, and therefore does not represent a joint confidence band. Events occurring before $t = 0.1$ d are shown for reference but excluded from target construction and parameter estimation. (c) Quality-controlled 30 d productivity is plotted against ComCat preferred mainshock magnitude. Filled symbols denote the primary dataset and open symbols the 101 additional sensitivity-only sequences; the horizontal dotted lines mark the selection thresholds $N_{fit} = 20$ and $N_{fit} = 10$. The displayed regression is a descriptive fit to the response-truncated primary sample, with slope $b = 0.51$, 95% bootstrap interval 0.41–0.59, and $R^2 = 0.44$, and should not be interpreted as a truncation-corrected physical productivity law. (d) Common-bin histograms show the fraction of primary and sensitivity-only sequences as a function of p, while the lower panel compares the magnitude-type composition of primary mainshocks and aftershocks used in the 0.1–30 d fit window. Aftershock magnitudes are dominated by m_b, which is the prevailing fit-window magnitude type in 168 of the 175 primary sequences.*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

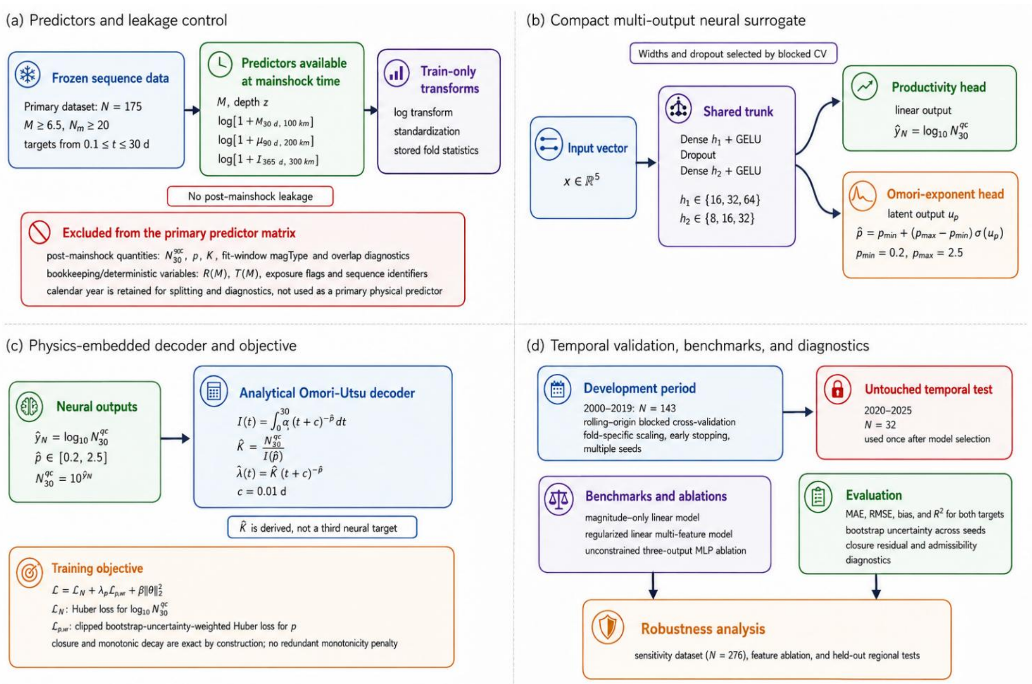

Figure 3. *Architecture and training workflow of the physics-informed neural surrogate (PISAS v1.0). (a) Predictor definition and leakage control. The primary modelling dataset contains $N = 175$independent aftershock sequences with mainshock magnitude $M \geq 6.5$ and at least $N_m \geq 20$ fit-window aftershocks, with targets defined over $0.1 \leq t \leq 30d$. The predictor vector is restricted to quantities available at mainshock occurrence, namely mainshock magnitude M, depth z, and three pre-mainshock background-rate descriptors at progressively larger space–time scales. Post-mainshock variables, target quantities, fit-window magnitude-type information, overlap diagnostics, and deterministic bookkeeping variables are excluded from the primary predictor matrix to prevent information leakage. Train-only preprocessing consists of logarithmic transformation, standardization, and storage of fold-specific statistics. (b) Compact multi-output neural surrogate. A shared multilayer perceptron receives the five-dimensional standardized input vector and feeds two output heads: a linear head for $\hat{y}_N = \widehat{log_{10} N_{30}^{qc}}$, and a bounded Omori-exponent head for $\hat{p}$, obtained through a sigmoid mapping to the admissible interval $[p_{min}, p_{max}] = [0.2, 2.5]$. Hidden widths and dropout are selected by blocked cross-validation. (c) Physics-embedded decoder and training objective. The neural outputs $\hat{y}_N$ and $\hat{p}$ are passed to an analytical Omori–Utsu decoder, whereby $N_{30}^{qc} = 10^{\hat{y}_N}$, $\hat{K}$ is derived analytically from N_{30}^{qc} and $\hat{p}$, and the rate function is reconstructed as $\hat{\lambda}(t) = \hat{K}(t + c)^{-\hat{p}}$ with $c = 0.01d$. Thus, $\hat{K}$ is not an independent neural target. The loss function combines a Huber loss for $log_{10} N_{30}^{qc}$, a clipped bootstrap-uncertainty-weighted Huber loss for p, and L_2 regularization; by construction, closure and monotonic temporal decay are enforced without a separate monotonicity penalty. (d) Temporal validation, benchmarks, and diagnostics. Model development is performed on the 2000–2019 subset ($N = 143$) using rolling-origin blocked cross-validation with fold-specific scaling, early stopping, and multiple random seeds, whereas the 2020–2025 subset ($N = 32$) is retained as an untouched temporal test set. Benchmark models include a magnitude-only linear model, a regularized linear multi-feature model, and an unconstrained three-output multilayer perceptron ablation. Model evaluation is based on MAE, RMSE, bias, and R^2 for both targets, together with bootstrap uncertainty, closure residuals, and admissibility diagnostics. Additional robustness analyses are performed using the expanded sensitivity dataset ($N = 276$), feature-ablation experiments, and held-out regional tests.*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

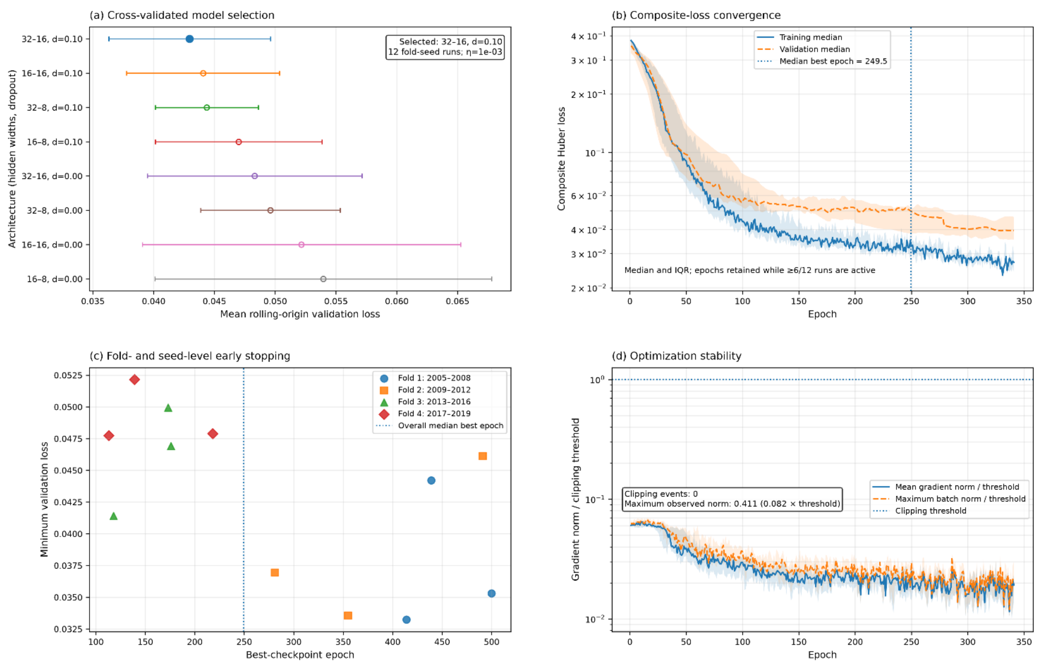

Figure 4. *Cross-validated model selection, convergence, and optimization stability of PISAS v1.0. (a) Mean composite validation loss for the eight candidate multilayer-perceptron configurations evaluated over four rolling-origin temporal folds and three random seeds per fold. Horizontal bars represent one standard deviation across the 12 fold–seed runs. Labels report the widths of the two hidden layers and the dropout fraction; all configurations use a learning rate of* 10^{-3}*. The selected architecture contains 32 and 16 hidden units, uses a dropout fraction of 0.10, comprises 754 trainable parameters, and achieves a mean composite validation loss of* 0.0430 ± 0.0067*. (b) Median training and validation composite Huber losses for the selected architecture, with interquartile envelopes across the fold–seed runs. Curves are retained only while at least 6 of the 12 runs remain active, thereby limiting late-epoch summaries based on a small number of unconverged runs. The vertical dotted line marks the median best-checkpoint epoch of 249.5, rounded to the 250-epoch budget subsequently used for final training on the complete 2000–2019 development dataset. (c) Minimum validation loss as a function of the best-checkpoint epoch for each temporal fold and random seed. Symbols distinguish the four validation intervals: 2005–2008, 2009–2012, 2013–2016, and 2017–2019. The spread among points quantifies the combined effects of temporal non-stationarity and stochastic initialization on convergence time and attainable validation loss. (d) Median and interquartile ranges of the mean and maximum batch-gradient norms, normalized by the clipping threshold of 5.0. The maximum observed gradient norm is 0.411, corresponding to 0.082 of the clipping threshold, and no clipping events occur during either cross-validation or final-model training, indicating numerically stable optimization.*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

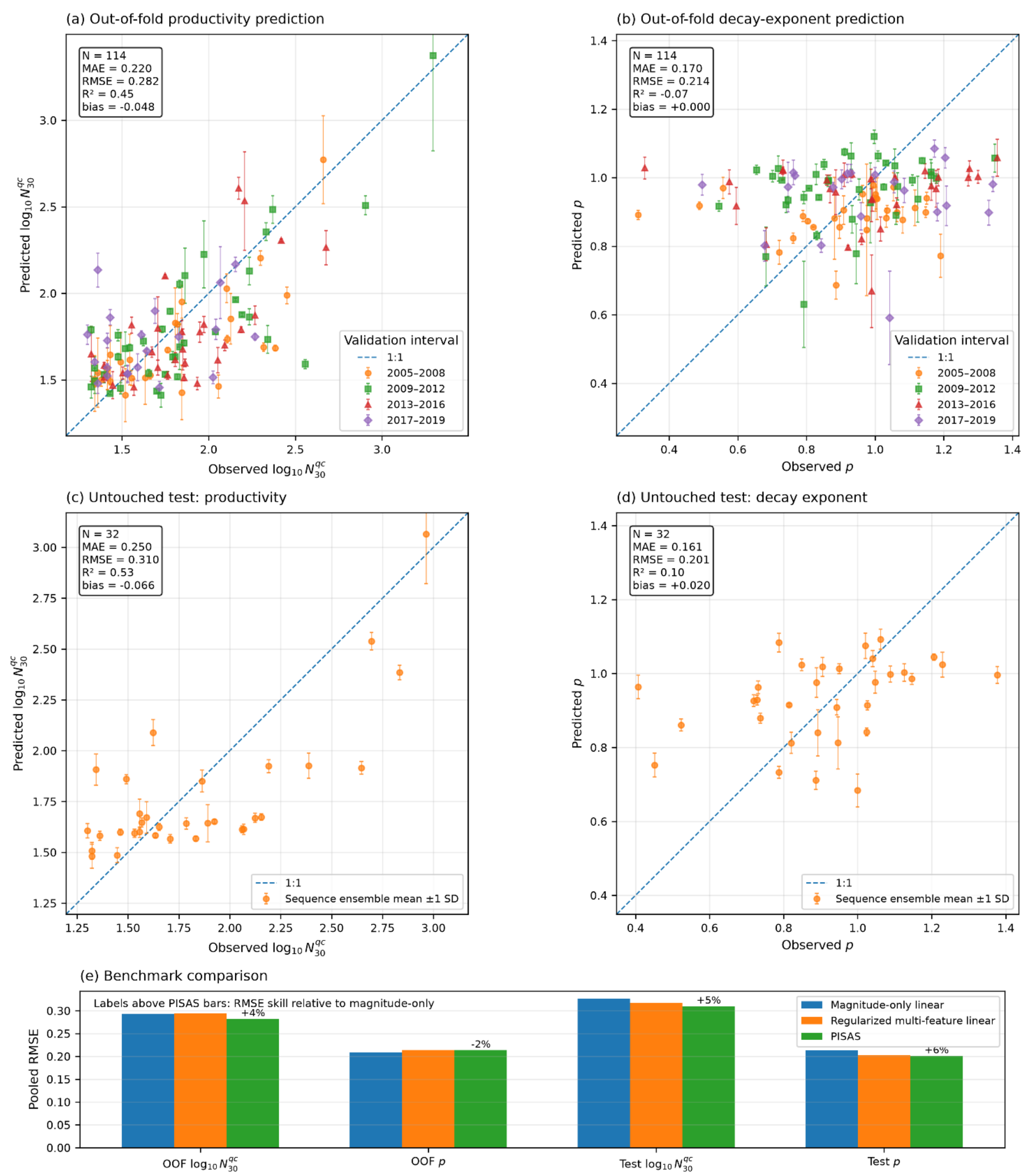

Figure 5. *Predictive skill of PISAS v1.0 under rolling-origin cross-validation and an untouched temporal test. (a) Out-of-fold ensemble-mean predictions of the quality-controlled 30 d aftershock productivity, expressed as $log_{10}N_{30}^{qc}$, for the 114 sequences entering the four rolling-origin validation intervals. Symbols distinguish the validation periods 2005–2008, 2009–2012, 2013–2016, and 2017–2019, while vertical error bars denote one standard deviation across the three independently initialized models fitted for each fold. The dashed line indicates perfect agreement. The pooled out-of-fold performance is $MAE = 0.220$, $RMSE = 0.282$, $bias = -0.048$, and $R^2 = 0.45$. (b) Corresponding out-of-fold predictions of the Omori–Utsu decay exponent p. Predictions cluster around the central part of the observed range, yielding $MAE = 0.170$, $RMSE = 0.214$, negligible aggregate bias, and $R^2 = -0.07$, indicating that the five mainshock-time predictors provide little out-of-fold skill for sequence-to-sequence variability in p. (c) Ensemble-mean productivity predictions for the 32 sequences in the untouched 2020–2025 temporal test set. Error bars represent one standard deviation*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

across five independently initialized final models trained on the complete 2000–2019 development dataset. Test performance is $MAE = 0.250$*,* $RMSE = 0.310$*,* $bias = -0.066$*, and* $R^2 = 0.53$*. (d) Corresponding temporal-test predictions of* p*, with* $MAE = 0.161$*,* $RMSE = 0.201$*,* $bias = +0.020$*, and* $R^2 = 0.10$*. (e) Pooled RMSE comparison between PISAS, a magnitude-only linear model, and a regularized multi-feature linear model for both targets and evaluation settings. Percentages above the PISAS bars quantify skill relative to the magnitude-only benchmark, defined as* $1 - RMSE_{PISAS}/RMSE_{mag}$*. PISAS improves productivity RMSE by 3.9% in out-of-fold validation and by 5.0% on the temporal test. For* p*, PISAS performs 2.1% worse than the magnitude-only model out of fold but improves test RMSE by 5.7%; however, the low or negative* R^2 *values show that this reduction in absolute error does not correspond to strong sequence-level predictive skill. RMSE values should be compared only within the same target because* $\log_{10} N_{30}^{qc}$ *and* p *have different numerical scales.*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

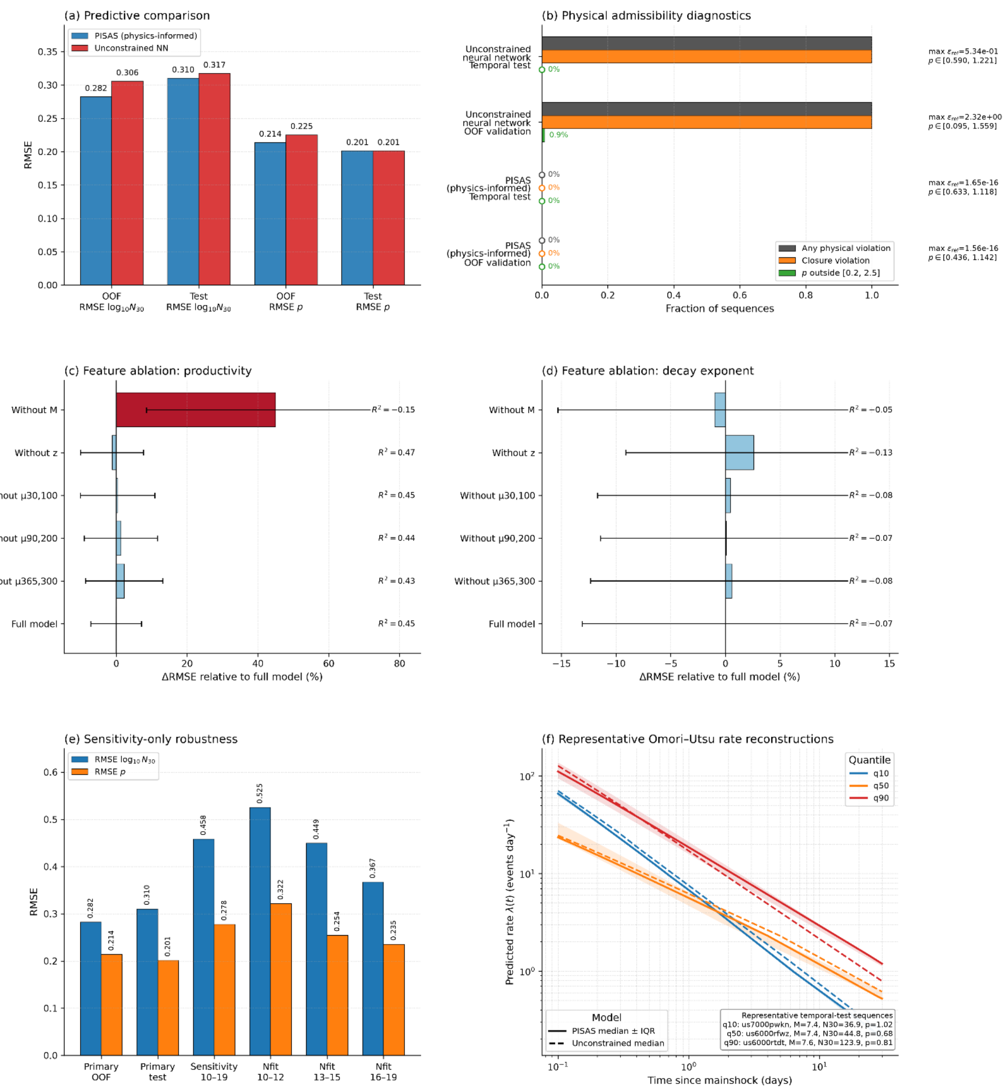

Figure 6. *Physical consistency, feature importance, and robustness of the physics-informed surrogate. (a) Predictive comparison between PISAS and the unconstrained neural-network baseline, reported as RMSE for $log_{10}N_{30}^{qc}$ and p on the out-of-fold (OOF) validation set and on the untouched temporal test set. (b) Physical admissibility diagnostics for both models on the OOF validation and temporal test sets, showing the fraction of sequences with any physical violation, closure violation, and predicted p outside the admissible interval [0.2, 2.5]; the maximum relative closure error and the range of predicted p are reported at right. (c) Feature-ablation analysis for productivity prediction, expressed as percentage change in RMSE relative to the full model after removing each predictor in turn; horizontal bars denote mean ΔRMSE and whiskers indicate variability across fold-seed realizations, while the corresponding R^2 values are reported at right. (d) Same as in (c), but for decay-exponent prediction. (e) Robustness on the sensitivity-only dataset, showing RMSE for $log_{10}N_{30}$ and p for the full sensitivity subset and for subsets stratified by the number of aftershocks used in the Omori–Utsu fit (N_{fit} = 10–12, 13–15, and 16–19), together with the primary OOF and temporal-test references. (f) Representative Omori–Utsu rate reconstructions for three temporal-test sequences corresponding to the lower, median, and upper portions of the predicted-productivity distribution (q10, q50, and q90); solid lines and shaded envelopes denote the PISAS median and interquartile range, whereas dashed*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

lines denote the unconstrained model median. Overall, the figure shows that the physics-informed formulation preserves predictive skill while enforcing strict physical admissibility, identifies mainshock magnitude as the dominant predictor for productivity, and remains stable across lower-information sensitivity subsets.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

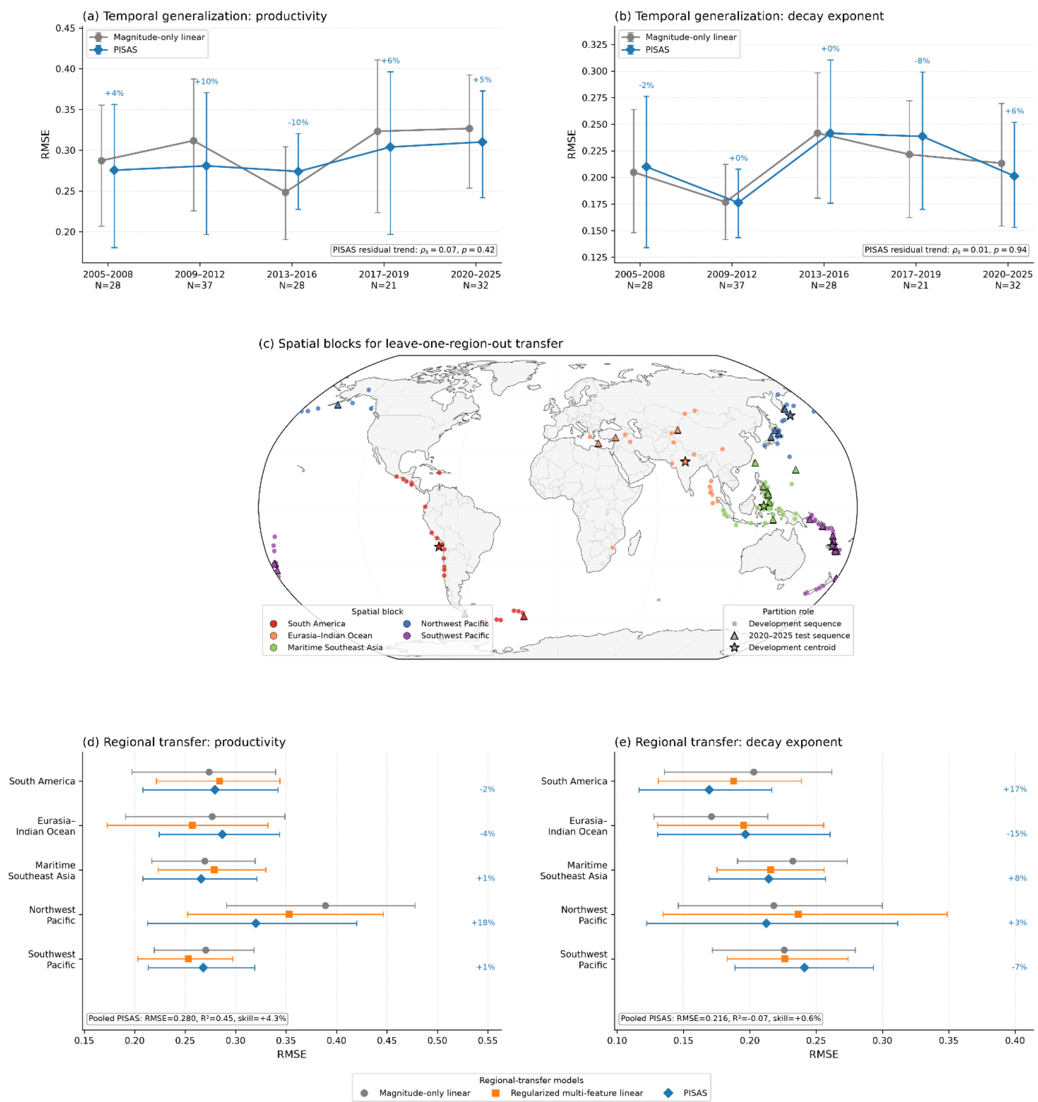

Figure 7. *Temporal and regional generalization of the PISAS surrogate. (a) Temporal generalization for productivity, expressed as root-mean-square error (RMSE) of predicted $\log_{10}N_{30}^{qc}$ across consecutive temporal blocks. (b) Same as (a), but for the Omori–Utsu decay exponent p. In both panels, the physics-informed surrogate (PISAS) is compared with the magnitude-only linear benchmark; symbols denote block-wise median RMSE and vertical bars denote bootstrap 95% confidence intervals. Percent labels above the PISAS series report RMSE skill relative to the magnitude-only benchmark, with positive values indicating lower RMSE for PISAS. The boxed annotations summarize the Spearman rank correlation between PISAS residuals and block order. (c) Spatial partition adopted for leave-one-region-out transfer tests. Colored circles indicate development sequences, open triangles indicate sequences belonging to the 2020–2025 temporal test interval, and stars denote development centroids for the five regional blocks: South America, Eurasia–Indian Ocean, Maritime Southeast Asia, Northwest Pacific, and Southwest Pacific. (d) Regional transfer performance for productivity $\log_{10}N_{30}^{qc}$. (e) Regional transfer performance for decay exponent p. In panels (d) and (e), each regional block is excluded in turn from model fitting and used as a hold-out transfer set; three models are compared: the magnitude-only linear benchmark, a regularized multi-feature linear model, and PISAS. Horizontal bars denote bootstrap 95% confidence intervals, and the blue percentage labels report RMSE skill*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

of PISAS relative to the magnitude-only benchmark. The boxed summaries at the bottom of panels (d) and (e) report pooled leave-one-region-out performance for PISAS, including RMSE, R^2, and aggregate skill.

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

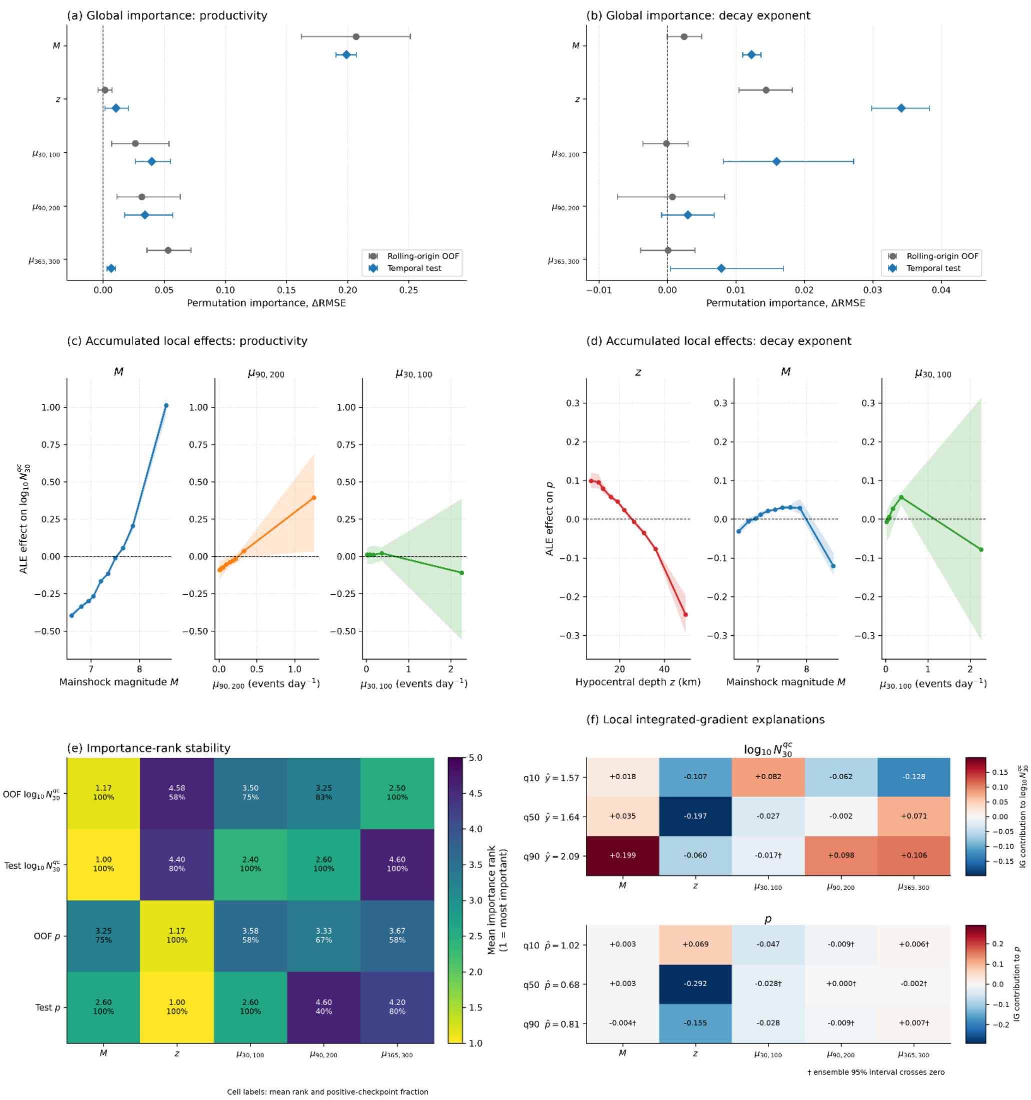

Figure 8. *Ensemble-aware interpretation of PISAS. (a) Repeated permutation importance for the quality-controlled 30-day aftershock productivity, expressed as* $\log_{10}N_{30}^{qc}$*. Symbols represent the mean increase in root-mean-square error,* $\Delta RMSE$*, obtained after independently permuting each predictor within the evaluation set; horizontal bars denote 95% bootstrap intervals across the 12 rolling-origin fold–seed checkpoints and the five final-model checkpoints evaluated on the untouched 2020–2025 temporal test. Mainshock magnitude* M *is the dominant productivity predictor under both protocols, with* $\Delta RMSE = 0.207$ *out of fold and* 0.199 *on the temporal test. (b) Corresponding permutation importance for the Omori–Utsu decay exponent* p*. Hypocentral depth* z *ranks first under both evaluation protocols, with* $\Delta RMSE = 0.014$ *out of fold and* 0.034 *on the temporal test; the substantially smaller importance values relative to the productivity target are consistent with the limited predictive skill obtained for* p*. (c) First-order accumulated local effects for the three predictors with the lowest mean importance rank for productivity: mainshock magnitude, the 90-day/200-km background rate* $\mu_{90,200}$*, and the 30-day/100-km background rate* $\mu_{30,100}$*. Curves show the mean response of the five final-model seeds over the complete 175-sequence primary predictor distribution, and shaded envelopes denote seed-wise 95% intervals. (d) Corresponding accumulated local effects for the three highest-ranked predictors of* p*: hypocentral depth, mainshock magnitude, and* $\mu_{30,100}$*. The horizontal dashed lines in panels (c) and (d)*

https://doi.org/10.5194/egusphere-2026-4241
Preprint. Discussion started: 3 September 2026

mark zero accumulated effect. (e) Stability of feature-importance ranks across evaluation protocols and prediction targets. Each cell reports the mean checkpoint rank and, beneath it, the percentage of checkpoints for which feature permutation increased RMSE; rank 1 denotes the most important predictor. (f) Integrated-gradient explanations for representative low-, median-, and high-productivity temporal-test sequences: $q10 =$*us7000pwkn (*$M = 7.4$, $\widehat{N}_{30} = 36.9$, $\hat{p} = 1.02$*);* $q50 =$*us6000rfwz (*$M = 7.4$, $\widehat{N}_{30} = 44.0$, $\hat{p} = 0.68$*); and* $q90 =$*us6000rtdt (*$M = 7.6$, $\widehat{N}_{30} = 122.8$, $\hat{p} = 0.81$*). Cell values are ensemble-mean signed contributions relative to the development-feature mean baseline, with positive and negative values increasing and decreasing the corresponding prediction, respectively. Daggers identify contributions whose seed-wise 95% interval includes zero. The productivity and* pheat *maps use independent symmetric colour scales because integrated-gradient magnitudes are expressed in the units of their respective outputs and should not be compared directly across targets.*