Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Leakage-Controlled Time-Series Forecasting Benchmark for GOES Soft X-ray and Proton Flux

Wonkeun Jo
Korea Astronomy and Space Science Institute
Eunsu Park
Korea Astronomy and Space Science Institute
Ji-Hye Baek
jhbaek@kasi.re.kr

Korea Astronomy and Space Science Institute

Article

Keywords: solar flare forecasting, soft X-ray flux, energetic proton flux, GOES, deep learning, time-series regression, space weather, benchmark, data leakage

Posted Date: October 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11110677/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Leakage-Controlled Time-Series Forecasting Benchmark for GOES Soft X-ray and Proton Flux

Wonkeun Jo[1], Eunsu Park[2], Ji-Hye Baek[1,3*]

[1*]Center for Space Exploration Technology, Korea Astronomy and Space Science Institute, 776 Daedeok-daero, Yuseong-gu, 34055, Daejeon, Republic of Korea.
[2]Center for Heliophysics Observations and Technology, Korea Astronomy and Space Science Institute, 776 Daedeok-daero, Yuseong-gu, 34055, Daejeon, Republic of Korea.
[3]Astronomy and Space Science, University of Science and Technology (UST), 217 Gajeong-ro, Yuseong-gu, 34113, Daejeon, Republic of Korea.

*Corresponding author(s). E-mail(s): jhbaek@kasi.re.kr;
Contributing authors: jowonkeun@kasi.re.kr; eunsupark@kasi.re.kr;

Abstract

Forecasting solar soft X-ray (SXR) and proton flux underpins space-weather operations, yet machine-learning work largely frames the task as flare classification, not continuous-flux prediction, and the few regression studies use one solar cycle, under-specified splits, and no baselines. We present a time-series benchmark for multi-step regression of two drivers, the GOES 1–8 Å SXR flux (2002–2026) and the $\geq$10 MeV proton flux (1999–2026), on a 5-minute cadence with historical horizons in days and forecast horizons in hours. A leakage-controlled, term-based rotating k-fold assigns half-year terms atomically, and event skill is graded against persistence and the operational National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC) forecast with imbalance- and extreme-aware metrics. Across 36 settings a segment-recurrent model (SegRNN) attains the lowest root-mean-square error (RMSE) in 35, yet the aggregate ranking conceals an event-skill artefact: its SXR M-class per-day Heidke skill score (HSS) near 0.10 sits on the persistence floor, far below occurrence-persistence near 0.46, matched by the operational reference, and no model separates from the proton floor near 0.79. The impulsive SXR regime stays essentially unpredicted at 12–24 h, reaching parity only at 3–6 h despite low global error. We release the dataset and code.

Keywords: solar flare forecasting, soft X-ray flux, energetic proton flux, GOES, deep learning, time-series regression, space weather, benchmark, data leakage

1 Introduction

Solar soft X-ray (SXR) irradiance, monitored almost continuously since the 1970s by the Geostationary Operational Environmental Satellite (GOES) X-Ray Sensor (XRS) in the 0.5–4 Å and 1–8 Å bands, is the defining observable of solar flares and a primary driver of the day-side ionosphere and thermosphere; the associated flux of solar energetic protons poses radiation hazards to spacecraft, aviation, and astronauts. The 1–8 Å channel underpins the National Oceanic and Atmospheric Administration (NOAA) A/B/C/M/X flare classification, and the flux level together with its time integral governs the ionospheric and thermospheric response. Operational users therefore need more than a binary statement that a flare will occur. They need to know how the flux will evolve — its rise, peak, and decay. The operational need motivates forecasting the flux as a continuous quantity, not merely detecting or grading discrete events.

Despite this, the machine-learning (ML) flare-forecasting literature is dominated by classification, predicting the occurrence or NOAA class of a flare within a fixed window (e.g., the DeFN family [1, 2], SolarFlareNet, and magnetogram-based classifiers such as Bobra and Couvidat [3]; recent variants use multimodal large language models [4] and transformers [5]; see the review by Shao et al. [6]). Classification discards magnitude information, hinges on threshold choices, and its headline metrics are notoriously sensitive to class imbalance. Direct regression of the SXR flux is comparatively rare, and the existing attempts are individually incomplete as benchmarks.

Yi et al. [7] forecast the 30-minute flux profile with a sequence-to-sequence network, but on essentially a single cycle and with a random, event-level split. Dey and Fuentes [8] autoregress the 1–8 Å flux 1–5 h ahead but do not specify whether their split is chronological. van der Sande and Muñoz-Jaramillo [9] recast forecasting as regression yet reduce the target to a single 24-h maximum. Models that ingest the X-ray curve itself still predict flare occurrence rather than the continuous flux [10, 11]. For the energetic-proton channel, sequence models already regress the integral flux profile directly, from GOES [12–14] and other monitors [15, 16], yet on the proton driver alone and without a leakage-controlled multi-cycle protocol. Thus, while continuous flux regression itself is not new, no existing study combines it with the components that make a benchmark reliable.

Two methodological pathologies recur across the field. First, data leakage: random shuffling, active-region overlap between train and test, and model selection on test metrics all inflate apparent skill. Several studies demonstrate this on their own data, with the True Skill Statistic (TSS) or the area under the receiver operating characteristic curve (AUC) dropping sharply once a leakage-controlled split is imposed [17–19]. Second, absent baselines and metric artefacts: many deep models are never compared against an operational reference, and reliance on TSS alone obscures poor

Paper	(A) Continuous Regression	(B) Multi-Cycle	(C) Leakage-Controlled Split	(D) Operational Reference	XRS Target
Our	Yes	Yes	Yes	Partial	Yes
Yi [7]	Yes	No	No	No	Yes
Dey [8]	Yes	No	No	Partial	Yes
van der Sande [9]	No	Yes	Yes	No	No
SuryaBench [24]	No	Yes	Yes	No	No
SolarFlareNet [25]	No	Yes	Partial	No	No

Table 1 Positioning against prior benchmarks. No prior study satisfies (A) continuous flux regression, (B) multi-cycle coverage, and (C) a leakage-controlled split together on the XRS 1–8 Å channel; (D) marks a secondary reference comparison to the operational NOAA/SWPC forecast, not a target. Cells are Yes / No / Partial; the per-study qualifiers (regression cadence, split type, target quantity) are given in the text.

precision under class imbalance, the "TSS trap" [20–22], whose regression analogue is the degradation of extreme-quantile skill behind strong global scores [23].

We address these gaps with a reproducible benchmark for continuous GOES flux forecasting. Yi et al. [7] and Dey and Fuentes [8] already regress the continuous flux. Our contribution is the first benchmark, summarized in Table 1, to jointly enforce continuous flux regression (A), multi-cycle coverage (B), and a leakage-controlled term-based split (C) on a single flux-forecasting task, while additionally reporting a reference comparison to the operational NOAA Space Weather Prediction Center (SWPC) forecast (D). The benchmark further provides a systematic model-vs-model comparison of modern deep time-series models across an SXR track, a proton track, and a multivariate (joint SXR + proton) track.

The column headers are (A) continuous flux regression, where the model outputs the flux time series rather than a scalar summary or a class label; (B) coverage of more than one solar cycle; (C) a split that prevents temporal leakage; and (D) a reference comparison to the operational NOAA/SWPC forecast. The regression studies [7, 8] satisfy A, at 30-minute and 1–5-hour resolution, but run on a single cycle with random or unspecified splits (B, C). The clean-split benchmarks [9, 24] secure B and C but replace continuous 1–8 Å regression with a 24-hour maximum or a flare-class label (A). Only the present work satisfies A–C together on the XRS 1–8 Å channel; the operational comparison (D) is reported as a secondary reference rather than a target.

Released as an open, end-to-end pipeline rather than a static leaderboard, the benchmark fixes the evaluation protocol once so that two communities which rarely share one co-develop a single, leakage-controlled task. Space-weather practitioners contribute physically grounded targets and event thresholds, while machine-learning and time-series researchers contribute architectures and training methods, each held to the identical leakage-controlled evaluation. The contributions are as follows.

- **A public, real-world ML time-series forecasting benchmark** for continuous GOES flux (SXR 1–8 Å and $\geq$10 MeV proton): task, leakage-controlled split, metrics, and a leaderboard of 10 modern deep time-series models, all reconstructed from NOAA GOES products as mirrored in the operational space-weather archive of the

Korea Space Weather Research Center (KSWRC, https://kswrc.kasi.re.kr/ko) at the Korea Astronomy and Space Science Institute (KASI).

- **A multi-cycle, leakage-controlled protocol**: atomic half-year (`year_half`) terms in a rotating k-fold, so no input/label window crosses a split boundary and every fold spans all solar-cycle phases.
- **Imbalance- and extreme-aware evaluation**: event skill scored with metrics that avoid the TSS trap and its regression analogue, referenced to persistence floors and, as a secondary comparison, the operational NOAA/SWPC forecast.
- **A systematic model-vs-model comparison** (leaderboard) across the SXR-univariate, proton-univariate, and multivariate (joint) tracks, surfacing a marked SXR event-skill degradation under competitive aggregate error, and showing that joint cross-driver input lowers aggregate error without raising event skill.
- **An open, community-facing benchmark**. The task, data-reconstruction pipeline, term-atomic splits, and scoring are released on GitHub (https://github.com/merchen911/GOES-time-series), with a public leaderboard that invites external submissions under the identical leakage-controlled protocol.

2 Results

2.1 Dataset Composition and Dynamical Regimes

The leakage-controlled rotating k-fold described in Methods partitions the record into five interleaved train/validation/test rotations, each spanning all solar-cycle phases. In the reference fold, counted at each channel's native cadence, the SXR target (1-minute) contributes 7.10M training, 2.19M validation, and 2.02M test observations across 30 / 9 / 9 half-year terms, and the proton target (5-minute) 1.40M / 0.48M / 0.45M across 33 / 11 / 11 terms. The models operate on the common 5-minute grid onto which the 1-minute SXR channel is aggregated, so the SXR modelling timesteps number about one-fifth of its native count. The partition is near 60 / 20 / 20 by term, and the rotation preserves it across all five folds.

Supplementary Fig. S2 shows that the held-out sets remain event-bearing: the reference test set carries 25,292 M-class, 1,021 X-class, and 9,325 solar-proton-event samples. Across the full record the impulsive-event classes are rare and strongly imbalanced: M-class ($\geq 10^{-5}$ W m^{-2}) accounts for 0.79% of SXR samples, X-class ($\geq 10^{-4}$ W m^{-2}) for 0.037%, and solar proton events (SPE, $\geq$10 pfu) for 2.74% of proton samples. The per-fold test event rate stays non-zero and of the same order across the rotation, ranging 0.59–1.25% for M-class, 0.026–0.051% for X-class, and 2.0–4.5% for SPE, with a cross-fold coefficient of variation of 0.24–0.37. A contiguous forward holdout, by contrast, concentrates the test set in a single cycle phase and can leave one or more event classes near-empty, unlike the interleaved atomic-term rotation detailed in the Methods (Section 4).

Figure 1 places the flux on the complexity–entropy plane, where it separates into two dynamical regimes. Quiet proton background occupies the high-entropy, low-complexity noise corner, with mean permutation entropy $H = 0.95$ and statistical complexity $C = 0.09$, whereas SPE windows migrate to lower entropy and higher complexity at $H = 0.91$ and $C = 0.15$. For the SXR target the two regimes share

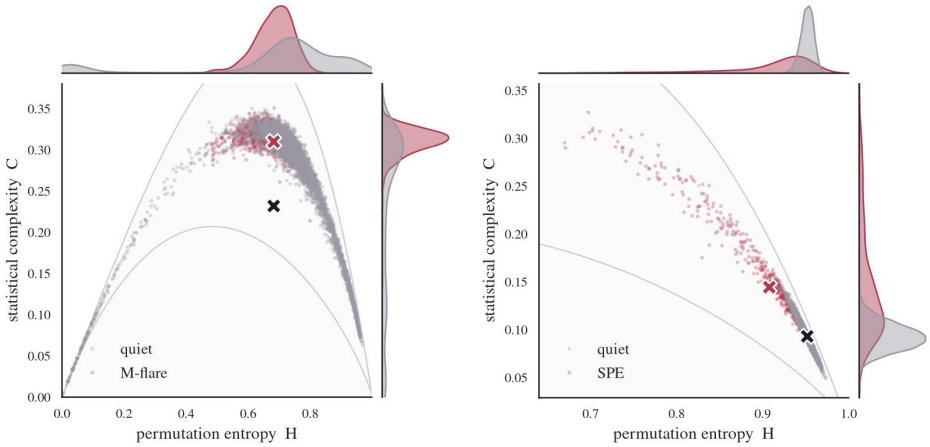


Fig. 1 Two dynamical regimes (complexity–entropy plane). The left panel shows the XRS 1–8 Å target; the right panel the proton ≥10 MeV target. The × marker locates the per-regime mean. The central scatter plots permutation entropy H versus statistical complexity C for sliding sequence-length windows (within the theoretical $C_{\min}/C_{\max}$ envelope), with per-axis kernel-density-estimate marginals by regime. Proton quiet windows concentrate in the high-H, low-C noise corner (mean $H, C = 0.95, 0.09$); SPE windows shift to lower entropy and higher complexity (0.91, 0.15). For XRS the two regimes share almost the same entropy (mean $H \approx 0.68$) but M-flare windows sit at markedly higher complexity (C rising from 0.23 to 0.31).

almost the same entropy near $H \approx 0.68$, but M-flare windows sit at markedly higher complexity, C rising from 0.23 to 0.31. The extremes therefore form a structurally distinct regime rather than a larger-amplitude background. The companion forecastability panel (Supplementary Fig. S1) shows that increment forecastability is low in both regimes, so the low aggregate root-mean-square error (RMSE) reflects persistence of the level rather than predictable increments, the data-intrinsic basis for the SXR event-skill degradation documented in Section 2.3 and the Discussion.

2.2 Forecasting Performance

2.2.1 Model Comparison

Table 2 summarizes the RMSE over all historical and forecasting combinations. On aggregate accuracy, SegRNN attains the lowest 5-fold-mean RMSE in 35 of the 36 base cells (mean rank 1.03; the exception is proton forecasting at the 3-h horizon from a 1-day history, where iTransformer attains a marginally lower RMSE) and is the only model robust across historical horizons (Section 2.2.2). The large gaps in the averaged table are mostly long-input divergences. At a fixed sensible cell (1-day history, 12-h horizon) the top models lie within ∼0.02 log-flux units (≈5% in linear flux); models that diverge at the 7-day historical horizon are penalized by the averaging.

A Friedman test over the fold-collapsed per-configuration log-flux RMSE rejects rank equality across the ten models ($\chi^2_9 = 150.3$, $p < 10^{-27}$; $N = 24$ fold-averaged configurations spanning the four horizons 3, 6, 12, 24 h, the complete-case set at the

[1] xPatch diverges at the longest (7-day) historical horizon; its means exclude those cells.

Rank	Model	XRS 1-8Å	Proton	Multi
1	SegRNN	$\mathbf{0.206 \pm 0.018}$	$\mathbf{0.185 \pm 0.021}$	$\mathbf{0.183 \pm 0.022}$
2	xPatch[1]	0.228 ± 0.019	0.207 ± 0.027	0.192 ± 0.020
3	PatchMixer	0.451 ± 0.144	0.563 ± 0.476	0.224 ± 0.035
4	iTransformer	0.219 ± 0.022	0.228 ± 0.049	0.235 ± 0.047
5	PatchTST	0.225 ± 0.026	0.303 ± 0.065	0.243 ± 0.058
6	RLinear	0.233 ± 0.030	0.618 ± 0.574	0.267 ± 0.081
7	TiDE	0.235 ± 0.031	0.588 ± 0.537	0.272 ± 0.082
8	DLinear	0.754 ± 0.183	0.482 ± 0.239	0.276 ± 0.067
9	TSMixer	0.820 ± 0.159	0.327 ± 0.077	0.706 ± 0.175
10	FReTS	0.328 ± 0.056	0.471 ± 0.103	1.759 ± 3.561

Table 2 Mean RMSE of log-flux. Log-flux RMSE, mean ± 95% confidence interval across the five folds, by model and track (`uni_b` SXR, `uni_a` proton, `multi` joint), averaged over the three historical horizons (1, 3, and 7 days) and the four forecast horizons (3, 6, 12, and 24 h); the lowest RMSE in each track is in bold, and rows are ordered by the multivariate-track RMSE. The wide intervals on the lower-ranked models reflect fold-to-fold divergence at the longest historical horizon.

1- and 3-day historical horizons where every model produced a valid run). The five folds are averaged within each configuration before ranking because adjacent stride-5 folds share most of their training terms and are not independent replicates; the complete-case set omits the 7-day historical-horizon configurations, where SegRNN's RMSE margin is largest, so the omission is conservative. SegRNN holds the top average rank of 1.04, first in 23 of the 24 configurations, and its gap to the second-ranked xPatch at average rank 3.33 is below the Nemenyi critical difference of 2.77 ($\alpha = 0.05$), so the diagram joins SegRNN with the top cluster of xPatch and iTransformer while separating it by more than the critical difference from the remaining models, among which the divergent linear (DLinear) and multilayer-perceptron (MLP; FReTS, TSMixer) models rank lowest. The rank test certifies the consistency of the ordering across configurations rather than the size of the RMSE gaps; Holm-corrected Wilcoxon signed-rank tests place SegRNN ahead of every other model at $p < 2 \times 10^{-6}$ (Figure 2).

2.2.2 Sensitivity to Historical & Forecast Horizon

The historical horizon sharply separates the model families. As the historical horizon lengthens from 1 to 7 days, SegRNN is the only model whose per-track RMSE stays flat or improves, while the linear and MLP families (DLinear, PatchMixer, FReTS, TSMixer) destabilize, their RMSE growing by up to an order of magnitude, and the transformers degrade mildly; xPatch diverges at the longest historical horizon (per-track detail in Supplementary Fig. S3). The benefit is track-dependent: a longer historical horizon helps proton most (SegRNN's proton RMSE improves from 0.206 to 0.172 at the 12 h forecast horizon), is neutral for soft X-ray, and slightly degrades the multivariate track. Across the forecast horizon the RMSE ranking is preserved, every model degrading monotonically from 3 to 24 h.

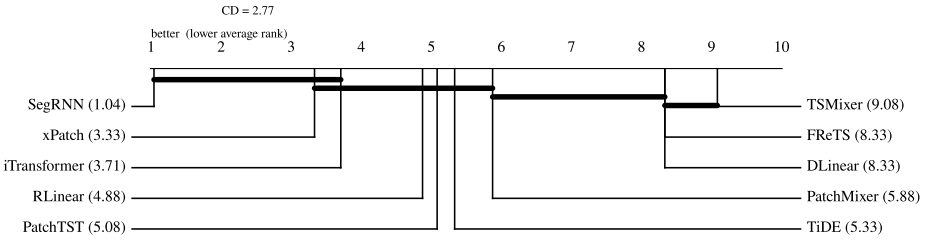


Fig. 2 Critical-difference diagram of the RMSE ranking. Average rank of each model over the $N = 24$ fold-averaged benchmark configurations spanning the four horizons 3, 6, 12, 24 h (Friedman $\chi^2_9 = 150.3$, $p < 10^{-27}$); lower is better, and models joined by a bar differ by less than the Nemenyi critical difference (CD = 2.77, $\alpha = 0.05$). SegRNN holds rank 1.04, first in 23 of the 24 configurations; it joins the top cluster of xPatch and iTransformer within the critical difference and is separated by more than the critical difference from the remaining models. xPatch, divergent at the longest (7-day) historical horizon, is ranked on the fully populated historical horizons.

Aggregate error is only half the horizon story, because event skill scales differently. Figure 3 tracks the best per-day M-class event skill against the two persistence floors across the four horizons. On soft X-ray the shortfall to occurrence-persistence is horizon-graded. At 12–24 h the models fall far below the occurrence floor, with a per-day Heidke skill score (HSS) of 0.217 ± 0.159 at 12 h and 0.135 ± 0.122 at 24 h against occurrence floors of 0.462 ± 0.075 and 0.502 ± 0.087. As the horizon shortens the gap closes, and at 3–6 h the best model (DLinear at every horizon) reaches statistical parity with occurrence-persistence, 0.304 ± 0.081 against 0.341 ± 0.092 at 3 h and 0.343 ± 0.122 against 0.391 ± 0.075 at 6 h (paired Wilcoxon $p = 0.63$ and $p = 0.19$). The best model's mean stays above flat-value persistence at every horizon. The gap is also robust to the historical horizon: re-scoring the best model at the 3-day and 7-day historical horizons leaves its per-day soft X-ray skill below the occurrence-persistence floor at every forecast horizon (per-fold scores are released with the benchmark). Proton event skill stays within its persistence floors at every horizon, and the aggregate per-point skill decay (Supplementary Fig. S4) is shallow for proton and steep for soft X-ray, quantifying the same horizon sensitivity on the leaderboard metric.

2.3 Forecast Evaluation — Event-Skill Decoupling

Figure 4 shows the principal result, the decoupling of accuracy and event skill on the impulsive SXR target. Minimizing log-flux RMSE rewards fitting the smooth quiet background, so the lowest-RMSE model, SegRNN, forecasts M-class crossings at a small fraction of their observed frequency (frequency bias ≈ 0.06) and its per-day event skill sits on the flat-value persistence floor. Averaged over the five folds its per-day HSS is 0.096 ± 0.039 and TSS 0.057 ± 0.027, inside the interval of the flat-value floor at HSS 0.112 ± 0.032, far below occurrence-persistence at HSS 0.462 ± 0.075 and the operational reference at HSS 0.47 and TSS 0.44. A volatility-retaining model reaches higher event skill despite worse RMSE. DLinear attains the highest SXR event-skill mean, per-day HSS 0.217 ± 0.159, at a worse RMSE. Its mean is nearly double the trivial flat-value floor at 0.112 ± 0.032, yet the five-fold design cannot resolve the gap,

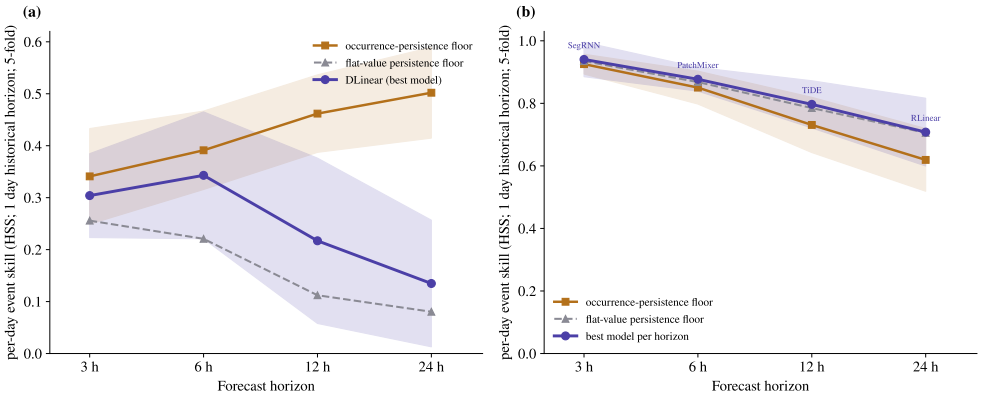


Fig. 3 Event skill versus forecast horizon. Best-model per-day M-class HSS (5-fold mean, 95% CI band) against the occurrence-persistence and flat-value persistence floors, at the 1-day (reference) historical horizon, for (a) soft X-ray and (b) proton. On soft X-ray the shortfall to occurrence-persistence closes as the horizon shortens, reaching parity at 3–6 h, and the best event-skill model is DLinear at every horizon; on proton the model tracks both floors at every horizon. Per-day granularity (the operational contingency unit), not the per-point aggregate of Supplementary Fig. S4.

a paired signed-rank test over the folds giving $p = 0.12$ with four of five folds favouring DLinear, so the flat-value floor is reported as a descriptive reference rather than a rejected alternative. Against the strong occurrence-persistence baseline at 0.462 ± 0.075 the negative result is instead resolved: DLinear falls below it in all five folds, and its upper confidence bound of 0.376 sits under the baseline's lower bound of 0.387, so no model reaches occurrence-persistence on the per-day SXR target.

On the persistent proton target the two axes re-couple. The RMSE-strong models also reach high event skill, SegRNN at per-day HSS 0.774 ± 0.109 alongside the other high-scoring models between 0.76 and 0.79, though the cluster overlaps the flat-value floor at 0.785 ± 0.089 because solar proton events persist for days. The trap is explicit on proton. TSS instead ranks the higher-frequency forecasters first: PatchMixer reaches per-day TSS 0.752 ± 0.127 and DLinear 0.732 ± 0.091, both above SegRNN's 0.650 ± 0.144, tracking their higher frequency bias (0.90 and 0.93 against SegRNN's 0.66), whereas HSS leaves the three within one interval, 0.76 to 0.79.

Accuracy and event skill therefore diverge only where the events are impulsive, and the accuracy leaderboard alone would rank an event-blind model first on SXR. The decoupling holds at the 24 h horizon. At the 24-hour horizon the best SXR M-class forecast reaches only a per-day HSS of 0.135 ± 0.122, well under the occurrence-persistence floor, while the proton forecasts hold near a per-day HSS of 0.70 and track their flat-value floor at 0.706 ± 0.046 (all re-scored at the 24-h forecast horizon across the five folds).

The same evaluation applies to the multivariate track. The benchmark's third track (`multi`) forecasts both drivers jointly. Supplementary Fig. S5 applies the identical event-skill evaluation to each channel, one panel per channel, with the numbers in Supplementary Table S4. Per-channel RMSE is at or below the univariate track, yet

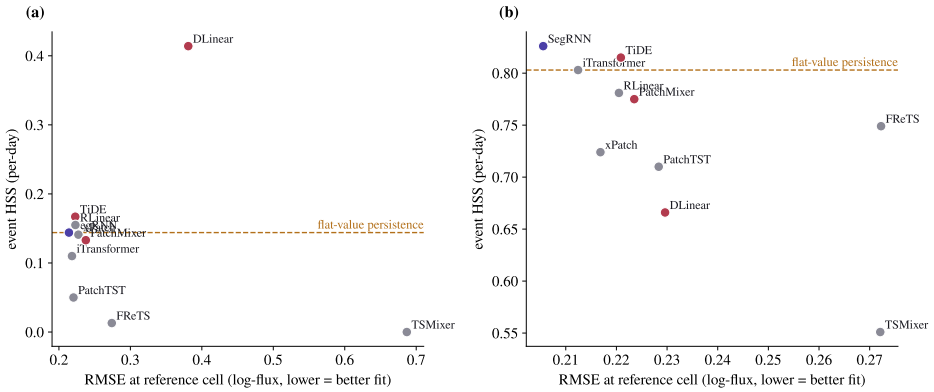


Fig. 4 Accuracy and event skill decouple only for impulsive events. Per-model reference-cell RMSE (x) versus per-day event HSS (y); the dashed line marks flat-value persistence. (a) On the impulsive SXR target the lowest-RMSE model (SegRNN) sits on the persistence floor while a worse-RMSE model (DLinear) attains higher event skill. (b) On the persistent proton target the two axes re-couple. Reference fold (fold 0); cross-fold intervals are in Supplementary Table S5.

event skill follows the same pattern. On the impulsive SXR channel every model sits near the persistence floor regardless of fit, and on the persistent proton channel skill tracks persistence. The RMSE–event-skill decoupling is therefore a property of the task, reproduced whether a target is forecast alone or jointly.

Cross-driver input shifts aggregate error but not event skill. The best univariate SXR event-skill model, DLinear, keeps a comparable per-window margin under joint input, HSS 0.216 ± 0.125 single-driver and 0.194 ± 0.109 joint, both near twice their persistence floor but with intervals that overlap it, while proton event skill is preserved. Both input regimes are reported for every target, so the benchmark scores the three tracks on the same footing.

2.4 Skill versus Persistence and the Operational Reference

Two persistence baselines must be distinguished (Supplementary Fig. S6; Supplementary Table S5). Flat-value persistence (flux held at its last value) issues few event predictions for impulsive flares, whereas occurrence-persistence (an event recurs if one just occurred) exploits the multi-day persistence of event activity. On SXR the operational M-class skill, HSS 0.47 and TSS 0.44 [26], matches occurrence-persistence at HSS 0.462 ± 0.075 and TSS 0.456 ± 0.070; the same study reports that SWPC does not consistently outperform persistence and identifies the TSS trap, and the SXR flux-regression models fall below both.

On proton, solar proton events persist for days, so the forecasts cluster near the persistence floors. The best-performing models reach per-day HSS between 0.76 and 0.79, overlapping the flat-value floor at 0.785 ± 0.089 and the occurrence floor at 0.731 ± 0.088, so no model separates from persistence across folds. In TSS, the metric the operational proton forecast is published in, the models reach per-day 0.65–0.75, competitive with the operational NOAA/SWPC proton skill of 0.49–0.69 [27] and with the persistence floors, 0.664 ± 0.126 for flat-value and 0.744 ± 0.103 for occurrence; at

the 2.74% SPE base rate TSS is far less degenerate than at the 0.79% M-class rate, and a strict head-to-head aligned to the SWPC horizon and contingency unit is left to future work.

The comparison is horizon-dependent. At 24 h the occurrence-persistence floor rises for SXR, from a per-day HSS of 0.462 at 12 h to 0.502 ± 0.087, but falls for proton, from 0.731 to 0.619 ± 0.100, so the SXR models fall further behind while the best proton forecasts stay near their floors.

The X-class tail degrades further. X-event days remain scarce, a few dozen across the five folds, so per-day X-class skill is essentially zero for every model, and at per-window granularity HSS falls toward the floor, the best flux model reaching only 0.046 ± 0.015 against the flat-value-persistence floor at 0.053 ± 0.025 and the occurrence-persistence floor at 0.204. Only the Symmetric Extremal Dependence Index (SEDI), which stays defined as the base rate approaches zero, resolves residual extremal dependence; across folds the better models reach 0.44–0.48, comparable to the persistence floors at 0.457 (flat-value) and 0.515 (occurrence), so no model attains operationally useful X-class skill.

Two uncertainty analyses confirm that the decoupling is not an artefact of the reference fold. Across the five folds the per-day event skill carries wide intervals and no model separates from flat-value persistence on either target: SegRNN reaches SXR HSS 0.096 ± 0.039 against the floor at 0.112 ± 0.032 and proton HSS 0.774 ± 0.109 against 0.785 ± 0.089 (Supplementary Table S5). A complementary day-level bootstrap over the pooled test coverage agrees, with SegRNN's per-day TSS interval overlapping that of persistence on every track: proton 0.650 (0.603–0.698) against 0.671 (0.622–0.714), and SXR 0.058 (0.045–0.072) against 0.063 (0.050–0.079).

3 Discussion

A multi-cycle, leakage-controlled regression benchmark

The benchmark establishes a leakage-controlled, multi-cycle, continuous-flux regression task on GOES soft X-ray (SXR) and proton observations, evaluated across ten modern deep time-series architectures on a strictly term-partitioned split spanning 1999–2026. By predicting the continuous 1–8 Å flux (`xrs_long`) and the $\geq$10 MeV proton flux (`p_gt10`) as regression targets, rather than reframing the problem as daily classification or window-maximum estimation, the benchmark addresses a gap in the flare-forecasting literature. We do not claim the first SXR flux regression: continuous flux-regression precedents such as Yi et al. [7] and Dey and Fuentes [8] already regress the continuous flux, but cover a single solar cycle and leave their split protocols underspecified, so the joint requirements of multi-cycle coverage and leakage-controlled evaluation remain open. van der Sande and Muñoz-Jaramillo [9] applies rigorous forward splitting but predicts a 24-h flux maximum, and SolarFlareNet [25], published in this journal, exemplifies the categorical framing from which the present continuous-regression formulation departs. Across RMSE and mean absolute error (MAE) on the log scale, the segment-recurrent SegRNN architecture attains the lowest RMSE in 35 of the 36 evaluated cells, outperforming the transformer and linear model families most clearly on the long SXR and proton sequences. SegRNN's advantage is its robust

use of a long historical horizon, holding or improving accuracy as the history grows to 7 days where the linear and MLP families become unstable under the shared training budget. Because a single learning rate is used for every model, an optimization component in those long-input divergences cannot be excluded without per-model tuning, so most architectures are best kept short at this budget. The best-model choice is therefore conditional on historical horizon and forecast horizon, which a benchmark of this scope must surface rather than reduce to the single value the flare-forecasting literature routinely reports.

Aggregate error and event skill decouple

The principal result is that RMSE and event skill decouple on the impulsive soft X-ray target. Minimizing log-flux RMSE rewards fitting the quiet background, which dominates the record, so the error-optimal solution converges toward a smoothed reconstruction that under-represents impulsive excursions. The RMSE-best model, SegRNN, issues almost no threshold-crossing M-class predictions; its per-day precision approaches 1 while its frequency bias falls to $\approx$0.06, so it forecasts events at a small fraction of their observed frequency and its M-class event skill descends to the persistence floor (per-day HSS 0.096 ± 0.039 against the floor at 0.112 ± 0.032). A volatility-retaining architecture such as DLinear preserves more of the high-frequency structure and attains the highest soft X-ray event-skill mean (per-day HSS 0.217 ± 0.159) at worse RMSE, a direct exchange between aggregate error and threshold-crossing detection, though its wide cross-fold interval overlaps the persistence floor. Including the linear and DLinear baselines follows the LTSF-Linear (long-term time-series forecasting) caution [28] that trivial baselines expose over-optimistic claims of complex architectures, and here they do more than that, since a linear baseline posting the highest event skill directly shows that aggregate accuracy and threshold-crossing detection reward different architectures. The same tension recurs across forecasting domains: for financial series, Huang et al. [29] add a direction-aware loss because minimizing squared error alone does not secure the practically decisive trend direction.

On the persistent proton target the two axes re-couple. SegRNN attains the lowest RMSE while its proton event skill (per-day HSS 0.774 ± 0.109) sits within the persistence cluster, and a base-rate-dependent score would instead rank the higher-frequency forecasters first, so a single accuracy figure cannot rank models for operational use across both regimes.

The multivariate track reinforces the divergence. Joint cross-driver input lowers aggregate error yet does not raise event skill on either channel. An additional information channel does not overcome the over-smoothing the squared-error objective imposes on impulsive onsets.

A partly intrinsic event-skill shortfall

The two targets diverge in event skill because their event dynamics differ in predictability. Solar proton events persist for days once initiated, so level-persistence is structurally predictable, and consequently every model together with the persistence baseline clusters at SEDI 0.88–0.95 on the proton target. SXR flares, by contrast, are

impulsive. The complexity–entropy and forecastability analysis (Section 2.1; Supplementary Fig. S1) shows that increment forecastability, the ForeCA spectral measure Ω evaluated on the differenced log-flux, is uniformly low in both quiescent and event regimes. Because a squared-error objective converges toward the conditional mean under low-forecastability increments, and the conditional mean on rare, information-poor onsets is the persistent level, the SXR event-skill shortfall is partly intrinsic to the data rather than solely a modelling deficiency.

The horizon grading of the occurrence-persistence floor sharpens the interpretation. Occurrence-persistence measures the predictability of flare activity rather than of individual onsets, since active regions remain flare-productive over several days, so an M-class event in one interval forecasts one in the next and the skill of the rule rises as the forecast window widens, from a per-day HSS of 0.46 at 12 h to 0.50 at 24 h. At a 3–6 h window the rule instead rests on the timing of the next onset within hours, which the broad flare waiting-time distribution leaves largely stochastic, so the floor is lowest there. The flux-regression model recovers little of the day-scale activity state from the soft X-ray history alone, and it approaches the floor only at the shortest horizons where that state is least informative, whereas at the operational 12–24 h horizons the attainable skill is governed by an activity persistence the flux stream does not encode.

The regime contrast, a widening weighted-permutation-entropy minus permutation-entropy (WPE−PE) gap and rising complexity within event windows, identifies the extremes as a distinct high-amplitude regime that the aggregate error under-weights. Such under-weighting motivates onset-sensitive objectives (peak-weighted, quantile, or extreme-value losses) and exogenous precursors (magnetograms, coronal mass ejections, extreme-ultraviolet observations) over larger models trained under the same mean-squared-error objective. Architectures that condition the residual model on the trend state, letting the predicted variance rise where the series is most volatile, are one concrete path in this direction [30].

Metric selection under class imbalance

Metric selection follows directly from the base-rate structure of the targets. As the event base rate approaches zero, the True Skill Statistic (TSS) degenerates toward the probability of detection (POD), so a model that raises POD by over-forecasting can post a favourable TSS while incurring an impractical false-alarm rate. TSS remains the correct statistic for comparing forecasts across datasets at differing base rates, the cross-cycle setting for which Bloomfield et al. [20] recommend it; within a single fixed-base-rate benchmark the degeneration surfaces only as a ranking pathology, where its base-rate independence no longer discriminates and over-forecasting inflates the score. Event skill is therefore reported as two complementary scalars: the Heidke skill score (HSS) and the True Skill Statistic (TSS). HSS is the recommended metric, which penalizes false alarms, stays defined at every operating point, matches the flare-forecasting literature, and preserves the accuracy-versus-event-skill ranking; TSS is retained for comparability with the operational and flare-forecasting literature and to exhibit its own degeneration, the gap between the two serving as direct evidence for the trap. The Symmetric Extremal Dependence Index (SEDI), for the extreme tail where HSS and TSS both fall toward zero, together with precision and frequency bias

make up the supporting panel. Reporting both per-window and per-day, with forecasts binarized and scored on the SWPC daily contingency unit following the Ali et al. [27] protocol, prevents temporal aggregation from masking the M-class result, for which the representative per-day HSS is ≈0.10 across folds.

Positioning against operational practice

Set against operational practice, the two tracks occupy different positions. No SXR flux-regression model in the benchmark reaches occurrence-persistence or the cited NOAA/SWPC operational skill at threshold crossings, whereas the proton regression track is competitive with operational proton forecasts. The comparison is indicative rather than definitive, because the operational numbers are cited at a daily, categorical framing and full alignment with the SWPC horizon and contingency unit remains pending. A concurrent effort by Jeong-Heon et al. [31], known to us only through a public Zenodo deposit while its manuscript is under review, is complementary rather than directly comparable: its predicted quantity is the Flare Irradiance Spectral Model 2 (FISM2) multi-wavelength irradiance reconstruction (with GOES XRS as an input, not the target), and it is organized around the post-peak decay phase of individual flare events rather than the continuous flux stream, so its evaluation protocol cannot be verified and no metric comparison is drawn.

Limitations

Several limitations bound the scope of the present results. The released dataset is a single reconstruction from NOAA GOES products, mirrored operationally at KSWRC, subject to inter-generation cross-calibration, data gaps, saturation of the largest flares, and scarcity of extreme events. In the present reconstruction, fill and saturation values are masked rather than imputed, and no explicit cross-generation rescaling is applied beyond the archived operational products; a formal cross-calibration continuity audit across the GOES-generation boundary is left to future work. The term-based rotating k-fold, while leakage-controlled, is interleaved rather than strictly forward. Because impulsive events are aperiodic rather than phase-locked to the solar cycle, a held-out term's events cannot be interpolated from neighbouring terms, so event prediction remains genuine forecasting; the continuous-flux envelope is autocorrelated across solar rotations and retains some interpolation structure, so a forward-holdout replication is warranted to bound the gap. Because such residual structure inflates apparent accuracy rather than deflating it, the finding that no model separates from persistence is conservative, and the leaderboard remains a per-split relative comparison applied identically to all ten models. The forecasts are point estimates rather than probabilistic distributions, and the operational head-to-head is not yet aligned to a common horizon and contingency unit. The qualification on the operational reference, in particular, means the proton-track competitiveness should be read as provisional until the operational head-to-head is aligned to the SWPC horizon and contingency unit.

Outlook

Two directions follow. First, probabilistic forecast heads would replace point estimates with calibrated distributions scored by the Brier score, reliability diagrams, and precision–recall area under the curve (PR-AUC), which are better matched to the rare-event structure than deterministic RMSE. Foundation-model time-series approaches and deep state-space models may extend to continuous flux regression. Foundation transformers have already been applied to GOES SXR for event classification [32], and deep state-space models to flare forecasting [33], so adapting either to the continuous regression target is a natural next step. Expectations should be tempered by Kieokaew et al. [34], where a zero-shot foundation model underperformed persistence on a related space-weather flux target.

Second, the benchmark is intended as a continually updated, community-facing challenge rather than a one-off comparison. The task definition, the leakage-controlled term-based split, and the evaluation harness are released publicly with a maintained leaderboard, so that external forecasts are scored under the identical protocol (the same atomic-term folds, historical horizons, forecast horizons, and imbalance-aware metrics) and remain free of the leakage that has confounded earlier flare-forecasting work; the Supplementary Information documents how to add a model to the harness. In the spirit of open time-series benchmarks such as GIFT-Eval, the SXR event-skill shortfall documented here (M-class HSS $\approx$0.10) is precisely the kind of open problem a shared leaderboard should attract from both the machine-learning and space-weather communities.

4 Methods

Data

The benchmark draws two operational GOES targets. The source products are the NOAA GOES soft X-ray and $\geq$10 MeV integral-proton flux, publicly archived at the National Centers for Environmental Information (NCEI) and National Geophysical Data Center (NGDC) and mirrored operationally in the KSWRC space-weather database; the released pipeline reconstructs the continuous benchmark time series from that KSWRC operational mirror of the public products. The soft X-ray (SXR) 1–8 Å channel supplies the `xrs_long` target at a native 1-minute cadence, spanning 2002-12-03 to 2026-06-25 (23.6 yr) with 11.48M primary-satellite samples, a 92.7% duty cycle relative to the 1-minute grid, and 98.5% valid-positive coverage. The $\geq$10 MeV integral proton flux supplies the `p_gt10` target at a 5-minute cadence, spanning 1999-02-27 to 2026-06-25 (27.3 yr) with 2.72M samples at a 94.8% duty cycle. Coverage spans roughly two-and-a-half solar cycles (Figure 5): the proton record (from 1999) captures the rise, maximum, and decline of solar cycle 23 (SC23), while the SXR record begins in the SC23 declining phase (2002); both continue through the full SC24 and the rise-to-maximum of SC25. Flux values are $\log_{10}$-transformed and resampled onto a common 5-minute grid used by every track, so all historical and forecast horizons are counted in 5-minute steps (288 steps = 1 day); the natively 1-minute SXR channel is aggregated to this grid.

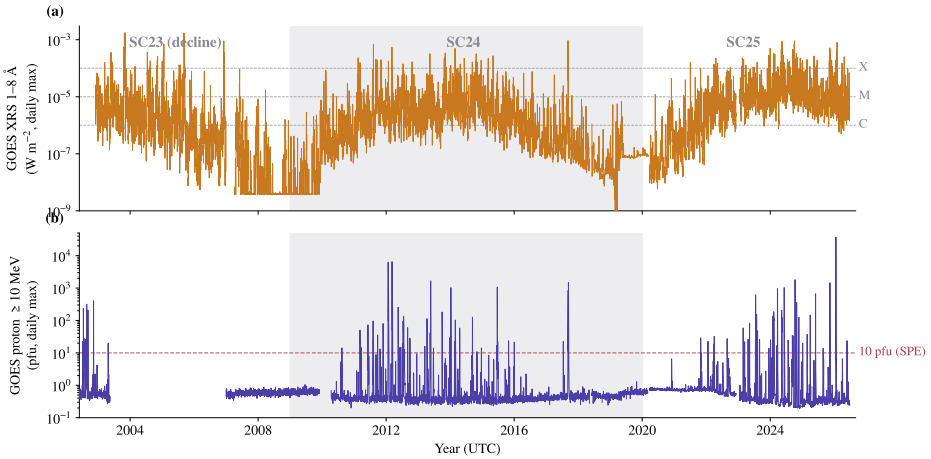


Fig. 5 Data overview. Daily-maximum GOES observations across the benchmark span. (a) The soft X-ray 1–8 Å flux (`xrs_long`), with the C, M, and X flare thresholds marked; (b) the ≥10 MeV integral proton flux (`p_gt10`), with the 10 pfu solar-proton-event threshold. Coverage runs from the declining phase of solar cycle 23, through the full cycle 24 (shaded), to the rise of cycle 25.

Split (leakage control)

Half-year `year_half` tagging (H1 = Jan–Jun, H2 = Jul–Dec) is followed by a rotating k-fold ($n = 5$): terms are sorted chronologically and interleaved with stride 5, so `val = terms[(3+i)%5::5]`, `test = terms[(4+i)%5::5]`, and the rest are train. Whole terms are atomic, so no input/label window crosses a split boundary and every fold spans all cycle phases. The split figure (Figure 6) sets the present protocol against three prior evaluation families for cyclic time series. Rolling-origin (walk-forward) cross-validation [35, 36] is applied in space-weather forecasting by Nishizuka et al. [37], Cao et al. [38], Nagashima and Sugiura [39], and Kieokaew et al. [34]. Expanding-window (accumulative) cross-validation [35, 40, 41] is a second family. The chronological or active-region-disjoint splits standard in astronomy and solar-flare forecasting [9, 42–46] are the third. The interleaved atomic-term rotation used here spans all solar-cycle phases in every fold while remaining leakage-controlled.

Models

Ten deep time-series models are evaluated: SegRNN [47], xPatch [48], iTransformer [49], PatchTST [50], TiDE [51], RLinear [52], PatchMixer [53], DLinear [28], TSMixer [54], and FReTS [55]. The set spans the segment-recurrent, patch-transformer, inverted-transformer, dense-MLP, linear, MLP-mixing, and frequency-domain families that define the current long-horizon forecasting literature.

The benchmark grid crosses three forecasting tracks, three historical horizons, and four forecast horizons, and every configuration is evaluated over five folds. The three tracks are the SXR-univariate track (`uni_b`), the proton-univariate track (`uni_a`), and the joint multivariate track (`multi`). The three historical horizons are 1, 3, and 7 days (288, 864, and 2016 five-minute steps), and the four forecast horizons are 3, 6,

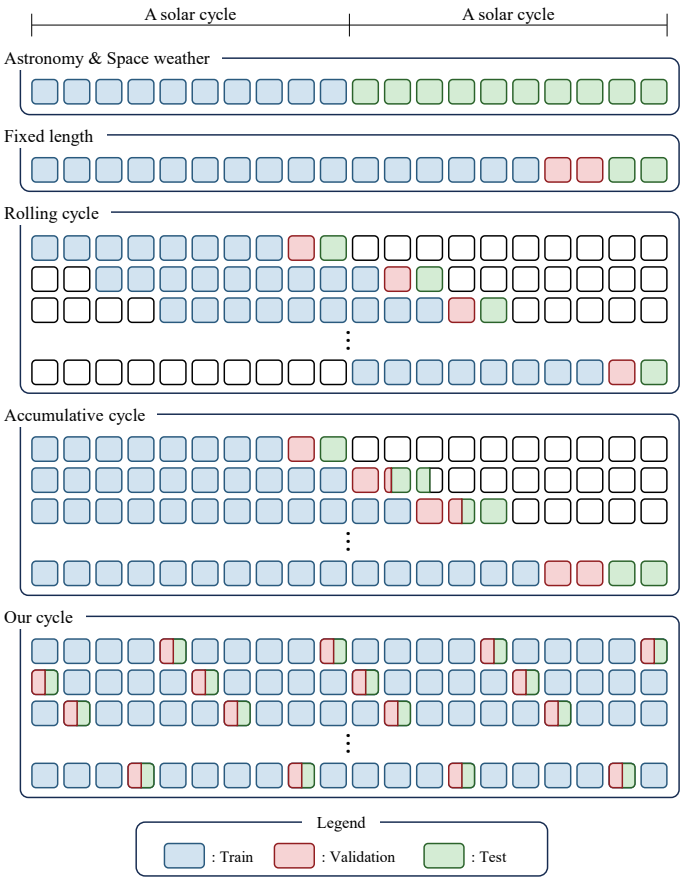


Fig. 6 Evaluation split protocols. Train (blue), validation (red), and test (green) assignment over two solar cycles for five schemes: the chronological astronomy/space-weather holdout, a fixed-length holdout, rolling-origin (walk-forward) cross-validation, expanding-window (accumulative) cross-validation, and the interleaved atomic-term rotation used here (bottom). Only the last keeps every fold's train, validation, and test sets spanning all cycle phases while remaining leakage-controlled.

12, and 24 h ahead (36, 72, 144, and 288 steps). Crossing the three tracks, the three historical horizons, the four horizons, and the five folds produces one hundred eighty configurations for every model.

The inputs are prepared through a fixed preprocessing pipeline. Each target series is transformed to base-10 logarithmic flux before modelling. Data gaps, missing samples, and saturation of the largest flares are handled during series preparation. Normalization statistics are fit on the training split alone and then applied unchanged to the validation and test splits, so that no information from the evaluation periods leaks into the inputs. Inputs and outputs are formed by sliding windows, with historical horizons of 1, 3, or 7 days and forecast horizons of 3, 6, 12, or 24 h.

All ten models share a single training configuration, so the leaderboard compares architectures under a common budget rather than individually tuned pipelines. Each

model minimizes the mean squared error on the $\log_{10}$-flux scale with the Adam optimizer (learning rate 10^{-3}, no weight decay) and a batch size of 64, under a PyTorch Lightning trainer on a single GPU. Training runs to at most 10,000 epochs with early stopping on the validation loss (patience 10 epochs, minimum improvement 10^{-6}), and the lowest-validation-loss checkpoint is retained. The shared capacity settings are a model dimension of 128, two layers, and a dropout of 0.1; no per-model hyperparameter search is performed.

The domain anchor is the operational NOAA/SWPC forecast, taken from the NCEI Report of Solar-Geophysical Activity (RSGA) daily-report archive (`YYYYMMDDRSGA.txt`, Part III Event Probabilities, 1966–2026) plus live SWPC feeds. Our forecasts are binarized and re-scored on SWPC's horizon and daily contingency unit (the Ali et al. [27] protocol); persistence baselines are trivial floors only. Class C is not archived in RSGA (M/X/proton only), and operational skill values are cited, not re-measured.

Metrics

Regression accuracy is RMSE and MAE on the log scale. Event skill is derived by binarizing the forecasts at the operational thresholds and scoring on the 2×2 contingency table. Event skill is reported as two complementary scalars, HSS (recommended; drawing on all four cells and penalizing false alarms) and TSS (the community-standard flare-forecasting score, retained for comparability and to exhibit its degeneration at low base rate), at per-window and per-day granularity on the SWPC daily contingency unit (the Ali et al. [27] protocol; rationale in Section 2.3). The remaining panel (the critical success index [CSI], Matthews correlation coefficient [MCC], SEDI, precision, and frequency bias) is tabulated in Supplementary Tables S6–S17, where SEDI, being base-rate-robust, carries the X-class extreme tail for which HSS and TSS both fall toward zero. Formal definitions of every score are collected in the Supplementary Information.

Ranking significance is assessed with a Friedman omnibus test on the fold-collapsed per-configuration RMSE, a Nemenyi critical-difference diagram, and Holm-corrected Wilcoxon signed-rank tests between the top-ranked model and each of the others over the complete-case configurations across the four horizons of the extended grid (the 1- and 3-day historical horizons, $N = 24$). The five folds are averaged within each configuration before ranking because the 120 per-fold evaluations are not independent: adjacent stride-5 folds share about two-thirds of their training terms, the horizons of a target are correlated, and the SXR channel is shared by the univariate and multivariate tracks; fold dispersion is instead carried by the 95% cross-fold intervals reported with every RMSE and event-skill table. Random seeds, had they been swept, would enter as within-configuration averaging on the same footing as the folds rather than as an additional blocking axis, since seed variance reports optimization stochasticity on a fixed split whereas the Friedman blocks are defined to sample generalization across distinct forecasting problems.

Per-day event-skill uncertainty is a nonparametric bootstrap over test days reported as 95% percentile intervals. Because a model and its persistence baseline are scored on the same folds and days, overlapping marginal intervals understate the paired

difference, so a non-overlap between a model and a baseline is read as a resolved separation while an overlap is treated as unresolved rather than as evidence of equality; the resolved per-day SXR separations against occurrence-persistence are additionally checked with a paired Wilcoxon signed-rank test over the five folds.

Dynamical-regime analysis

Two diagnostics characterize the dataset. Split balance (Supplementary Fig. S2) is the per-fold event count computed under the exact `term_split` fold indices. The complexity–entropy analysis (Figure 1) places each sliding window (length = a 1-day historical horizon at the 5-minute cadence, stepped by the 12-h forecast horizon and labelled quiet or event) on the complexity–entropy causality plane [56].

The permutation entropy H [57] is the Shannon entropy of the ordinal-pattern probability distribution P for embedding dimension $d = 5$ (delay $\tau = 1$), normalized by $\ln d!$ so that $H \in [0, 1]$. The statistical complexity C is the Martín–Plastino–Rosso form $C = H \cdot Q_J[P, P_e]$, the product of H with the normalized Jensen–Shannon divergence of P from the uniform distribution P_e. C vanishes for both a perfectly ordered and a perfectly random series and is maximal at intermediate structure, so each window falls within the theoretical $C_{\min}/C_{\max}$ envelope. Topological descriptors such as Euler-characteristic surfaces offer a complementary, information-theory-independent route to the same periodic-versus-chaotic distinction [58]. Both are evaluated on the raw log-flux window (ordinal statistics are invariant to monotone transforms) with `ordpy.complexity_entropy(·, dx=5)`; per-regime means are marked in Figure 1.

The forecastability panel (Supplementary Fig. S1) adds the ForeCA spectral predictability Ω [59] on the differenced log-flux, the weighted-minus-plain permutation-entropy gap [60], and the detrended-fluctuation scaling exponent α [61]. Implementations use the `ordpy`, `antropy`, and `nolds` libraries.

Supplementary information

The Supplementary Information, which contains Supplementary Figures S1–S6, Supplementary Tables S1–S17, and additional methods and analysis, is provided as a separate file.

Acknowledgements

The GOES soft X-ray and energetic-particle measurements analysed in this work are provided by the National Oceanic and Atmospheric Administration (NOAA). The science-quality records are archived and distributed by the NOAA National Centers for Environmental Information (NCEI), the operational streams are issued by the NOAA Space Weather Prediction Center (SWPC), and the working series were accessed through the operational mirror maintained by the Korea Space Weather Research Center (KSWRC) of the Korea Astronomy and Space Science Institute (KASI). NOAA GOES data are in the public domain. We also thank the authors of the benchmarked models for releasing the open-source reference implementations on which the leaderboard is built.

Declarations

- **Funding**. W. Jo and J.-H. Baek were supported by the Development of Adaptive Imaging Platform using Edge AI of the Korea Astronomy and Space Science Institute (KASI) grant funded by the Korea government (Ministry of Science and ICT, MSIT; 2026-1-810-01). E. Park was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (MoE; RS-2024-00337363).

- **Conflict of interest/Competing interests**. The authors declare that they have no competing interests.
- **Ethics approval and consent to participate**. Not applicable. The study uses only publicly archived geophysical measurements (GOES soft X-ray and $\geq$10 MeV proton flux) and involves no human participants, human data, or animals.
- **Consent for publication**. Not applicable.
- **Data availability**. The datasets generated and/or analyzed during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.22245285.
- **Materials availability**. Not applicable.
- **Code availability**. The benchmark code, the leakage-controlled term split, the model set, the scoring, the static leaderboard, and the data-reconstruction pipeline that rebuilds the targets from the KSWRC operational mirror of the public NOAA GOES products are available at https://github.com/merchen911/GOES-time-series under the Apache-2.0 license.
- **Author contribution**. W. Jo designed the benchmark, implemented the framework and ran the experiments, carried out the analysis, and drafted the manuscript. E. Park contributed to the data preparation and preprocessing and to the evaluation methodology, and reviewed and edited the manuscript. J.-H. Baek supervised the study, provided the computational resources and research environment, and reviewed and edited the manuscript. All authors read and approved the final version.
- **Use of generative AI**. During the preparation of this manuscript the authors used a generative AI assistant (a large language model) to improve the English-language expression of the text and to check the consistency of cross-references and citations between the main article and the Supplementary Information. The tool was not used to generate scientific content, data, analyses, results, or conclusions; the authors reviewed and verified all output and take full responsibility for the content of the publication.

References

[1] Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Ishii, M.: Deep flare net (defn) model for solar flare prediction. The Astrophysical Journal **858**(2), 113 (2018) https://doi.org/10.3847/1538-4357/aab9a7

[2] Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., Ishii, M.: Reliable probability forecast of solar flares: Deep flare net-reliable (defn-r). The Astrophysical Journal **899**(2), 150 (2020) https://doi.org/10.3847/1538-4357/aba2f2

[3] Bobra, M.G., Couvidat, S.: Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. The Astrophysical Journal **798**(2), 135 (2015) https://doi.org/10.1088/0004-637X/798/2/135

[4] Shao, M., Wang, H., Li, Y., Lin, J., Liu, J., Tan, B., Guo, J., Zhang, Y., Huang, J., Su, J., Sun, Y., Xu, H., Chen, J., Liu, S., Deng, Y., Tong, L., Bai, Y., Wang, C., Ji, K., Zhou, Y.: JW-Flare: Accurate Solar Flare Forecasting Method Based on Multimodal Large Language Models (2025). https://arxiv.org/abs/2511.08970

[5] Ferreira, J.S., Gradvohl, A.L.S., Antunes da Silva, A.E., Coelho, G.P., Cinto, T.: Short-term forecasting of medium- and high-intensity solar flares: A comparison between transformers and classical models. Advances in Space Research **77**(10), 10889–10905 (2026) https://doi.org/10.1016/j.asr.2026.03.067

[6] Shao, M., Liu, S., Xu, H., Jia, P., Wang, H., Tong, L., Bai, Y., Yang, C., Li, Y., Li, N., Lin, J.: Advances and challenges in solar flare prediction: a review. Astronomical Techniques and Instruments (2026) https://doi.org/10.3724/ati2025080

[7] Yi, K., Moon, Y.-J., Shin, G., Lim, D.: Forecast of major solar x-ray flare flux profiles using novel deep learning models. The Astrophysical Journal Letters **890**(1), 5 (2020) https://doi.org/10.3847/2041-8213/ab701b

[8] Dey, S., Fuentes, O.: Predicting solar x-ray flux using deep learning techniques. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–7 (2020). https://doi.org/10.1109/IJCNN48605.2020.9207284

[9] Sande, K., Muñoz-Jaramillo, A.: The struggles of developing an operational machine learning model for flare forecasting: Recasting the problem as a regression onto x-ray flux. The Astrophysical Journal **984**(1), 87 (2025) https://doi.org/10.3847/1538-4357/ad8de5

[10] Gao, Y., Zhang, L., Xu, L.: An intelligent solar flare prediction model based on x-ray flux curves using long short-term memory. Astronomical Techniques and Instruments **2**, 65–72 (2025) https://doi.org/10.61977/ati2024067

[11] Wei, M.: Prediction based on convolutional neural networks and vision transformer for goes-xrs solar flare time series. Academic Journal of Computing & Information Science **6**(8), 158–164 (2023) https://doi.org/10.25236/AJCIS.2023.060819

[12] Nedal, Mohamed, Kozarev, Kamen, Arsenov, Nestor, Zhang, Peijin: Forecasting solar energetic proton integral fluxes with bi-directional long short-term memory neural networks. J. Space Weather Space Clim. **13**, 26 (2023) https://doi.org/10.1051/swsc/2023026

[13] Yi, K., Shen, B., Li, Q., Wang, H., Moon, Y.-J., Lee, J., Lee, H.: Comparing lstm-based sequence-to-sequence forecasting strategies for 24-hour solar proton flux profiles using goes data. In: 2025 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1–6 (2025). https://doi.org/10.1109/ICDMW69685.2025.00225

[14] Yu, Y., Chen, Y., Zhao, L., Whitman, K., Manchester, W., Gombosi, T.: Realtime forecasting of solar energetic particle event and proton flux using multi-source solar observations and multi-task deep learning. Scientific Reports (2026) https://doi.org/10.1038/s41598-026-66110-2

[15] Stumpo, M., Laurenza, M., Benella, S., Marcucci, M.F.: Predicting the energetic proton flux with a machine learning regression algorithm. The Astrophysical Journal **975**(1), 8 (2024) https://doi.org/10.3847/1538-4357/ad7734

[16] Torres, J., Chan, P.K., Zhao, L., Zhang, M.: A machine learning approach to predicting sep proton intensity and events using time series of relativistic electron measurements. Space Weather **23**(2), 2024–003921 (2025) https://doi.org/10.1029/2024SW003921 . e2024SW003921 2024SW003921

[17] Landa, V., Reuveni, Y.: Low-dimensional convolutional neural network for solar flares goes time-series classification. The Astrophysical Journal Supplement Series **258**(1), 12 (2022) https://doi.org/10.3847/1538-4365/ac37bc

[18] Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., Shin, G.: Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms. The Astrophysical Journal **869**(2), 91 (2018) https://doi.org/10.3847/1538-4357/aaed40

[19] Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M.: Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. The Astrophysical Journal **835**(2), 156 (2017) https://doi.org/10.3847/1538-4357/835/2/156

[20] Bloomfield, D.S., Higgins, P.A., McAteer, R.T.J., Gallagher, P.T.: Toward reliable benchmarking of solar flare forecasting methods. The Astrophysical Journal Letters **747**(2), 41 (2012) https://doi.org/10.1088/2041-8205/747/2/L41

[21] Liu, H., Liu, C., Wang, J.T.L., Wang, H.: Predicting solar flares using a long short-term memory network. The Astrophysical Journal **877**(2), 121 (2019) https://doi.org/10.3847/1538-4357/ab1b3c

[22] Francisco, G., Berretti, M., Chierichini, S., Mugatwala, R., Fernandes, J., Barata, T., Del Moro, D.: Limits of solar flare forecasting models and new deep learning approach*. The Astrophysical Journal **985**(1), 108 (2025) https://doi.org/10.3847/1538-4357/adc56d

[23] Hong, J., Pandey, C., Aydin, B.: Uncertainty-Aware Solar Flare Regression (2026). https://arxiv.org/abs/2603.06712

[24] Roy, S., Hegde, D.V., Schmude, J., Lal, R., Gaur, V., Lin, A., Mandal, K., Singh, T., Muñoz-Jaramillo, A., Yang, K., Pandey, C., Hong, J., Aydin, B., McGranaghan, R., Kasapis, S., Upendran, V., Bahauddin, S., Silva, D., Freitag, M., Gurung, I., Pogorelov, N., Watson, C., Maskey, M., Bernabe-Moreno, J., Ramachandran, R.: Suryabench: Benchmark dataset for advancing machine learning in heliophysics and space weather prediction. Scientific Data **13**(1), 712 (2026) https://doi.org/10.1038/s41597-026-06552-5

[25] Abduallah, Y., Wang, J.T.L., Wang, H., Xu, Y.: Operational prediction of solar flares using a transformer-based framework. Scientific Reports **13**(1), 13665 (2023) https://doi.org/10.1038/s41598-023-40884-1

[26] Camporeale, E., Berger, T.E.: Verification of the NOAA space weather prediction center solar flare forecast (1998–2024). Space Weather **23**(10), 2025–004546 (2025) https://doi.org/10.1029/2025SW004546

[27] Ali, A., Sadykov, V., Kosovichev, A., Kitiashvili, I.N., Oria, V., Nita, G.M., Illarionov, E., O'Keefe, P.M., Francis, F., Chong, C.-J., Kosovich, P., Marroquin, R.D.: Predicting solar proton events of solar cycles 22–24 using goes proton and soft-x-ray flux features. The Astrophysical Journal Supplement Series **270**(1), 15 (2024) https://doi.org/10.3847/1538-4365/ad0a6c

[28] Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'23/IAAI'23/EAAI'23. AAAI Press, ??? (2023). https://doi.org/10.1609/aaai.v37i9.26317 . https://doi.org/10.1609/aaai.v37i9.26317

[29] Huang, J., Sun, Q., Zhang, X., Du, H., Xu, Z.: Decomposition-enhanced network for financial time series forecasting. Scientific Reports **16**, 23554 (2026) https://doi.org/10.1038/s41598-026-63484-1

[30] Wang, Y., Chen, X., Chen, J., Liu, M.: Ctmnet: causal trend evolution and adaptive modulation for time series forecasting. Scientific Reports **16**, 21169 (2026)

https://doi.org/10.1038/s41598-026-51651-3

[31] Jeong-Heon, K., Junmu, Y., Sung-Hong, P., Seungwoo, A., Young-Sil, K., Yong-Jae, M., P.C., C.: Jeongheon-Kim/2026SW_Solar Irradiance Prediction Model:Dataset. Zenodo (2026). https://doi.org/10.5281/zenodo.21400071

[32] Riggi, S., Romano, P., Pilzer, A., Becciani, U.: Solar flare forecasting with foundational transformer models across image, video, and time-series modalities. Astronomy and Computing **55**, 101042 (2026) https://doi.org/10.1016/j.ascom.2025.101042

[33] Takagi, Y., Nagashima, S., Sugiura, K.: Flare-ssm: Deep state space models with influence-balanced loss for 72-hour solar flare prediction. In: Taniguchi, T., Leung, C.S.A., Kozuno, T., Yoshimoto, J., Mahmud, M., Doborjeh, M., Doya, K. (eds.) Neural Information Processing, pp. 419–433. Springer, Singapore (2026)

[34] Kieokaew, R., Guezzi, R., Ginisty, F., Mariaccia, H.: Forecasting megaelectron-volt electron flux in the Earth's outer radiation belt using supervised machine learning algorithms and a timeseries foundation model (2026). https://arxiv.org/abs/2605.15752

[35] Tashman, L.J.: Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting **16**(4), 437–450 (2000) https://doi.org/10.1016/S0169-2070(00)00065-0 . The M3- Competition

[36] Bergmeir, C., Benítez, J.M.: On the use of cross-validation for time series predictor evaluation. Information Sciences **191**, 192–213 (2012) https://doi.org/10.1016/j.ins.2011.12.028 . Data Mining for Software Trustworthiness

[37] Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., Ishii, M.: Operational solar flare prediction model using deep flare net. Earth, Planets and Space **73**(1), 64 (2021) https://doi.org/10.1186/s40623-021-01381-9

[38] Cao, J., Xu, T., Deng, L., Wan, M., Zhao, X., Peng, Y., Zhou, X., Wang, W., Huang, L., Wang, W., Yang, L., Zhou, W.: Multi-source forecast of solar cycle flare activity using the novel informer-based models. Space Weather **23**(4), 2024–004322 (2025) https://doi.org/10.1029/2024SW004322 . e2024SW004322 2024SW004322

[39] Nagashima, S., Sugiura, K.: Deep space weather model: Long-range solar flare prediction from multi-wavelength images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)

[40] Hyndman, R., Athanasopoulos, G.: Forecasting: Principles and Practice, 3rd edn. OTexts, Australia (2021)

[41] Cerqueira, V., Torgo, L., Mozetič, I.: Evaluating time series forecasting models: an

empirical study on performance estimation methods. Machine Learning **109**(11), 1997–2028 (2020) https://doi.org/10.1007/s10994-020-05910-7

[42] Camporeale, E.: The challenge of machine learning in space weather: Nowcasting and forecasting. Space Weather **17**(8), 1166–1207 (2019) https://doi.org/10.1029/2018SW002061

[43] Angryk, R.A., Martens, P.C., Aydin, B., Kempton, D., Mahajan, S.S., Basodi, S., Ahmadzadeh, A., Cai, X., Filali Boubrahimi, S., Hamdi, S.M., Schuh, M.A., Georgoulis, M.K.: Multivariate time series dataset for space weather data analytics. Scientific Data **7**(1), 227 (2020) https://doi.org/10.1038/s41597-020-0548-x

[44] Wang, X., Chen, Y., Toth, G., Manchester, W.B., Gombosi, T.I., Hero, A.O., Jiao, Z., Sun, H., Jin, M., Liu, Y.: Predicting solar flares with machine learning: Investigating solar cycle dependence. The Astrophysical Journal **895**(1), 3 (2020) https://doi.org/10.3847/1538-4357/ab89ac

[45] Li, X., Zheng, Y., Wang, X., Wang, L.: Predicting solar flares using a novel deep convolutional neural network. The Astrophysical Journal **891**(1), 10 (2020) https://doi.org/10.3847/1538-4357/ab6d04

[46] Saini, K., Alshammari, K., Hamdi, S.M., Filali Boubrahimi, S.: Classification of major solar flares from extremely imbalanced multivariate time series data using minimally random convolutional kernel transform. Universe **10**(6) (2024) https://doi.org/10.3390/universe10060234

[47] Lin, S., Lin, W., Wu, W., Zhao, F., Mo, R., Zhang, H.: Segrnn: Segment recurrent neural network for long-term time-series forecasting. IEEE Internet of Things Journal **13**(5), 9861–9871 (2026) https://doi.org/10.1109/JIOT.2025.3647705

[48] Stitsyuk, A., Choi, J.: xpatch: Dual-stream time series forecasting with exponential seasonal-trend decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 20601–20609 (2025)

[49] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M.: iTransformer: Inverted Transformers Are Effective for Time Series Forecasting (2024). https://arxiv.org/abs/2310.06625

[50] Nie, Y., H. Nguyen, N., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: International Conference on Learning Representations (2023)

[51] Das, A., Kong, W., Leach, A., Mathur, S.K., Sen, R., Yu, R.: Long-term forecasting with tiDE: Time-series dense encoder. Transactions on Machine Learning Research (2023)

[52] Li, Z., Qi, S., Li, Y., Xu, Z.: Revisiting long-term time series forecasting: an

investigation on affine mapping. Academia AI and Applications **2**(2) (2026) https://doi.org/10.20935/acadai8236

[53] Gong, Z., Tang, Y., Liang, J.: PatchMixer: A Patch-Mixing Architecture for Long-Term Time Series Forecasting (2023). https://api.semanticscholar.org/CorpusID:263334059

[54] Chen, S.-A., Li, C.-L., Yoder, N., Arik, S.Ö., Pfister, T.: Tsmixer: An all-mlp architecture for time series forecasting. CoRR **abs/2303.06053** (2023)

[55] Yi, K., Zhang, Q., Fan, W., Wang, S., Wang, P., He, H., Lian, D., An, N., Cao, L., Niu, Z.: Frequency-domain mlps are more effective learners in time series forecasting. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Curran Associates Inc., Red Hook, NY, USA (2023)

[56] Rosso, O.A., Larrondo, H.A., Martin, M.T., Plastino, A., Fuentes, M.A.: Distinguishing noise from chaos. Phys. Rev. Lett. **99**, 154102 (2007) https://doi.org/10.1103/PhysRevLett.99.154102

[57] Bandt, C., Pompe, B.: Permutation entropy: A natural complexity measure for time series. Phys. Rev. Lett. **88**, 174102 (2002) https://doi.org/10.1103/PhysRevLett.88.174102

[58] Sharma, B.N., Mandal, V., Luwang, S.R., Nurujjaman, M., Mitra, A.J., Majhi, S.: Interpretable classification of time series using euler characteristic surfaces. Scientific Reports (2026) https://doi.org/10.1038/s41598-026-64177-5

[59] Goerg, G.: Forecastable component analysis. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 64–72. PMLR, Atlanta, Georgia, USA (2013). https://proceedings.mlr.press/v28/goerg13.html

[60] Fadlallah, B., Chen, B., Keil, A., Príncipe, J.: Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. Phys. Rev. E **87**, 022911 (2013) https://doi.org/10.1103/PhysRevE.87.022911

[61] Peng, C.-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L.: Mosaic organization of dna nucleotides. Phys. Rev. E **49**, 1685–1689 (1994) https://doi.org/10.1103/PhysRevE.49.1685

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ScientificReportsRSForecastingBenchmark2.pdf