

From Daily Hazard to Multi-Horizon Wildfire Risk: A Calibrated Spatiotemporal Transformer for the Maule Region, Chile

Efraín Campusano*, Marcela Quezada-Araya*, Hernán Astudillo, Carla Taramasco, Pablo Schwarzenberg, Paulo Quinsacara and Ernesto Vivanco

Universidad Andrés Bello, Viña del Mar, Chile

Abstract

Spatiotemporal estimates of wildfire occurrence probability across multiple lead times can support regional allocation of surveillance and preparedness resources. We evaluate a calibrated spatiotemporal Transformer for the Maule Region of Chile using daily wildfire records and predictors restricted to information available before each forecast date. The regional domain contains 27 active cells on an approximately 40 km grid, and each forecast origin uses the preceding 120 days. A temporal attention block summarizes the history within each cell and a spatial block subsequently exchanges information among cell representations. The network produces 20 daily lead probabilities, which are aggregated through a noisy-OR rule into probabilities of at least one recorded wildfire within the next 5, 10, 15, or 20 days; a final monotonicity constraint preserves the ordering of the nested horizons. Preprocessing, early stopping, calibration, and operating thresholds are fitted before the held-out period, and the last 30% of the chronology is used only for final assessment. Across the four horizons, balanced accuracy ranges from 0.7715 to 0.7823, recall from 0.7161 to 0.7923, and F1 from 0.5373 to 0.6727. Threshold-free AUPRC ranges from 0.5545 to 0.6678 and remains above the horizon-specific positive-rate reference at every horizon. Expected calibration error ranges from 0.0384 to 0.0852 and Brier score from 0.0977 to 0.1561. All horizons improve on a horizon-specific climatological reference in Brier skill (0.238–0.288). The maps show broadly similar high- and low-probability areas as the horizon expands. The results characterize the complete forecasting pipeline and are intended for regional prioritization rather than precise ignition localization; they do not establish superiority of the Transformer architecture over simpler alternatives.

Keywords

Wildfire occurrence forecasting, spatiotemporal Transformer, probabilistic forecasting, calibration, rare events, regional prioritization, Maule Region

1. Introduction

Wildfire prevention and preparedness require decisions before the location and timing of the next ignition are known. Surveillance teams, personnel, and response resources cannot be distributed uniformly across a large region every day, so a forecasting system is most useful when it supports spatial prioritization under uncertainty. A probability of occurrence preserves more decision information than a binary label because locations can be ranked while operational thresholds remain adaptable to the costs of false alarms and missed events. The Maule Region of Chile provides a demanding setting for this task. Central Chile experienced the prolonged 2010–2015 megadrought, whose effects on climate and vegetation have been extensively documented [1], and the 2017 fire season demonstrated how climatic stress, landscape configuration, and human influence can converge to produce extreme fire activity [2]. Studies focused on Maule have also identified substantial seasonal and spatial variation in fire occurrence and ignition probability [3, 4]. These characteristics motivate a dynamic forecast that is updated with the forecast date rather than a static susceptibility surface.

ICAIW 2026: Workshops at the 9th International Conference on Applied Informatics 2026, October 13–17, 2026, Orlando, United States

*Corresponding author.

✉ e.campusanomelendez@uandresbello.edu (E. Campusano); m.quezadaaraya@uandresbello.edu (M. Quezada-Araya); hernan.astudillo@unab.cl (H. Astudillo); carla.taramasco@unab.cl (C. Taramasco); pablo.schwarzenberg@unab.cl (P. Schwarzenberg); p.quinsacaracofre@uandresbello.edu (P. Quinsacara); e.vivancotapia@uandresbello.edu (E. Vivanco)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Wildfire modeling covers occurrence, danger, spread, and impact-related tasks [5]. Random forests and other nonlinear models have been used to describe spatial fire patterns [6], while recent work learns representations from temporal or remotely sensed data [7, 8, 9]. Transformer components have also appeared in wildfire spatiotemporal studies [10, 11]. Evaluation is difficult because daily cell-level occurrence is imbalanced: a model can attain high ordinary accuracy by predicting non-events while remaining poor at detecting events. We therefore report balanced accuracy together with precision, recall, specificity, and F1 [12, 13, 14]. Discrimination is evaluated separately from probability quality using Brier score, log-loss, expected calibration error (ECE), and validation-only post-hoc calibration [15, 16, 17, 18]. Similar calibration challenges have been reported for probabilistic fire-weather products [19].

A second methodological issue is temporal leakage. Incident records contain quantities such as burned area and damage that are known only during or after an event. Rolling or neighborhood summaries can also leak future information if their time support overlaps the prediction window. We therefore treat feature availability as part of the forecasting design. Each prediction uses the previous 120 completed days from 27 active cells and produces 20 daily leads that are aggregated into 5-, 10-, 15-, and 20-day occurrence probabilities. Scaling, early stopping, calibration, and threshold selection are completed without access to the final test block. Throughout the paper, *wildfire occurrence probability* is the primary quantity of interest. When the shorter term “risk” is used, it refers only to occurrence probability and not to expected loss, damage severity, exposure, or vulnerability. The present study evaluates the complete pipeline; because no same-split baseline or ablation is included, the reported results should not be interpreted as evidence that the Transformer or any individual attention block is superior to simpler alternatives.

The remainder of the article is organized as follows. Section 2 reviews wildfire occurrence forecasting, spatiotemporal modeling, and probability calibration. Section 3 describes the Maule panel, predictors, and chronological split. Section 4 presents the multi-horizon probability formulation, Section 5 the Transformer and training procedure, and Section 6 the calibration and evaluation protocol. Section 7 reports the held-out results, Section 8 discusses their interpretation, Section 9 addresses prospective reproducibility, Section 10 summarizes threats to validity, Section 11 states limitations and future work, and Section 12 concludes the paper.

2. Related Work

2.1. Wildfire occurrence modeling in Chile

Historical studies of the Maule Region show that wildfires do not occur uniformly across the territory or with the same frequency throughout the year [3]. In particular, Azócar de la Cruz et al. analyzed ignition risk during the period of the 2017 megafires and highlighted the importance of factors associated with human activity and land cover [4]. At the national level, the availability of consolidated wildfire data for Chile has strengthened the basis for conducting more consistent and reproducible analyses [20]. Our use of these records is different from a static susceptibility study: the quantity of interest is recalculated at each forecast date as the 120-day history moves forward.

2.2. Deep spatiotemporal wildfire prediction

Machine learning in wildfire research spans conventional regression, tree ensembles, support-vector methods, boosting, recurrent networks, convolutional models, and multimodal systems [5]. Examples include deep models for fire danger [7], next-day spread prediction from remote sensing [8], multimodal spread modeling [9], and architectures that combine spatiotemporal feature extraction with Transformer blocks [10]. Other studies have examined explainability in Mediterranean occurrence models [21] and the robustness of geospatial predictors assembled from heterogeneous sources [22].

The present target should not be confused with spread. We predict whether a cell will contain at least one recorded wildfire during a future interval. In that respect, the task is closer to work on probability-

of-fire forecasting [23] and data-driven fire-activity prediction [24]. The use of 10–20 day windows is also consistent with the broader planning motivation seen in probabilistic week-two fire-danger research [25].

2.3. Probability calibration across nested horizons

Calibration is performed after the predictive model has been fitted. Temperature scaling alters confidence while usually leaving the rank ordering unchanged [17]; isotonic regression instead learns a monotone non-parametric mapping [26]. We inspect calibration with ECE together with proper probability scores such as Brier score and log-loss [15, 18]. Multi-horizon products introduce one extra requirement. Because the windows are nested, the probability attached to 20 days should not fall below the probability attached to the corresponding 5 days. All four horizons are therefore derived from the same daily-lead sequence, and any crossing created by calibration is removed before the forecasts are reported.

3. Data and Forecast Design

3.1. Wildfire panel and grid definition

We assembled a daily forecasting panel from georeferenced wildfire records in the Maule Region between 29 September 2011 and 21 June 2021. The region is represented by a regular grid with approximately 40 km spacing. Twenty-seven cells have sufficient activity to be included in the final panel. For day t and cell g , $Y_{t,g}$ equals one when at least one wildfire is recorded in that cell and zero otherwise. Importantly, the zeroes are retained rather than discarded; the model therefore sees the natural imbalance of event and non-event day–cell observations. The 40 km grid limits the level of detail that can be extracted from the results. It can help identify areas of the region where wildfire risk may require closer attention, but it cannot pinpoint where an ignition is likely to occur. Factors such as road access, fuel breaks, and exposure of individual properties can be important at the local level, but the grid is not detailed enough to represent them. The results should therefore be interpreted as a regional overview rather than as a tool for tactical decisions on the ground.

3.2. Predictors and leakage control

The reported forecasting run uses 15 dynamic channels derived from past wildfire occurrence and its temporal, regional, and neighborhood context. In the exact order used to construct the input tensor, these channels are: (1) the historical cell-level occurrence indicator (`occ`); (2) normalized log time since the last recorded occurrence (`tsl_n`); (3–6) normalized local rolling occurrence rates over 7, 30, 14, and 90 days (`r7_n`, `r30_n`, `r14_n`, and `r90_n`); (7–8) sine and cosine encodings of day of year (`doy_sin`, `doy_cos`); (9–10) sine and cosine encodings of day of week (`dow_sin`, `dow_cos`); (11–13) normalized regional fire activity for the current day and its 7- and 30-day rolling sums (`glob_total_n`, `glob_r7_n`, `glob_r30_n`); and (14–15) mean occurrence and mean 7-day occurrence rate among neighboring cells (`nb_occ`, `nb_r7`). Thus, the model input used for the reported results is based on occurrence history, seasonality, regional activity, and neighboring-cell activity rather than contemporaneous post-event impact variables.

All channels are constructed using information that precedes the forecast origin. For every issue date t , the network receives the 120 completed days immediately before t . Rolling statistics are calculated from historical observations only, and neighborhood summaries exclude the target interval. Continuous input channels are standardized using parameters estimated exclusively from the training block; these parameters remain fixed during validation and testing. Burned-area totals, vegetation-specific damage, and other quantities generated during or after an event are not used by the reported forecasting run.

Table 1

Forecast unit, targets, and safeguards against using future information.

Item	Definition used in the experiment
Spatial unit	27 active cells on an approximately 40 km regional grid.
Historical input	120 completed days before the forecast date.
Daily output	Binary occurrence label for leads $d = 1, \dots, 20$.
Window output	Indicator of one or more recorded wildfires in the next 5, 10, 15, or 20 days.
Future-data guard	Current-event burned area/damage and any rolling or neighborhood feature touching the target window are excluded.
Time split	Earlier dates are used for development; the last 30% of the chronology is reserved for testing.
Calibration policy	Probability mappings and action thresholds are estimated from validation predictions.

3.3. Chronological split

Time order is preserved throughout the split. The earliest 70% of dates is the development block. Its final 20% serves as validation, while the earlier portion is used to fit model parameters. The remaining 30% of the full chronology is not used during development and becomes the test period. We did not use random cross-validation because neighboring dates from the same fire season could then appear on both sides of a split, giving the evaluation information flow that would not exist in a real forecast.

The archived run used 1,060 forecast dates in the held-out test block. With 27 active cells, this gives 28,620 day–cell observations per horizon. The number of positive windows is 4,414 (15.42%) at 5 days, 6,280 (21.94%) at 10 days, 7,402 (25.86%) at 15 days, and 8,236 (28.78%) at 20 days. These counts explain the increasing prevalence as the forecast window expands.

4. Probabilistic Formulation

For each cell g and forecast date t , the network produces one probability for every lead $d \in \{1, \dots, 20\}$,

$$q_{t,g,d} = \sigma(z_{t,g,d}), \quad (1)$$

where $z_{t,g,d}$ is the logit for lead d . These outputs are trained with the corresponding future daily occurrence labels. Operationally, however, the question is usually whether any event will occur within a window of length h . We form that quantity as

$$P_{t,g}^{(h)} = 1 - \prod_{d=1}^h (1 - q_{t,g,d}), \quad h \in \{5, 10, 15, 20\}. \quad (2)$$

Under a sequential conditional-hazard interpretation, or under conditional independence of the daily events, Eq. (2) has an exact probabilistic interpretation. Otherwise we use it as an operational noisy-OR approximation. Either way, it gives the uncalibrated windows the desired nesting property. Independent calibration mappings can disturb that order, so the final outputs are projected back to

$$P_{t,g}^{(5)} \leq P_{t,g}^{(10)} \leq P_{t,g}^{(15)} \leq P_{t,g}^{(20)}. \quad (3)$$

No ordering violations remain in the test predictions after this step.

5. Spatiotemporal Transformer

5.1. Time-then-space architecture

At a given forecast date, each cell is represented by a $120 \times q$ matrix. Daily feature vectors are projected to 192 dimensions and augmented with learned encodings for temporal position and grid location.

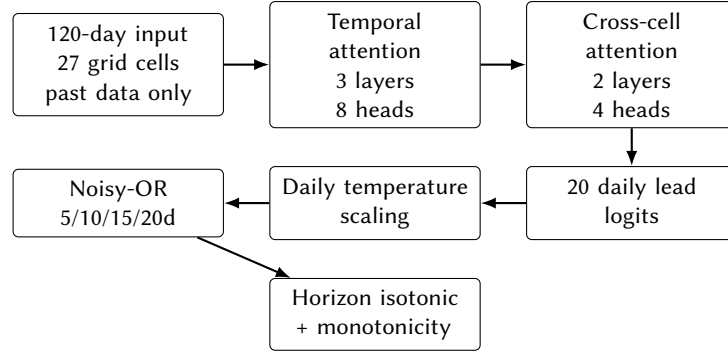


Figure 1: Model flow used for the reported run. Daily logits are temperature-scaled on validation, aggregated into four windows, and then corrected with horizon-specific isotonic regression and a final monotonicity constraint.

The first Transformer sees only the 120-day trajectory of one cell at a time. After that temporal block, each cell is summarized by a single representation. Those 27 representations then become the token sequence for a second Transformer that mixes information across cells. This ordering keeps temporal memory and regional interaction separate instead of applying one large attention operation over every day–cell pair.

The experiment reported here uses three temporal layers with eight heads per layer, followed by two spatial layers with four heads. Feed-forward layers have dimension 512 and dropout is 0.20. The output head applies layer normalization and returns 20 logits, one per lead day. Figure 1 gives the sequence of operations from historical input to calibrated windows.

5.2. Optimization objective and training

The loss was chosen to address both the rarity of positive lead days and the quality of the window probabilities. A focal-weighted binary term handles the daily labels, while a Brier-style term is evaluated on the four aggregated windows:

$$\mathcal{L} = 0.85 \ell_{\text{focal}} + 0.15 \ell_{\text{Brier}}. \quad (4)$$

We use $\gamma = 2$ for the focal component. Optimization uses AdamW with learning rate 2×10^{-4} and weight decay 2×10^{-4} ; the warm-up fraction is 0.08, dropout remains 0.20, and gradients are clipped at 1.0. Training is allowed to run for at most 60 epochs with 300 steps per epoch. Each batch contains 10 forecast days and gradients are accumulated over two batches. The random seed is 123. All tables, figures, and test metrics reported in this paper correspond to this single training run. Validation performance was best at epoch 31, and early stopping ended training at epoch 41. Because initialization sensitivity is not quantified with repeated seeds, the results are interpreted as a single-run characterization of the pipeline rather than an estimate of across-seed variability.

6. Calibration and Evaluation

All calibration parameters are estimated from validation output only. The reported run uses a two-stage procedure. First, temperature scaling [17] is fitted separately to the logits of each of the 20 daily lead times on the validation block. The temperature-scaled daily probabilities are then aggregated with the noisy-OR formulation in Eq. (2) to obtain the 5-, 10-, 15-, and 20-day window probabilities. Second, a horizon-specific isotonic regression [26] is fitted on the resulting validation window probabilities. Because independently calibrated horizons can cross, a final monotonic projection enforces $P_{t,g}^{(5)} \leq P_{t,g}^{(10)} \leq P_{t,g}^{(15)} \leq P_{t,g}^{(20)}$. The fitted temperature parameters, isotonic mappings, and operating thresholds are frozen before the test labels are evaluated. The ECE, Brier score, log-loss, mean probability, and Brier skill reported below correspond to these final calibrated and monotonic test probabilities.

Table 2

Architecture and training settings for the reported run.

Setting	Value
Grid / history	27 cells / 120 days
Daily leads / windows	1–20 / 5, 10, 15, 20 days
Latent / feed-forward size	192 / 512
Temporal block	3 Transformer layers, 8 heads
Spatial block	2 Transformer layers, 4 heads
Dropout	0.20
Optimization	AdamW; LR 2×10^{-4} ; WD 2×10^{-4}
Training length	at most 60 epochs; best epoch 31; stopped at epoch 41
Objective	0.85 focal daily loss + 0.15 Brier window term

We treat threshold-based performance and probability quality as separate aspects of the forecast. After applying the threshold selected on validation, we report balanced accuracy, precision, recall, specificity, and F1. Separately, ECE, Brier score, and log-loss evaluate the numerical probabilities. Brier score is also compared with a constant horizon-specific climatological forecast. If the empirical event rate for horizon h is π_h , then the climatological Brier score is $BS_{\text{clim}} = \pi_h(1 - \pi_h)$ and skill is

$$BSS_h = 1 - \frac{BS_{\text{model},h}}{\pi_h(1 - \pi_h)}. \quad (5)$$

A positive BSS means that the model reduces squared probability error relative to assigning the same base-rate probability to every day–cell example.

For the map-based analysis, calibrated probabilities are averaged by cell over the held-out dates. Each horizon has its own probability distribution, so the four panels use their respective 5th and 95th percentiles as display limits. The maps are therefore used for a descriptive comparison of the locations of higher- and lower-probability cells, not as a quantitative test of spatial rank stability.

6.1. Why several metrics are needed

No single number answers all of the evaluation questions in this experiment. Prevalence increases with window length, and the forecast itself is meant to be read as a probability. Let TP , TN , FP , and FN denote the confusion-matrix counts obtained after applying the threshold fixed on validation. We use

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad (6)$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad \text{F1} = \frac{2 \text{ Precision Recall}}{\text{Precision} + \text{Recall}}, \quad (7)$$

and compute balanced accuracy as the mean of recall and specificity. Ordinary accuracy is less useful here because a large number of non-event cells can dominate the count.

Those classification metrics describe what happens after a probability has been converted into an action. They do not establish that the probability scale itself is reliable. For N binary test cases with forecasts p_i and outcomes y_i , we also report

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2, \quad (8)$$

and

$$LL = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]. \quad (9)$$

Brier score is a squared error on the probability scale and can be compared directly with climatology. Log-loss is more severe when an incorrect forecast is made with high confidence. ECE provides an additional perspective on model calibration by comparing predicted probabilities with the frequencies of events actually observed across different reliability bins. However, its value can vary depending on how these bins are defined. For this reason, ECE is used as a complementary diagnostic measure rather than as a single criterion for evaluating the quality of the predicted probabilities.

This distinction is important from a practical perspective. A model may correctly rank two cells according to their probability of event occurrence while systematically overestimating both probabilities. In that case, the ranking could still be useful for deciding where to focus surveillance or allocate resources first, but a nominal “30%” threshold would no longer have its expected meaning: events would not actually occur with a frequency close to 30%. The opposite situation can also occur. A model may produce probabilities that are well calibrated overall and correctly reflect the average frequency of events, yet assign very similar values to all cells and therefore provide little ability to distinguish which locations have a higher probability of occurrence.

For this reason, we evaluate three aspects separately: first, whether the model can distinguish between periods or cells in which an event occurs and those in which it does not; second, whether the estimated probabilities reliably represent the observed frequencies; and third, whether a threshold defined before evaluating the test set provides a reasonable operational trade-off between the different decisions that may be made.

6.2. Operating thresholds fixed before testing

The test block is opened only after the choices that can adapt to outcomes have been made. This includes the scaling constants, the model checkpoint selected by early stopping, the calibration map, the nesting correction, and one decision threshold for each horizon. The reported threshold is consequently part of the frozen forecasting policy, not a cutoff selected retrospectively to make the test confusion matrix look favorable.

Different operations may still choose different cutoffs. A surveillance task can accept more false alarms if doing so improves recall; an intervention that mobilizes expensive field resources may prefer a stricter rule. We report one validation-selected threshold per horizon so that the results can be reproduced. The continuous probabilities are retained, so an agency can later define another action policy without retraining the network, provided the new policy is tuned without consulting the held-out outcomes.

The reported quantities answer complementary questions. Balanced accuracy, precision, recall, specificity, and F1 summarize the fixed validation-selected decision policy. AUPRC provides a threshold-free summary of precision–recall discrimination; for each horizon, the no-skill reference is the corresponding positive rate. Brier score, log-loss, ECE, and mean forecast probability versus prevalence describe probability quality. Brier skill compares the model with horizon-specific climatology, the monotonicity check verifies consistency of nested windows, and the maps provide a descriptive view of regional prioritization. Because nearby forecast dates and cells are not statistically independent and no block-bootstrap confidence intervals are reported, differences among horizons are interpreted descriptively rather than as statistically significant differences.

7. Results

7.1. Discrimination in the rare-event setting

Table 3 summarizes the fixed-threshold results on the held-out chronology. Balanced accuracy is similar across the four windows: 0.7715 at 5 days, 0.7746 at 10 days, 0.7763 at 15 days, and 0.7823 at 20 days. Recall moves upward from 0.7161 to 0.7923 as the horizon grows, while specificity decreases from 0.8268 to 0.7724. Precision rises from 0.4299 to 0.5845, and F1 follows from 0.5373 to 0.6727.

Table 3

Fixed-policy discrimination on the held-out chronology.

Hor.	Thr.	Bal.	Prec.	Rec.	Spec.	F1
5d	0.2519	0.7715	0.4299	0.7161	0.8268	0.5373
10d	0.3339	0.7746	0.5250	0.7365	0.8127	0.6130
15d	0.3667	0.7763	0.5614	0.7597	0.7930	0.6457
20d	0.3790	0.7823	0.5845	0.7923	0.7724	0.6727

Table 4

Probability diagnostics and climatological Brier skill on the test period.

Hor.	ECE	Brier	LogL	Mean p	Base	Clim. BS	BSS
5d	0.0384	0.0977	0.3598	0.1775	0.1542	0.1304	0.2509
10d	0.0529	0.1220	0.4331	0.2505	0.2194	0.1713	0.2876
15d	0.0670	0.1403	0.4903	0.3081	0.2586	0.1917	0.2682
20d	0.0852	0.1561	0.5379	0.3570	0.2878	0.2050	0.2384

The rise in precision and F1 should not be interpreted as a pure model improvement, because the target changes with the window. A 20-day interval simply provides more opportunities for at least one fire to be recorded than a 5-day interval; the positive rate correspondingly grows from 0.1542 to 0.2878. The nearly unchanged balanced accuracy is therefore the more useful cross-horizon observation. The validation-selected cutoffs also increase with horizon, which argues against reusing one threshold for all four products.

7.2. Calibration and skill relative to climatology

Calibration changes more noticeably with horizon than discrimination does. The 5-day product has ECE 0.0384 and Brier score 0.0977; at 20 days those values are 0.0852 and 0.1561 (Table 4). The average forecast is above the observed prevalence in every window. The absolute gaps are 0.0233, 0.0311, 0.0495, and 0.0692 from 5 to 20 days. Thus the longer-horizon maps remain useful for ordering cells, but their numerical probabilities are progressively more optimistic.

The climatological comparison puts that calibration error in perspective. Brier skill remains positive at every horizon: 0.2509, 0.2876, 0.2682, and 0.2384 for 5, 10, 15, and 20 days. Expressed another way, the spatiotemporal forecast reduces squared probability error by about 24–29% relative to a forecast that assigns the same horizon-specific prevalence to every case. The largest reduction occurs for the 10-day window.

7.3. Threshold-free precision–recall discrimination

Because wildfire occurrence is imbalanced, we also report the area under the precision–recall curve (AUPRC) for each horizon. The horizon-specific positive rate is the appropriate no-skill reference for AUPRC. Table 5 therefore reports both quantities and their ratio. AUPRC is above prevalence for every forecast window, with values from 0.5545 at 5 days to 0.6678 at 20 days. The corresponding lift over the no-skill reference is $3.60\times$, $2.91\times$, $2.57\times$, and $2.32\times$, respectively. Absolute AUPRC values should not be compared across horizons without accounting for the increasing positive rate.

After calibration and the nesting correction, none of the test cases violates the required horizon order. This check may appear minor, but it prevents an operationally confusing output such as reporting a smaller chance of at least one fire in 20 days than in the 5-day interval contained within it.

Table 5
Threshold-free precision–recall discrimination on the held-out test period. The positive rate is the no-skill AUPRC reference.

Hor.	AUPRC	Base rate	PR lift
5d	0.5545	0.1542	3.60×
10d	0.6392	0.2194	2.91×
15d	0.6647	0.2586	2.57×
20d	0.6678	0.2878	2.32×

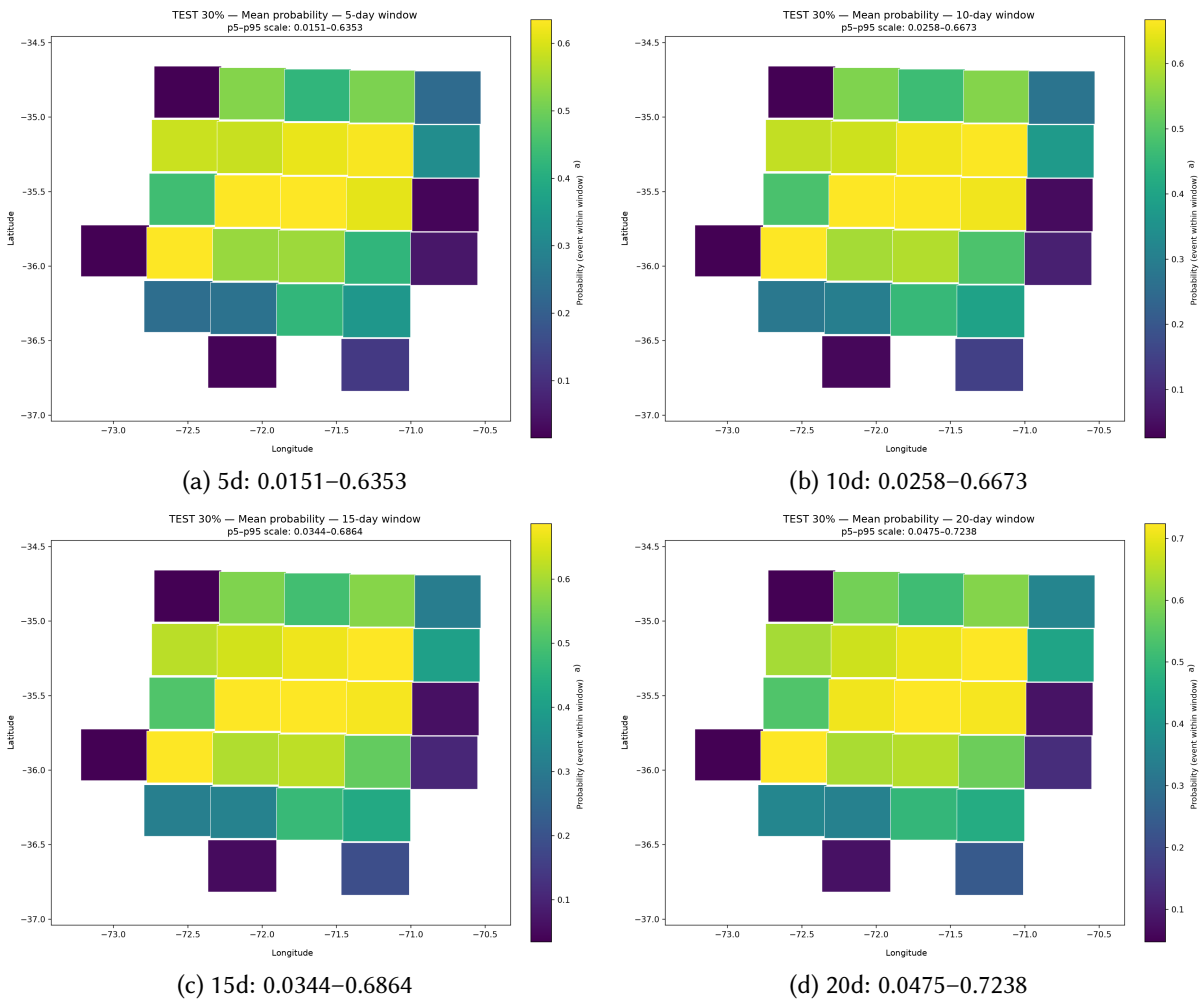


Figure 2: Cell-wise mean of the calibrated forecasts over the held-out dates. Values under the panels give the 5th–95th percentile display limits for that horizon. Since the scales differ, the figure is used to compare the location of high and low cells rather than equal color values across horizons.

7.4. Spatial pattern of the probability forecasts

Figure 2 shows the cell-wise mean probability from the calibrated predictions on the test set. Predicted probabilities increase as the horizon expands, as expected for nested “at least one event” windows. The 5th–95th percentile interval increases from 0.0151–0.6353 at 5 days to 0.0475–0.7238 at 20 days. Visual inspection indicates broadly similar high- and low-probability areas across the four panels: several cells that are high at 5 days remain comparatively high at longer horizons, while peripheral low-probability cells remain evident. This observation is descriptive. Because no rank-correlation statistic is reported, we do not interpret the maps as quantitative evidence of invariant spatial ordering.

These maps are screening products. They do not claim to identify the coordinates of a future ignition.

Table 6

Practical reading of the four forecast windows.

Hor.	Likely use	What is strongest	What requires caution
5d	Immediate surveillance	Lowest ECE and highest specificity.	Recall is below the longer windows.
10d	Short-range planning	Largest Brier-skill improvement over climatology.	The accumulated probability is already higher because the window is longer.
15d	Medium-range prioritization	Recall and F1 increase while balanced accuracy remains stable.	Mean probability shows a larger positive bias.
20d	Strategic readiness	Highest recall and F1 at the frozen cutoff.	Highest ECE; best treated mainly as a ranking product.

A cell that repeatedly receives a higher probability can be given additional attention, but local decisions still require information at finer scales, including current weather forecasts, fuel conditions, access, active incidents, and expert judgment.

7.5. Operational role of each forecast window

The four horizons represent different planning products rather than interchangeable versions of the same forecast. At 5 days, ECE is lowest (0.0384) and specificity is highest (0.8268), making the shortest horizon the most reliable of the four when the probability value itself is important. At 10 days, Brier skill reaches its maximum (0.2876), providing the largest reduction in squared probability error relative to climatology. At 15 days, recall and F1 increase, but the mean predicted probability exceeds observed prevalence by about 0.05. At 20 days, recall, precision, F1, and balanced accuracy are highest at the frozen operating point, while ECE is also highest (0.0852) and the positive probability bias is largest. The longer horizons are therefore more appropriate for prioritization and preparedness than for interpreting each numerical probability as a precise event frequency. Table 6 summarizes this operational reading.

Figure 2 is a diagnostic summary over the held-out period rather than the map an operator would consult on a specific day. In prospective use, the map would be regenerated from the newest 120-day history, and the forecast date, horizon, calibration version, and threshold policy should be stored with the output so that the issued forecast can later be reconstructed.

8. Discussion

The held-out evaluation shows a consistent discrimination profile across the four forecast windows, while calibration changes more strongly with horizon. Balanced accuracy remains close to 0.78, but longer windows contain more positive cases because they provide more opportunity for at least one wildfire to occur. For that reason, the increase in precision, recall, and F1 with horizon should not be interpreted as a pure improvement in model quality. The probability diagnostics provide a complementary view: ECE and the gap between mean forecast probability and observed prevalence increase toward 20 days, so longer-horizon probabilities require more cautious numerical interpretation even though their Brier skill remains positive relative to climatology.

The leakage controls are central to this result. Incident databases contain detailed descriptions of what happened during and after a fire, but consequence variables are unsuitable as contemporaneous predictors for pre-ignition forecasting. Excluding burned-area and damage fields, shifting rolling and neighborhood summaries backward, and fitting scaling, calibration, and thresholds without the final test block makes the retrospective protocol closer to information flow in prospective operation.

The maps provide a regional screening view rather than tactical localization. Their high- and low-probability areas appear broadly similar across horizons, but this is a visual description rather than

a measured rank-correlation result. Local decisions would still require finer-scale weather, fuel state, access, active incidents, exposed assets, and expert judgment.

We do not construct a numerical league table against earlier wildfire studies because targets, spatial units, predictor sets, geography, and event prevalence differ substantially. More importantly, the present manuscript does not contain a same-split baseline or architecture ablation. The contribution should therefore be read as an evaluation of a complete leakage-aware, calibrated, multi-horizon forecasting pipeline, not as evidence that temporal attention, cross-cell attention, or the Transformer class is necessary or superior. Controlled comparisons with simpler models remain future work.

9. Reproducibility in Prospective Operation

For an operational model, reproducibility means more than being able to rerun the training code. A forecast for a particular date should be reconstructable from the information that was genuinely available on that date. We therefore keep raw observations, derived historical features, fitted preprocessing parameters, model weights, calibration objects, and action thresholds as distinct artifacts. The retrospective experiment is operationally meaningful only if the same sequence can be repeated later without consulting future fields.

9.1. Daily forecasting sequence

A daily run starts after observation day t has closed. The pipeline reconstructs the 15-channel tensor from the previous 120 completed days, using cell-level occurrence history, rolling occurrence rates, seasonal encodings, regional activity, and neighboring-cell activity. The preprocessing parameters fitted during development are then applied without refitting. The temporal block produces one representation per cell and the spatial block exchanges information among the 27 cell representations. The network emits 20 daily logits. The frozen lead-specific temperature parameters are applied first, the daily probabilities are aggregated by noisy-OR into the four forecast windows, the frozen horizon-specific isotonic mappings are applied, and the final monotonicity constraint guarantees the nested ordering. Horizon-specific thresholds are applied only after these probability corrections. Both continuous probabilities and thresholded decisions are stored.

This order is part of the model specification. Temperature scaling is defined on the daily logits, whereas isotonic calibration is defined on the aggregated window probabilities. Rearranging these operations would create a different forecasting policy and would require a new validation record.

9.2. Checking temporal data integrity

The reported channels are derived from the historical occurrence panel and calendar, regional, and neighborhood context. Temporal integrity therefore requires more than excluding explicit future labels. For each forecast origin, the 120-day history must end before the target window; local rolling rates must not include target-period observations; regional activity summaries must be based only on completed days; and neighboring-cell features must be calculated from the same historical support. Standardization parameters are estimated from the training block only and are not recomputed on validation or test data.

This distinction is especially important when working with the broader wildfire database, which also contains post-event information such as burned area and damage. Such variables can be valuable for retrospective impact analysis but are not part of the 15-channel forecasting tensor reported here. Any future extension that introduces an external data source should preserve the same issue-time principle: a value is admissible only if it would have been available when the forecast was issued.

9.3. Versioning and calibration drift

Calibration can change over time. The relationship between a model score and the observed frequency of an event may shift, for example, as drought conditions, fuel characteristics, land use, prevention policies, reporting practices, or fire suppression strategies change. For this reason, we keep track of the feature schema, preprocessing settings, Transformer weights, calibration function, and thresholds used for each forecasting horizon. Each forecast should record which versions of these components were used to produce it.

Model performance can be checked at two different stages. Before issuing a forecast, simple checks can identify problems in the input data, including missing values, unrealistic ranges, or distributions that differ substantially from those seen during training. Later, once the actual outcomes are available, model performance can be reviewed to see whether its probabilities remain reliable and whether it can still distinguish between cells with and without events. If the main problem is calibration, updating the calibration function may be enough. If the model also becomes worse at ranking higher- and lower-probability cases, the underlying relationship between the predictors and event occurrence may have changed. In that situation, retraining the model should be considered.

Model and policy changes should also be recorded separately. An agency may raise or lower a threshold because staffing or intervention costs have changed even though the probability model is identical. On the other hand, a recalibrated model may keep the same action cutoff. Keeping those decisions separate avoids attributing an apparent improvement to a hidden retuning of the operational rule.

At minimum, prospective forecasts should version five artifacts separately: the feature schema, preprocessing parameters, Transformer weights, calibration object, and horizon-specific threshold policy. Silent changes to any of these components can alter the forecasting policy even when the remaining components are unchanged. Keeping them separate also distinguishes a model update from a policy change such as raising or lowering an operational threshold.

9.4. Transfer across regions and grid scales

The architecture can be retrained in another region, but the fitted Maule weights should not be treated as geographically universal. Fire prevalence, vegetation composition, weather regimes, and human ignition patterns differ across Chile. A transfer study would need to distinguish no-update transfer, probability recalibration with frozen network weights, and local fine-tuning or retraining.

Spatial scale creates a related trade-off. Smaller cells improve localization but sharply reduce the positive rate per cell, while larger cells increase event frequency at the cost of blending distinct local processes. A tactical grid near 5 km would therefore be a different learning problem rather than a simple resampling of the present maps. The current focal-weighted objective would need to be re-tuned for the stronger imbalance, and training might require class-aware sampling or weighting, minimum-data rules for active cells, and/or a hierarchical or multiscale representation that shares information across neighboring fine cells. Calibration and operating thresholds would also have to be estimated again at the new scale. These are design requirements for a future fine-grid system, not results established by the present 40 km experiment.

10. Threats to Validity

The chronological split and feature-availability rules reduce future-data leakage, but retrospective environmental databases can still contain later revisions that are difficult to identify from stored timestamps alone. The target is deliberately narrow: at least one recorded wildfire in a grid cell during a future window. It does not measure fire size, spread, exposure, damage, economic loss, or vulnerability. Evidence is specific to Maule, the historical period studied, and the approximately 40 km grid. Nearby dates and cells are not independent, and the paper reports no block-bootstrap confidence intervals; cross-horizon differences are consequently descriptive. Finally, the study evaluates a single Transformer

run with seed 123 and does not establish that this architecture is uniquely responsible for the reported performance.

11. Limitations and Future Work

The principal methodological limitation is the absence of same-split reference models. A simpler nonlinear classifier or a recurrent sequence model might reach similar probability quality, and the present results do not isolate the contribution of the temporal block, cross-cell attention, or the 120-day lookback. Future work should therefore compare classical baselines, LSTM/TCN-style sequence models, shorter lookbacks, and time-only versus time-then-space variants under exactly the same chronology, calibration protocol, and held-out test period. Such comparisons should retain threshold-free rare-event measures such as AUPRC and evaluate them under the same frozen chronology and calibration protocol.

Other extensions include finer or multiscale spatial representations, formal feature attribution, repeated-seed analysis, block-bootstrap uncertainty over test dates, seasonal recalibration, and transfer to other Chilean regions. A prospective trial in which every forecast is archived before its outcome is known would provide the strongest test of delayed data availability, database revisions, drift, and operational reliability.

12. Conclusions

We evaluated a calibrated multi-horizon wildfire occurrence forecasting pipeline for the Maule Region of Chile. Each forecast uses the previous 120 days from 27 regional cells. A temporal attention stage is followed by cross-cell attention, and 20 daily leads are aggregated into nested 5-, 10-, 15-, and 20-day occurrence probabilities. The final 30% of the chronology remains outside model fitting, calibration, and threshold selection.

On the held-out period, balanced accuracy remains close to 0.78 across the four horizons, and recall increases from 0.7161 to 0.7923. AUPRC ranges from 0.5545 to 0.6678 and exceeds the horizon-specific positive-rate reference at every window. Brier skill is positive relative to horizon-specific climatology for every window. Calibration is strongest at 5 days and becomes weaker as the horizon expands. The mean maps show broadly similar higher- and lower-probability areas across horizons, but this spatial comparison is descriptive rather than a formal rank-stability test. The final probabilities satisfy the required nesting order for every test case.

The intended use is regional prioritization and preparedness, not precise ignition localization. Because the paper contains no same-split baseline, ablation study, repeated-seed analysis, or confidence intervals, its conclusions are deliberately limited to the performance of the complete reported pipeline. Controlled comparisons with simpler models and alternative sequence architectures are the next step before making claims about architectural advantage.

Declaration on Generative AI

No generative artificial intelligence tools were used to generate, modify, optimize, debug, execute, or validate any of the computational code employed to obtain the experimental results reported in this article. All data preprocessing, model implementation, training, calibration, evaluation, metric computation, and figure generation associated with the reported results were carried out through the authors' computational workflow using conventional software libraries. Consequently, none of the numerical results, experimental outputs, or model predictions presented in this work were produced through generative AI-assisted programming.

References

- [1] R. D. Garreaud, C. Alvarez-Garreton, J. Barichivich, J. P. Boisier, D. Christie, M. Galleguillos, C. LeQuesne, J. McPhee, M. Zambrano-Bigiarini, The 2010–2015 megadrought in central chile: Impacts on regional hydroclimate and vegetation, *Hydrology and Earth System Sciences* 21 (2017) 6307–6327. doi:10.5194/hess-21-6307-2017.
- [2] D. M. J. S. Bowman, G. J. Williamson, J. T. Abatzoglou, C. A. Kolden, M. A. Cochrane, A. M. S. Smith, Human–environmental drivers and impacts of the globally extreme 2017 chilean fires, *Ambio* 48 (2019) 350–362. doi:10.1007/s13280-018-1084-1.
- [3] I. Díaz-Hormazábal, M. E. González, Spatio-temporal analysis of forest fires in the maule region, chile, *Bosque* 37 (2016) 147–158.
- [4] G. A. de la Cruz, G. Alfaro, C. Alonso, et al., Modeling the ignition risk: Analysis before and after megafire on maule region, chile, *Applied Sciences* 12 (2022) 9353. doi:10.3390/app12189353.
- [5] P. Jain, S. C. P. Coogan, S. G. Subramanian, M. Crowley, S. Taylor, M. D. Flannigan, A review of machine learning applications in wildfire science and management, *Environmental Reviews* 28 (2020) 478–505. doi:10.1139/er-2020-0019.
- [6] S. Oliveira, F. Oehler, J. San-Miguel-Ayanz, A. Camia, J. M. C. Pereira, Modeling spatial patterns of fire occurrence in mediterranean europe using multiple regression and random forest, *Forest Ecology and Management* 275 (2012) 117–129. doi:10.1016/j.foreco.2012.03.003.
- [7] S. Kondylatos, I. Prapas, M. Ronco, I. Papoutsis, G. Camps-Valls, M. Piles, M.-Á. Fernández-Torres, N. Carvalhais, Wildfire danger prediction and understanding with deep learning, *Geophysical Research Letters* 49 (2022) e2022GL099368. doi:10.1029/2022GL099368.
- [8] F. Huot, R. L. Hu, N. Goyal, T. Sankar, M. Ihme, Y.-F. Chen, Next day wildfire spread: A machine learning data set to predict wildfire spreading from remote-sensing data, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022) 1–13. doi:10.1109/TGRS.2022.3192974.
- [9] D. Shadrin, S. Illarionova, F. Gubanov, et al., Wildfire spreading prediction using multimodal data and deep neural networks, *Scientific Reports* 14 (2024) 2606.
- [10] X. Chen, Y. Tian, C. Zheng, X. Liu, Autost-net: A spatiotemporal feature-driven approach for accurate forest fire spread prediction from remote sensing data, *Forests* 15 (2024) 705. doi:10.3390/f15040705.
- [11] J. Zhu, et al., Unveiling spatiotemporal patterns of wildfire risk: A transformer-based earth system analysis, *Climate Dynamics* 63 (2025) 21. doi:10.1007/s00382-024-07481-y.
- [12] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, *PLOS ONE* 10 (2015) e0118432. doi:10.1371/journal.pone.0118432.
- [13] H. He, E. A. Garcia, Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering* 21 (2009) 1263–1284. doi:10.1109/TKDE.2008.239.
- [14] P. Branco, L. Torgo, R. P. Ribeiro, A survey of predictive modeling on imbalanced domains, *ACM Computing Surveys* 49 (2016) 31:1–31:50. doi:10.1145/2907070.
- [15] G. W. Brier, Verification of forecasts expressed in terms of probability, *Monthly Weather Review* 78 (1950) 1–3.
- [16] T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* 102 (2007) 359–378. doi:10.1198/016214506000001437.
- [17] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2017, pp. 1321–1330.
- [18] T. Dimitriadis, T. Gneiting, A. I. Jordan, Stable reliability diagrams for probabilistic classifiers, *Proceedings of the National Academy of Sciences* 118 (2021) e2016191118. doi:10.1073/pnas.2016191118.
- [19] S. Bohlmann, M. Laine, et al., Statistical calibration of probabilistic medium-range fire weather index forecasts in europe, *Natural Hazards and Earth System Sciences* 24 (2024) 4225–4243. doi:10.5194/nhess-24-4225-2024.

- [20] C. Vidal-Silva, R. Pizarro, M. Castillo-Soto, C. de la Fuente, V. Duarte, C. Sanguesa, A. Ibañez, R. Paredes, B. Ingram, Wildfire occurrence and damage dataset for chile (1985–2024): A real data resource for early detection and prevention systems, *Data* 10 (2025) 93. doi:10.3390/data10070093.
- [21] R. Cilli, M. Elia, M. D’Este, et al., Explainable artificial intelligence detects wildfire occurrence in mediterranean southern europe, *Scientific Reports* 12 (2022) 16349.
- [22] S. Illarionova, D. Shadrin, F. Gubanov, et al., Geospatial data and machine learning for robust wildfire occurrence prediction, *Scientific Reports* 15 (2025) 10712.
- [23] J. R. McNorton, F. D. Giuseppe, E. Pinnington, M. Chantry, C. Barnard, A global probability-of-fire forecast, *Geophysical Research Letters* 51 (2024) e2023GL107929. doi:10.1029/2023GL107929.
- [24] F. D. Giuseppe, J. R. McNorton, A. Lombardi, F. Wetterhall, Global data-driven prediction of fire activity, *Nature Communications* 16 (2025) 2918.
- [25] R. P. Worsnop, M. Scheuerer, F. D. Giuseppe, et al., Probabilistic fire-danger forecasting for week-two forecasts, *Weather and Forecasting* 36 (2021) 2113–2125.
- [26] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 694–699. doi:10.1145/775047.775151.