Estimating Wildfire Occurrence Probability in High-Risk Zones: The Maule Region, Chile

Efraín Campusano
Facultad de Ingeniería
Universidad Andres Bello
Santiago, Chile
e.campusanomelendez@uandresbello.edu

Paulo Quinsacara
Facultad de Ingeniería
Universidad Andres Bello
Santiago, Chile
p.quinsacarajofr@uandresbello.edu

Pablo Schwarzenberg
Facultad de Ingeniería
Universidad Andres Bello
Santiago, Chile
pablo.schwarzenberg@unab.cl

Carla Taramasco
Facultad de Ingeniería
Universidad Andres Bello
Viña del Mar, Chile
carla.taramasco@unab.cl

Hernán Astudillo
Facultad de Ingeniería
Universidad Andres Bello
Viña del Mar, Chile
hernan.astudillo@unab.cl

***Abstract*—Wildfires have become an increasingly recurrent hazard in central Chile, producing severe environmental and socioeconomic impacts and motivating probabilistic, spatially explicit tools for prevention and operational planning. Focusing on the Maule Region, this study estimates wildfire occurrence probability in high-risk zones using a spatiotemporal modeling approach that captures both local environmental conditions and their temporal evolution. We compile and harmonize a gridded dataset combining wildfire occurrence records with dynamic drivers (e.g., meteorological variables and vegetation-related proxies) and static descriptors (e.g., topography), and we formulate risk estimation as a sequence-to-probability task at the grid-cell level. We develop a Transformer-based spatiotemporal architecture that learns nonlinear dependencies across time and space and outputs calibrated probabilities of wildfire occurrence for multiple forecast windows. Performance is assessed on held-out periods using discrimination and calibration analyses, and the predicted probabilities are converted into risk-zonation maps to support consistent operational thresholding. Results show that the proposed model produces coherent probability fields that concentrate risk in historically fire-prone areas under conducive conditions, while delivering stable probabilistic outputs suitable for decision-making. The resulting risk maps provide actionable guidance for prioritizing surveillance and allocating prevention and response resources in the Maule Region, and the framework is readily transferable to other regions where wildfire risk exhibits strong spatiotemporal variability.**

***Index Terms*—Wildfire forecasting, Wildfire risk mapping, Spatiotemporal Transformer, Probabilistic forecasting, Probability calibration, Maule Region**

I. Introduction

Wildfires have become a major socio-environmental hazard in south-central Chile, where prolonged drought, extreme heat and wind events, and highly flammable plantation landscapes can align to produce fast-spreading, high-impact fire outbreaks. Maule is one of the most affected regions, in terms of burned area and fire impacts, and several studies have highlighted the roles of weather extremes, land-use patterns, and human drivers in shaping ignition and spread dynamics. Public agencies in Chile maintain operational risk products (e.g., risk categories for the wildland urban interface and ignition-risk alerts) that support prevention and territorial management, but these products typically need complementary, data-driven approaches that can deliver local, spatio-temporally explicit and probabilistic estimates with transparent performance characterization.

Most wildfire risk mapping and prediction studies in Chile have focused on ignition-risk modeling or susceptibility estimation using statistical learning and geospatial predictors, including recent work explicitly targeting the Maule region. While these approaches are valuable, there remains a gap between static or coarse-resolution risk products and multi-horizon probabilistic forecasts that update daily and explicitly leverage spatiotemporal dependence (persistence, seasonality, and neighborhood effects) in fire occurrence. From an operational standpoint, probabilities are most useful when they are not only accurate but also well calibrated, so that a predicted probability (e.g., 0.30) corresponds to an empirical frequency close to 30% under similar conditions.

To address this gap, we frame wildfire risk estimation in Maule as a grid-based spatiotemporal forecasting problem. We use geo-referenced incident records compiled for this study covering 29 Sep 2011 to 21 Jun 2021 (daily resolution), and discretize the territory into fixed cells to obtain a binary spatiotemporal process of wildfire occurrence per day and grid cell. We then train a spatio-temporal Transformer model, building on attention-based sequence modeling to estimate daily hazard probabilities for future lead times, and derive window probabilities for horizons (5, 10, 15, and 20 days) in a way that is coherent across horizons. Because decision-making depends strongly on probability quality, we incorporate post-hoc calibration strategies (temperature scaling and monotonic isotonic mapping) on a validation set, and report both discrimination and calibration metrics.

The main contributions of this work are: (i) a daily, gridded formulation of Maule wildfire occurrence suitable for multi-horizon risk estimation; (ii) a Transformer-based spatiotemporal model that outputs coherent probabilities for multiple future windows; and (iii) a calibration-and-evaluation protocol emphasizing decision-relevant probabilistic performance.

The remainder of this article is structured as follows: Section II surveys related work on wildfire risk mapping, spatiotemporal deep learning, probabilistic forecasting, and calibration; Section III introduces grid-based probabilistic formulation, coherent multi-horizon window probabilities, Transformer-based sequence modeling, calibration, and evaluation under class imbalance; Section IV describes the Maule wildfire dataset, input variables, feature construction, chronological train validation test split, model training, calibration strategy, and interpretability considerations; Section V presents the predictive performance from a wildfire risk perspective and discusses the mean risk maps over the held-out test period: Section VI summarizes the main limitations related to transferability, non-stationarity, interpretability, and operational use; and Section VII summarizes and concludes.

II. Related work

Wildfire risk estimation has been studied from complementary perspectives that differ in both the prediction target and the temporal scale. One line of work focuses on *susceptibility* or *ignition risk* mapping, aiming to estimate the likelihood of fire occurrence as a function of human activity, fuels/land cover, and topography. Another line addresses *dynamic forecasting*, ranging from near-real-time spread prediction to short-term (days) and seasonal outlooks, often leveraging weather and climate covariates. Our work lies at the intersection of these strands: it estimates spatio-temporally varying probabilities on a regular grid, and reports coherent multi-horizon risk windows that can support operational decision-making.

A. Ignition risk mapping and spatial susceptibility models

Spatial ignition risk models have commonly relied on statistical learning over historical ignition locations with covariates describing land cover, distance to infrastructure, population density, and terrain. In Maule, Azócar de la Cruz et al. [1] developed ignition risk maps and analyzed changes before and after the 2017 megafire episode, reporting strong contributions from anthropogenic and land-cover variables and illustrating how risk patterns can vary across seasons and land-use contexts. These studies provide important baselines for identifying persistent hotspots, but they typically (i) treat time coarsely (e.g., seasonal aggregation), or (ii) output risk scores that are not directly interpretable as well-calibrated probabilities for specific forward-looking horizons.

From a data perspective, Chile has also seen progress toward consolidated, reusable wildfire databases. Vidal-Silva et al. [2] introduced a national dataset spanning multiple decades with occurrence and damage attributes, supporting early detection and prevention research. These resources enable evaluation under realistic class imbalance and facilitate reproducible comparisons across modeling paradigms.

B. Spatiotemporal deep learning for wildfire modeling

Deep learning approaches for wildfire-related prediction tasks include CNN/LSTM hybrids, temporal convolution models, and attention-based architectures designed to fuse spatial context with temporal memory. Transformers [3] have become a central building block for long-range dependency modeling and are increasingly adopted in spatiotemporal settings. In remote-sensing-based spread prediction, for instance, AutoST-Net [4] combines 3D convolutions with Transformer components to capture local and global spatiotemporal features in fire spread imagery. While spread prediction differs from occurrence/hazard forecasting, the broader message is that attention mechanisms can effectively represent multi-scale spatiotemporal dependencies when sufficient data and careful inductive biases are provided (e.g., spatial tokens, positional/time embeddings, and exogenous signals).

In operational risk contexts, a key challenge is aligning model outputs with decision-relevant horizons. Many studies report single-step probabilities or categorical risk classes, whereas early warning systems often require *multi-horizon* risk estimates (e.g., 5/10/15/20 days) that remain internally consistent. This motivates modeling strategies that learn daily hazards and then aggregate them into window probabilities using probabilistic identities, avoiding ad-hoc post-processing that may distort calibration.

C. Probabilistic forecasting, calibration, and evaluation

For risk monitoring, probability quality is as important as discrimination. Modern neural networks can be miscalibrated, producing overconfident probabilities, which in turn can lead to unstable thresholding policies. Temperature scaling [5] is a widely used post-hoc calibration technique that adjusts logit sharpness while preserving ranking. For non-parametric calibration, isotonic regression [6] provides a monotone mapping from scores to probabilities and has been used effectively in classification settings. To quantify calibration, expected calibration error (ECE) and reliability diagrams are commonly used, with recent work [7] discussing stability and interpretation of reliability diagnostics under finite samples.

In environmental forecasting, calibration is also an active topic for probabilistic fire-danger products. For example, Bohlmann *et al.* [8] study statistical calibration of probabilistic medium-range Fire Weather Index (FWI) forecasts, illustrating practical benefits and methodological considerations for operational workflows. These results support the inclusion of explicit calibration stages in wildfire risk pipelines, especially under severe class imbalance and non-stationarity.

Summarizing, prior work motivates three design priorities for grid-based wildfire risk estimation in regions like Maule: (i) spatio-temporal architectures that capture cross-grid interactions and temporal persistence, (ii) multi-horizon risk outputs that are coherent by construction, and (iii) explicit calibration and probabilistic evaluation to ensure decision-ready probabilities.

III. Background

A. Grid-based spatiotemporal risk as probabilistic forecasting

Let the study area be discretized into a regular grid of G cells. For each day t, we define a binary variable $Y_{t,g} \in \{0,1\}$ indicating whether at least one wildfire occurred in cell g on day t. Given an input history window of length L (e.g., the previous L days), the goal is to estimate *probabilities* of future occurrence, not only hard labels. We use a *daily hazard* formulation:

$$p_{t,g,d} \;=\; \mathbb{P}(Y_{t+d,g}=1 \,|\, \mathcal{H}_{t,g}), \qquad d=1,\ldots,H_{\max}, \quad (1)$$

where $\mathcal{H}_{t,g}$ is the information available up to day t (historical activity and covariates on the grid).

Operationally, decision-making often uses *window probabilities* (e.g., the probability of at least one event within the next 5, 10, 15, or 20 days). If $p_{t,g,d}$ is interpreted as the conditional probability of an event on day $t+d$, then the probability of at least one event within a window of length h can be derived by the standard "OR" construction:

$$P_{t,g}^{(h)} \;=\; 1-\prod_{d=1}^{h}\bigl(1-p_{t,g,d}\bigr). \quad (2)$$

Equation (2) yields *coherent* multi-horizon probabilities by construction: if $h_1 < h_2$, then $P_{t,g}^{(h_1)} \leq P_{t,g}^{(h_2)}$.

B. Transformers for sequence modeling

Transformers were introduced as attention-based architectures that model dependencies in sequences without recurrent connections [3]. Given token embeddings, self-attention builds context-aware representations by weighting interactions between all token pairs. In time-series settings, Transformers can represent long-range temporal structure and nonlinear interactions through stacked attention and feed-forward blocks, while positional encodings (or learned embeddings) provide temporal order information.

C. Spatiotemporal Transformers

For gridded environmental risk, the signal is both temporal (seasonality, persistence, recent activity) and spatial (neighborhood dependence, correlated conditions). A common strategy is to factorize representation learning into (i) a temporal encoder that processes the past L days for each grid cell, and (ii) a spatial encoder that models interactions among cells at the prediction time. This "time-then-space" design is compatible with attention mechanisms and supports adding spatial context via learned grid embeddings and/or geographic coordinates.

D. Probability calibration

Even when a model ranks risk well, its raw probabilities can be miscalibrated (over/under-confident), which harms threshold-based decisions. Temperature scaling is a simple and effective post-hoc calibration method that rescales logits by a positive scalar T to adjust confidence while largely preserving ranking [5]. In multi-lead forecasting, calibration can be applied *per lead day* to account for changing base rates and uncertainty with horizon. As a more flexible monotone mapping, isotonic regression can transform predicted probabilities into better-calibrated estimates under minimal assumptions [6]. Calibration quality is commonly summarized with reliability diagnostics and scalar metrics such as ECE, alongside proper scoring rules like log-loss and the Brier score [7], [9].

E. Evaluation under class imbalance and operational thresholds

Wildfire occurrence at daily–grid resolution is typically sparse, producing severe class imbalance. For probabilistic models, evaluation should include both (i) discrimination (e.g., how well higher probabilities align with events) and (ii) probability quality (log-loss, Brier score, calibration). In operational settings, a practical approach is to select thresholds on a validation split that satisfy constraints such as a minimum precision (e.g., precision ≥ 0.80) and then report the associated recall and confusion-matrix metrics on a held-out test period. This bridges probabilistic forecasting with decision policies that prioritize false-alarm control.

IV. Experiments

A. Dataset and study design

We use georeferenced wildfire occurrence records for the Maule Region, Chile, covering 29 Sep 2011 to 21 Jun 2021 at daily resolution. The study area is discretized into a regular grid of 40 km cells, and each day-cell pair is labeled $Y_{t,g} \in \{0,1\}$ indicating whether at least one wildfire occurred in cell g on day t. Model inputs are built as sequences of length $L = 120$ days per cell. The model outputs daily hazard probabilities for lead days $d = 1,\ldots,20$, and window probabilities for horizons $h \in \{5,10,15,20\}$ are computed using Eq. (2).

B. Input variables and feature construction

The input representation combines wildfire occurrence history with environmental and territorial descriptors that are relevant to ignition and short-term hazard variation. The dynamic variables include meteorological conditions such as temperature, relative humidity, wind speed, and wind direction. These variables are used because they directly affect fuel drying, fire spread potential, and the short-term favorability of ignition conditions. Vegetation and land-cover descriptors include plantation-related variables, natural vegetation classes, agricultural surfaces, waste areas, and total affected or exposed area. Static terrain descriptors include slope, aspect/exposition, and topographic class, which provide information about spatial heterogeneity in fire behavior and landscape exposure.

All variables are harmonized at the daily grid-cell level before sequence construction. Continuous variables are normalized using parameters estimated exclusively from the training period, while categorical or discrete descriptors are encoded consistently across train, validation, and test periods. This procedure prevents temporal leakage and ensures that the validation and test splits simulate a prospective forecasting setting. Table I summarizes the main variable groups and the final modeling configuration used in the experiments.

C. Temporal split

We apply a chronological split to prevent leakage: the first 70% of the available timeline is used for train and validation, and the final 30% is used for testing. Validation is taken as the last 20% of the train and validation block. All normalization parameters are computed on the training period only.

D. Model and training

The final model follows a time-then-space Transformer design. First, a temporal attention encoder processes the 120-day sequence associated with each grid cell, allowing the model to learn persistence, seasonality, and delayed effects in the local risk signal. Second, a spatial attention encoder models interactions among grid cells, allowing the prediction for a given cell to be informed by regional spatial structure and neighboring risk patterns. The output layer produces daily hazard logits for lead days $d = 1, \ldots, 20$, which are subsequently transformed into calibrated daily probabilities and coherent window probabilities.

The training objective combines a daily binary classification loss over the 20 lead days with a Brier-type regularization term over the derived window probabilities. This design encourages both discrimination and probability quality. Because wildfire occurrence is sparse at the daily grid-cell level, mini-batches are balanced by sampling temporal instances with and without at least one event in the next 20 days. This sampling strategy reduces the dominance of non-event cases during optimization while preserving chronological separation between train, validation, and test periods.

The Transformer model was selected as the main architecture because the forecasting task involves three coupled sources of complexity: temporal dependence across the 120-day input sequence, nonlinear interactions among meteorological, vegetation, topographic, and historical-occurrence variables, and spatial dependence among grid cells. In this context, simpler models such as logistic regression and random forest are useful reference baselines for tabular or static risk estimation, but they do not explicitly model sequential multi-lead dynamics unless additional lagged features are manually engineered. Similarly, recurrent models such as LSTM networks are natural temporal baselines because they can represent sequential dependence, but they do not directly provide the same cross-cell attention mechanism used here to represent spatial interactions among grid locations.

For these reasons, the present study focuses on the Transformer-based probabilistic formulation and its calibrated multi-horizon outputs. A systematic comparison with logistic regression, random forest, and LSTM baselines is considered an important ablation study for future work, particularly to quantify the marginal contribution of temporal attention, spatial attention, and coherent probability aggregation.

E. Calibration and evaluation

Calibration is performed on the validation split using temperature scaling per lead day and, optionally, isotonic regression per lead day. The calibrated daily hazards are converted to coherent window probabilities. We evaluate on the held-out test period using log-loss, Brier score, and ECE (reliability bins), and report confusion-matrix metrics using thresholds selected on validation. For simplicity and comparability, we summarize results by reporting only the average performance for the four operational horizons 5, 10, 15, and 20 days.

F. Interpretability and operational transparency

Although the Transformer model is not intrinsically interpretable in the same way as linear models or tree-based feature-importance methods, operational transparency is addressed in three ways. First, the input variables are grouped into physically meaningful categories: weather, vegetation/land cover, topography, and historical occurrence. Second, the model outputs calibrated probabilities rather than only binary labels, which allows users to inspect continuous risk gradients and define decision thresholds according to operational constraints. Third, the predicted probabilities are summarized as spatial risk maps across multiple horizons, making it possible to verify whether the model concentrates risk in plausible and historically fire-prone areas.

A more detailed attribution analysis, for example using permutation importance, SHAP values, or attention-based diagnostics, is an important extension for future work. Such an analysis would help quantify the relative contribution of meteorological, vegetation, and topographic drivers and would improve the interpretability of the model.

V. Wildfire Risk Interpretation of Results

This section interprets the predictive results from a wildfire risk perspective, emphasizing what the reported probabilistic metrics imply for operational use. All results are summarized at the four decision horizons of 5, 10, 15, and 20 days, using window probabilities derived from calibrated daily hazard estimates.

A. Overall predictive skill under class imbalance

Daily wildfire occurrence at grid resolution is a sparse-event problem, so raw accuracy can be misleading. For this reason, we emphasize balanced accuracy, which averages sensitivity (recall) and specificity and therefore reflects performance on both event and non-event cases. Across the four horizons, balanced accuracy is consistently close to 0.78, ranging from 0.7703 (5 days) to 0.7846 (15 days), with an average of 0.7794. This stability indicates that the model maintains a robust separation between higher-risk and lower-risk grid-days even as the forecast window expands.

The stability of balanced accuracy is explained by a relatively stable trade-off between recall and specificity. Recall increases from 0.7121 at 5 days to values around 0.79 at longer horizons, while specificity remains between 0.7692 and 0.8286. In practical terms, the model detects a substantial

TABLE I: Summary of input variable groups and final modeling configuration.

Component	Description	Role in the model
Wildfire occurrence history	Daily binary occurrence per grid cell, indicating whether at least one wildfire was recorded in the corresponding cell and day.	Temporal target/history
Meteorological variables	Temperature, relative humidity, wind speed, and wind direction. These variables represent short-term atmospheric conditions associated with fuel dryness and fire spread potential.	Dynamic covariates
Vegetation and land-cover descriptors	Pine plantations by age class, eucalyptus, other plantations, wooded areas, scrub, pastureland, agricultural surfaces, waste areas, and aggregated surface categories.	Spatial/environmental covariates
Topographic descriptors	Slope, exposition/aspect, and topographic class. These variables describe terrain-related spatial heterogeneity.	Static covariates
Spatial discretization	Regular grid of 40 km cells over the Maule Region.	Prediction unit
Input sequence length	$L = 120$ previous days per grid cell.	Temporal context
Forecast structure	Daily hazards for lead days $d = 1, \ldots, 20$, aggregated into 5, 10, 15, and 20 days window probabilities.	Multi-horizon output
Train/validation/test split	Chronological split: first 70% for train/validation and final 30% for testing; validation corresponds to the last 20% of the train/validation block.	Leakage prevention
Calibration	Temperature scaling per lead day, with optional isotonic regression per lead day, fitted on validation predictions.	Probability correction

fraction of event windows while still correctly rejecting most non-event windows, which is essential for surveillance prioritization where false alarms must be controlled but missed events are also costly.

B. Precision, thresholds, and decision-oriented reporting

Because operational actions have different costs, we report results under a validation-selected threshold that targets decision stability. Precision ranges from 0.4310 (5 days) to 0.5849 (20 days), and F1 increases accordingly from 0.5370 to 0.6727. The increase in precision and F1 with horizon is expected because the base rate of observing at least one event in the window rises as the window length increases (from 0.1542 at 5 days to 0.2878 at 20 days). Importantly, the balanced accuracy does not degrade as base rate increases, indicating that the model is not merely benefiting from more frequent positives, but preserves a consistent discrimination ability across horizons.

C. Discrimination versus probability quality

In addition to threshold-based metrics, discrimination measures (AUROC and AUPRC) are strong across horizons, with AUROC around 0.85 to 0.87 and AUPRC around 0.55 to 0.67. However, for risk monitoring, probability quality matters because probabilities are directly consumed for ranking, mapping, and thresholding. The reported Brier scores decrease slightly with horizon (0.2424 to 0.2289), consistent with improved mean squared probability error as the event becomes less rare. Calibration summaries show ECE values from 0.2460 down to 0.2090, suggesting improved reliability at longer horizons, although the absolute ECE indicates that calibration remains an important component of the pipeline.

Log-loss increases with horizon (0.9235 to 2.3273), which is consistent with the fact that log-loss penalizes confident errors and the overall probability mass increases with the window length. For this reason, we interpret log-loss together with Brier score and calibration diagnostics, rather than in isolation.

D. Implications for risk mapping and monitoring

From a wildfire risk standpoint, the most relevant conclusion is that the model provides stable and decision-relevant performance across 5, 10, 15, and 20 days windows. Balanced accuracy near 0.78 indicates robust detection and rejection capability under class imbalance, while the probability outputs enable continuous risk ranking and mapping rather than binary alarms. The following subsection presents mean risk maps over the test period to illustrate how these probabilistic outputs translate into spatial risk zonation products at different horizons.

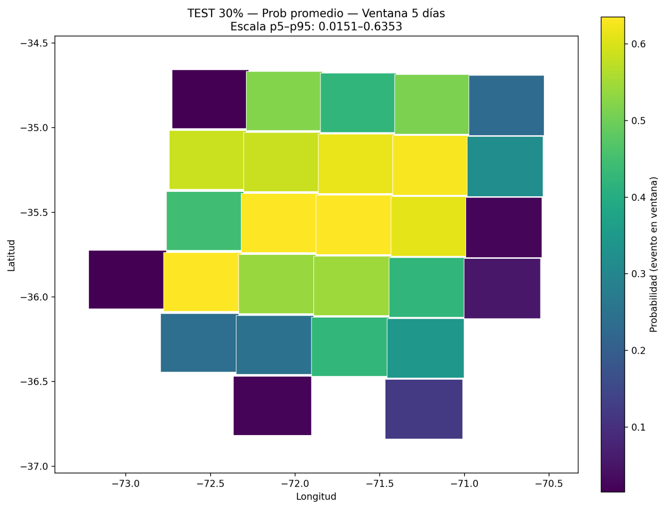
Figure 1 summarizes the spatial distribution of risk by averaging the predicted window probability over all test days for each horizon. Because window probabilities are computed as an “at least one event” probability over an increasing time window, the expected probability level increases with the horizon; therefore, the main comparison across horizons is the stability of the spatial pattern rather than absolute color intensity (each map uses its own p5 to p95 color limits).

For the 5-day horizon (Figure 1a), the highest mean probabilities concentrate in the central portion of the modeled grid, indicating a persistent hotspot where the model expects short-term occurrence to be more likely. Peripheral cells show lower mean probabilities, consistent with lower historical activity and weaker short-lead risk signals.

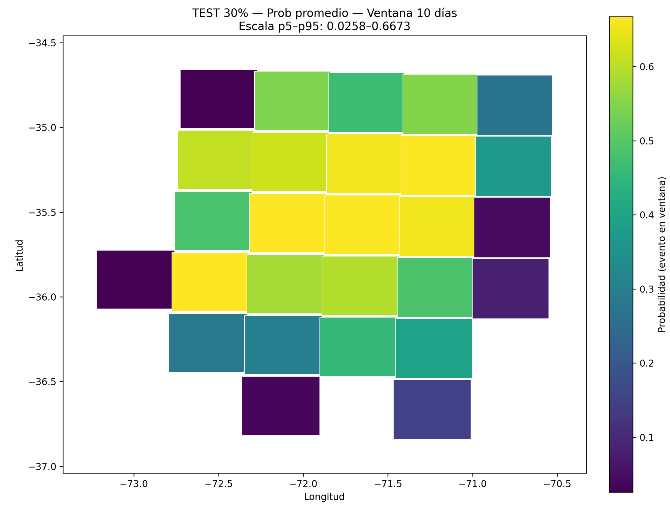
For the 10-day horizon (Figure 1b), the spatial structure remains largely unchanged, with the same central cluster dominating the risk map. Relative to 5 days, additional neighboring cells show moderate increases in mean probability, which is consistent with the longer window capturing more potential event opportunities.

For the 15-day horizon (Figure 1c), the map continues to emphasize the same core high-risk area, while mid-risk regions become more prominent around the hotspot. This suggests that the model preserves spatial ranking while increasing the estimated probability mass as the horizon expands.

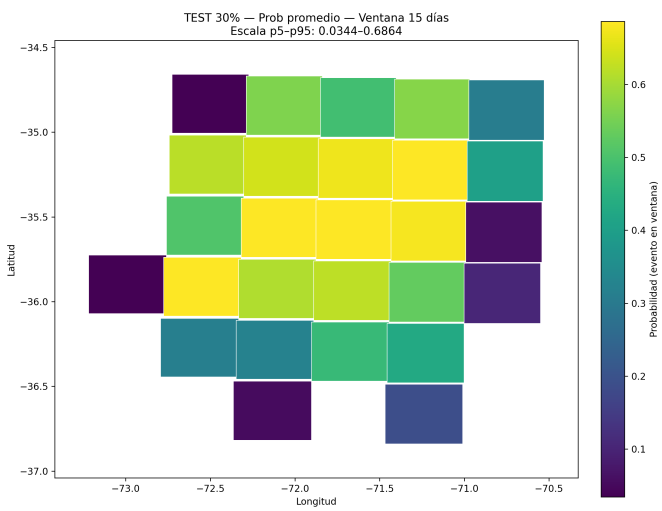
For the 20-day horizon (Figure 1d), the maximum mean probabilities are observed, but the dominant pattern remains stable: a central high-risk block and lower-risk margins. This



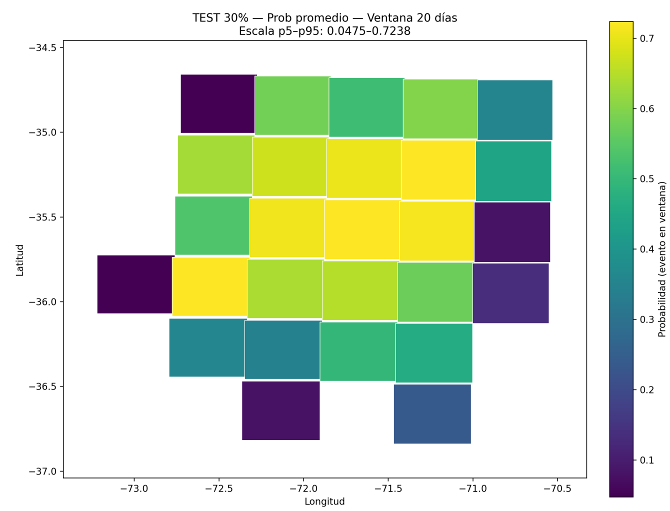
(a) Mean probability, 5-day window. Color scale uses p5 to p95: 0.0151 to 0.6353.



(b) Mean probability, 10-day window. Color scale uses p5 to p95: 0.0258 to 0.6673.



(c) Mean probability, 15-day window. Color scale uses p5 to p95: 0.0344 to 0.6864.



(d) Mean probability, 20-day window. Color scale uses p5 to p95: 0.0475 to 0.7238.

Fig. 1: Mean calibrated wildfire occurrence probability over the held-out test split, corresponding to the final 30% of the chronological dataset. Panels show the average window probability for the 5, 10, 15, and 20 days horizons. Since each horizon represents the probability of at least one event within an increasingly long time window, probability levels are expected to increase with horizon length. Color limits are independently scaled between the 5th and 95th percentiles within each horizon to improve spatial contrast; therefore, the maps should be compared mainly in terms of spatial pattern stability rather than absolute color intensity. Blank cells correspond to grid locations not included in the modeling set.

stability across 5, 10, 15, and 20 days supports the interpretation that the model captures persistent spatial heterogeneity, while the horizon mainly controls the overall probability level through the window aggregation.

VI. Limitations

Several limitations should be considered when interpreting the results. First, the study is focused on the Maule Region, and therefore the learned spatial patterns may reflect region-specific combinations of land cover, topography, climate, and human activity. Although the proposed formulation is transferable, its direct application to other regions should be validated with local data and recalibrated when necessary.

Second, wildfire occurrence is affected by non-stationary processes, including land-use change, fuel accumulation, drought conditions, forest management practices, and changes in prevention or suppression capacity. These factors may alter the relationship between historical predictors and future fire occurrence. For this reason, model calibration and threshold selection should be periodically updated when the framework is used operationally.

Third, the present work emphasizes probabilistic forecasting, calibration, and risk mapping rather than detailed causal interpretation. The model identifies statistical regularities useful for risk estimation, but it should not be interpreted as providing causal effects of individual variables. Similarly, although the Transformer architecture is suitable for modeling nonlinear spatiotemporal dependencies, additional ablation studies against simpler baselines and formal feature-attribution analyses are needed to further quantify the marginal contribution of each modeling component.

Finally, the risk maps summarize mean predicted probabilities over the test period. They are useful for visualizing persistent spatial risk structure, but operational use should rely on day-specific forecasts, updated covariates, and decision thresholds selected according to the cost of false alarms and missed events.

VII. Conclusions

This work presented a grid-based spatiotemporal framework for estimating wildfire occurrence probabilities in the Maule Region, Chile, using a Transformer-based architecture and coherent multi-horizon probability aggregation. By formulating the problem as daily hazard prediction followed by an "at least one event" window-probability construction, the proposed approach produces internally consistent probabilities for 5, 10, 15, and 20 days horizons. This structure is particularly relevant for operational wildfire risk monitoring, where decision-makers require stable probabilities over actionable time windows rather than isolated binary predictions.

The experimental results show stable predictive behavior under strong class imbalance. Balanced accuracy remained close to 0.78 across all horizons, indicating a consistent trade-off between detecting event windows and rejecting non-event windows. Discrimination metrics were also robust, while Brier score and ECE provided complementary evidence about probability quality and the importance of calibration. The resulting mean risk maps showed persistent high-risk zones across forecast horizons, suggesting that the model captures stable spatial heterogeneity in wildfire occurrence while adapting the overall probability level as the forecast window expands.

The study also highlights the importance of evaluating wildfire forecasting systems beyond raw accuracy. In sparse-event settings, probability calibration, threshold selection, and spatial interpretability are essential for operational usefulness. The proposed framework therefore contributes not only a predictive model, but also a probabilistic risk-mapping protocol that can support surveillance prioritization, prevention planning, and resource allocation.

Future work will extend the framework in four directions. First, we will evaluate simpler baselines and ablated versions of the model to quantify the specific contribution of the Transformer architecture, calibration stage, and coherent window-probability construction. Second, we will incorporate formal interpretability analyses, such as permutation importance or SHAP-based diagnostics, to assess the relative influence of meteorological, vegetation, topographic, and historical-occurrence variables. Third, we will test the framework in additional regions of Chile to evaluate geographic transferability and recalibration requirements under different fire regimes. Finally, we will study model updating strategies under non-stationary conditions associated with climate variability, land-use change, and evolving prevention or suppression practices.

References

[1] "Modeling the ignition risk: Analysis before and after megafire on maule region, chile," *Applied Sciences*, vol. 12, no. 18, p. 9353, 2022, full author list available via DOI on the publisher page.

[2] C. Vidal-Silva, R. Pizarro, M. Castillo-Soto, C. de la Fuente, V. Duarte, C. Sangüesa, A. Ibañez, R. Paredes, and B. Ingram, "Wildfire occurrence and damage dataset for chile (1985–2024): A real data resource for early detection and prevention systems," *Data*, vol. 10, no. 7, p. 93, 2025.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[4] X. Chen, Y. Tian, C. Zheng, and X. Liu, "Autost-net: A spatiotemporal feature-driven approach for accurate forest fire spread prediction from remote sensing data," *Forests*, vol. 15, no. 4, p. 705, 2024.

[5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[6] B. Zadrozny and 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 (KDD)*, 2002.

[7] T. Dimitriadis, T. Gneiting, and A. I. Jordan, "Stable reliability diagrams for probabilistic classifiers," *Proceedings of the National Academy of Sciences*, vol. 118, no. 8, 2021.

[8] "Statistical calibration of probabilistic medium-range fire weather index forecasts in europe," *Natural Hazards and Earth System Sciences*, vol. 24, pp. 4225–4243, 2024, full author list available via DOI on the journal page.

[9] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.