AUDITING TABULAR FOUNDATION MODELS FOR WILDFIRE FORECASTING: TRANSFER, LOCAL ADAPTATION, AND WEATHER-RESPONSE STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Wildfire forecasts must rank rare fire days from weather predictions, often before a region has local fire labels. We evaluate TabPFN and TabFM against CatBoost, LightGBM, logistic regression and the Fine Fuel Moisture Code (FFMC), using complete fire journals from two Russian regions and held-out seasons in British Columbia (BC). Neither foundation model beats its selected classical comparator in any of six zero-label transfer contrasts; FFMC recalls 3–5 percentage points more fires at a 5% alert budget. With 64 local BC labels, averaging five contexts improves both foundation models, but averaged trees reach the same recall. On BC 2023 the foundation models exceed the classical family selected by single-context validation by about 0.9 average-precision points and 1.8–2.0 recall points. The prespecified rule for confirming this advantage on untouched BC 2024 fails: only 2 of 16 simultaneous intervals are positive, although context averaging still helps. Weather-response categories disagree across contexts for 27–28% of queries at 4,096 labels, compared with 17% for logistic regression. Agreement across contexts selects FFMC-teacher response directions no better than response magnitude at matched coverage. Ranking, response stability and FFMC-equation fidelity therefore require separate evaluation.

1 INTRODUCTION

A regional fire agency may record only a few hundred fires per year across tens of thousands of cell-days. For a newly covered territory, it has no local labels. TabPFN [1; 7] and TabFM [8] can predict from examples supplied at inference time and compete with tuned gradient boosting on small benchmark tables [38]. Fire forecasts test whether that performance survives geographic shift and rare outcomes.

Agencies act on a short warning list, so a useful model must rank fires within a fixed alert budget. Its ranking must also survive a move to a new region or improve quickly when local labels arrive. Forecasters may read the model's response to rain, wind or heat as an explanation, yet that response can change when the labelled context changes. Ranking, context stability and agreement with a known fire-weather equation therefore require separate measurements.

We audit TabPFN 3.5 and TabFM 1.0 under five frozen evaluation settings (Table 1): zero-label transfer from two Russian regions to British Columbia (BC), adaptation with 64–4,096 local BC labels, synthetic and sequential FFMC teachers, archived station-weather FFMC controls, and an out-of-year replication on the untouched BC 2024 season. CatBoost, LightGBM and logistic regression receive the same labelled rows, validation budgets and context averaging, and a label-free FFMC ranking is the physical reference. The primary predictive endpoint integrates average precision (AP) across 14 forecast leads and a fixed curve of training budgets; warning recall is measured when only 5% or 10% of cells may be flagged. The BC protocol, target construction, simultaneous interval families, figure plan and interpretation rules were frozen before any aggregate BC metric was computed. Russian experiments were developed after initial inspection and are reported as exploratory.

We compare both foundation models with matched classical models and a label-free fire-weather index under a frozen, region-separated protocol. We then vary the labelled context, measure predic-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion and weather-response disagreement, and test whether cross-context agreement selects FFMC-consistent directions better than response magnitude. Every learner receives the same context averaging. An untouched BC 2024 season tests whether the BC 2023 adaptation results persist; its protocol was fixed before 2024 labels were joined to weather. Together these experiments audit warning utility and the reproducibility of weather-response explanations under finite context.

2 Related work

TabPFN [1; 7], TabFM [8] and TabICL [39] use labelled examples in context and compete with tuned boosting on curated benchmarks such as TabArena [38]. Trees remain strong on ordinary tabular data [37], and performance can deteriorate under temporal, group and geographic shift [28; 29]. Work on geospatial TabPFN [25], context sampling [3], active acquisition [35], retrieval and fine-tuning [32], and context optimization [33] shows that context construction matters. Here we hold the label budget and context rule fixed across foundation and classical models.

Models with similar predictive scores can disagree on labels, counterfactuals and attributions [17; 18; 19; 5]. Rashomon-set sampling [13] and fine-tuning studies of tabular LLMs [31] measure this multiplicity; explanations for tabular foundation models are also available [24]. In-context learning adds another choice at inference time: which labelled rows to supply. We measure how that choice changes responses to weather and whether agreement across choices predicts the FFMC equation's response direction.

Tree ensembles remain competitive for fire occurrence across Russian regions [36]. Wildfire-danger studies have examined explanations and physical consistency [4; 20; 21; 30], while global fire-activity forecasts combine weather, fuel and ignition predictors [40]. Ensemble weather forecasts have been tested for Canadian fire-weather components at 1–15 days [41]. Foundation models have been studied for post-fire hazards [26] and next-day spread [34]; WILDFIRE-FM calls for controlled evaluation [12]. Consensus ranking [22; 23] motivates prediction averaging, but does not show whether it improves an in-context learner more than a tree ensemble. We test both under the same label budgets and warning rule.

3 Data and forecasting task

The target is at least one registered fire in an administrative-region–H3-resolution-4-cell–calendar-day. It represents recorded occurrence, not every ignition, burned area or intensity. The data owner confirms complete journals for Chelyabinsk (CHE) and Kurgan (KUR) in 2019–2023. Sverdlovsk completeness is unknown, so it supports no negative-label claim. Administrative boundaries define the denominator independently of fire locations, and days with unlocated incidents are masked.

Weather inputs are forecasts issued before the target day, never observed future weather: AIFS-derived features at leads 1–7 days for the Russian comparison and an archived GFS corpus [6] with 31 initialization dates per season, leads 1–14 and two daily aggregation windows for BC. The historical dissemination time of each file is not reconstructed, so every result is a retrospective forecast from archived fields. TimesFM 2.5 [9] supplies temperature and humidity from station histories ending before initialization, without rainfall or wind.

The external endpoint contains 5,496 BC Wildfire Service records typed Fire in 2021–2023 [14], released under the Open Government Licence–British Columbia [15] and gridded to 593 H3 cells within the provincial boundary and a land mask [16]. The catalogue documents the ignition date as a discovery date without a timezone; we treat it as a local calendar date and test ± 1-day shifts of the target date. The panel holds 96,760/96,432/96,268 unique cell-days with 1,038/1,001/1,640 fire cell-days in 2021/2022/2023, a 2023 prevalence of 1.7%. CHE, KUR and their union supply 2021 training labels and 2022 validation labels. No BC label enters fitting, calibration or selection before evaluation on BC 2023.

After the zero-label result was inspected, a frozen local experiment used BC 2021 for training, BC 2022 for selection and BC 2023 only for evaluation, with contexts of 25% positive and 75% negative unique cell-days because natural-prevalence contexts of 64 rows are frequently one-class. Five nested contexts are shared by TabPFN, TabFM, CatBoost, LightGBM and logistic regression. Because case enrichment changes the class prior, we evaluate rankings and fixed-budget warnings

Table 1: Evaluation settings used for the main claims. Each held-out split answers a different question.

Evaluation setting	Training and selection	Held-out evaluation	What it establishes
Zero-label BC transfer	CHE, KUR or CHE+KUR in 2021; source-only 2022 validation	BC 2023; no BC labels before test	Geographic transfer and warning utility
BC local adaptation	BC 2021 contexts; BC 2022 selection	Untouched BC 2023	Benefit of 64–4,096 target labels
Synthetic and sequential FFMC teachers	Five nested contexts; fixed model settings	1,200 synthetic states and 2,688 sequential weather rows	Response direction when the equation supplies truth
Station FFMC control	Five contexts from 85,750 episodes at 147 stations	2,048 episodes at 144 stations	Equation-conditioned fidelity on archived weather states
BC 2024 replication	BC 2021–2022 contexts; BC 2023 selection	Untouched BC 2024; protocol frozen before any 2024 label was joined	Out-of-year test of context averaging, FM advantage and physical features

rather than calibrated probabilities. Both tabular FMs were released after the target years [7; 8], so all results are retrospective.

4 Evaluation

Table 1 separates the transfer, local-label and FFMC-teacher experiments so that each supports its own claim.

4.1 Model selection, metrics and uncertainty

All models receive the same labelled cell-days; for training budget N, validation adds max(32, N/4) labels, contexts are nested, and multiple source regions share a fixed total budget. Source-validation AP selects candidate settings, with log loss as a tie-break; at N=64 the five-input weather comparison has too few positives to tune and uses fixed settings.

Raw AP measures ranking and Brier score measures probability error. Warnings flag ceil(5% or 10% of eligible cells) at each initialization and lead; ties are integrated analytically. We report recall, precision and realised alert fraction. Repeated forecast outcomes are forecast cell-days rather than unique ignitions.

Two admissible 64-label contexts can give different rankings: within-issue rank correlation is 0.53 for TabPFN, 0.42 for TabFM, 0.40 for CatBoost and 0.48 for LightGBM, against 0.76 for logistic regression. At 4,096 labels every model exceeds 0.94. We average raw predictions over K contexts drawn by the same rule and compare the result with the mean single-context score. Every learner, including the classical baselines, receives the same five-context treatment. Averaging costs K inferences and has no tuned weight.

External primary intervals resample the 31 initialization dates per season in paired 35-day moving blocks, keeping repeated cells, windows and leads from one initialization together; simultaneous bands cover 18 prespecified contrasts. The many forecast rows therefore do not represent independent time replicates. A 15-day block analysis and $\pm$1-day target-date shifts test sensitivity. Domestic intervals use 1,000 paired temporal-block/context draws and are exploratory. An interval containing zero is unresolved; it does not establish equivalence.

A label-free shift diagnostic (median normalized 1-Wasserstein distance between source and BC weather, 0.59–0.71 against 0.23–0.26 for source-temporal controls) documents covariate shift without attributing model failure to it.

4.2 WEATHER-RESPONSE TESTS

We perturb temperature by 2°C at fixed dewpoint or fixed relative humidity, wind by 2 m/s and rainfall by 5 mm, recompute derived humidity, and query baseline and perturbed rows in the same order. A log-odds response is negative, flat or positive at thresholds ± 0.05; category disagreement records changes between fitted models, and opposite-active disagreement requires non-flat responses of opposite sign. These conditional probes are diagnostic, not coupled atmospheric and fuel trajectories.

Three controls supply a known response direction. A synthetic teacher applies the canonical FFMC update [27] through two pinned implementations [10; 11] and maps FFMC F to probability sigmoid((F-85)/5); five generated training sets give nested 64/256/1,024-label contexts, evaluated on 1,200 states and 4,800 perturbations per fit. A sequential control propagates FFMC for 14 days over archived GFS sequences (2,688 training and 2,688 test rows, one-step prediction from the true previous state). A station control uses 85,750 complete CWFIS weather-to-FFMC episodes from 147 BC stations, with 2,048 test episodes from 144 stations selected by a weather-independent hash; CWFIS FFMC is computed from weather, not measured fuel moisture.

The finite-context Direction Certificate (CDC) reports a non-flat response direction only when all five contexts agree. We compare its FFMC-equation fidelity with a magnitude rule that retains the same number of responses by absolute median response. Two thousand state-cluster bootstrap draws preserve the four perturbations. Agreement among sampled contexts does not guarantee a correct direction on new weather states.

5 RESULTS

5.1 LOCAL LABELS AND CONTEXT AVERAGING RECOVER USEFUL RANKINGS

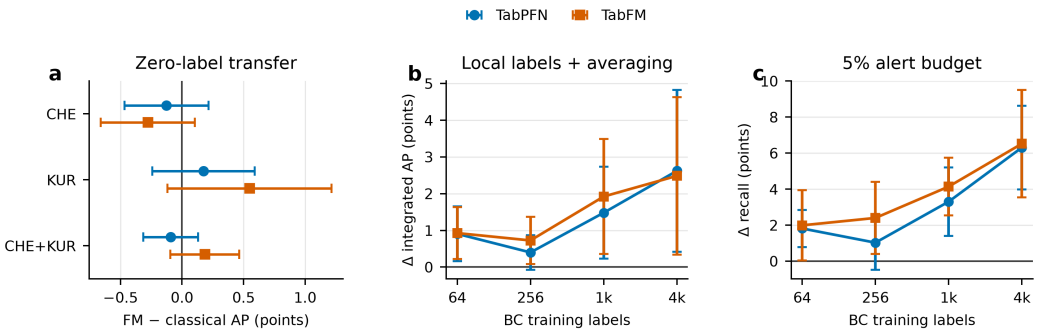

Figure 1: Zero-label transfer is unsupported; local labels improve BC 2023 ranking. (a) FM-minus-classical integrated AP after source-only training; all simultaneous intervals cross zero. (b,c) Five-context FMs minus five-context logistic regression, selected by single-context BC-2022 validation. Seven of eight intervals are positive for AP and warning recall. Bagged trees are a stronger small-budget comparator (Table 2). Whiskers are 35-day simultaneous intervals; differences are percentage points.

None of the 18 prespecified zero-label contrasts supports positive FM transfer (Figure 1a). All six FM-minus-classical intervals cross zero; five of six comparisons with vapour-pressure deficit (VPD) are adverse, while all comparisons with the surface wind×VPD proxy are unresolved. This proxy differs from canonical HDW [2]. The 15-day blocks and shifted fire dates leave those decisions unchanged. At 5% alerts, FFMC-85 recalls 3.2–5.5 percentage points more fire cell-days than either FM across the three source sets.

With BC labels, integrated AP across four budgets exceeds the validation-selected classical model by 0.87 points for TabPFN (35-day simultaneous interval [+0.14, +1.61]) and 0.95 for TabFM ([+0.03, +1.88]). The mean advantages at 64/256 labels remain unresolved, and the small-minus-large interactions are adverse. Single-context results do not support a special small-budget FM advantage.

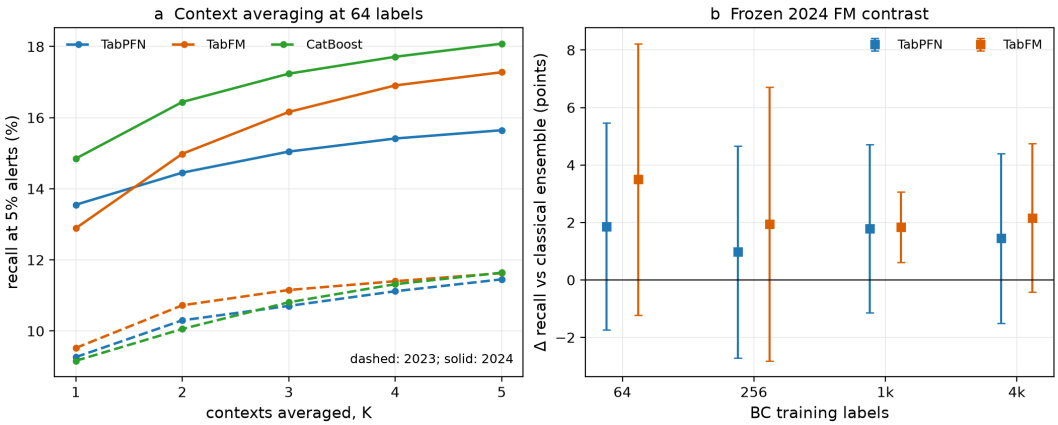

Figure 2: Context averaging in BC. (a) Recall at 5% alerts with 64 labels over all subsets of K among five contexts. Dashed lines are 2023, solid lines 2024; TabPFN, TabFM and CatBoost are shown. (b) Frozen BC 2024 five-context FM minus the classical family selected by single-context BC-2023 validation for 5% recall. Whiskers are 35-day simultaneous intervals; intervals crossing zero are unresolved. AP contrasts are reported in Section 5.4 and the supplementary material.

We then average predictions from five contexts, with the averaging rule fixed before seeing 2023 outcomes (Figure 1b,c). Against the logistic ensemble selected by single-context validation AP, seven of eight AP intervals and seven of eight 5%-alert recall intervals are positive. At 64 labels, TabPFN and TabFM gain 0.91 and 0.93 AP points, and 1.8 and 2.0 recall points. The comparator here is selected logistic regression; bagged trees require a separate comparison.

At 64 labels, five-context recall reaches 11.5% for TabPFN and 11.6% for TabFM, CatBoost and LightGBM (Figure 2a, Table 2). Logistic regression moves from 9.9% to 9.6%. Averaging helps the learners whose single-context rankings disagree most, including both trees; these data cannot isolate an effect of pretraining. FM recall point estimates exceed the bagged trees from 1,024 labels (13.8–14.6% versus 11.2–11.3%) and remain higher at 4,096 (15.8–16.0% versus 13.8–14.2%). At 64 and 256 labels, the direct tree comparison shows no distinct FM advantage.

Table 2 reports both the mean single-context and five-context recall for every model. A random 5% alert set recalls 5% of fire cell-days; FFMC-85 recalls 8.25%. The absolute model values are descriptive, while the intervals in the last two rows test FM gains against the selected logistic ensemble. Even the best single-context classical learner, chosen using test outcomes and therefore unavailable in practice, beats both FMs at 64 labels by 0.6–0.9 recall points. The FMs lead that oracle by 1.5–2.4 points from 1,024 labels.

5.2 Predictive fit and FFMC-equation response separate

With 4,096 BC labels, mean response disagreement falls to 27.3% for TabPFN and 28.1% for TabFM, from 39.8% and 42.2% at 64 labels (Figure 3a). Wind is the least reproducible perturbation: the 4,096-label disagreement remains 43.8% and 41.0%. Logistic regression has 16.9% mean disagreement. At 64 labels, changing the context changes response categories for 45–63% of queries; changing only the fit seed changes 5–17%. The labelled context matters more than fit randomness for these explanations.

On independent synthetic FFMC states, probability MSE falls from 0.043 to 0.015 for TabPFN and from 0.047 to 0.014 for TabFM between 64 and 1,024 labels. Yet rain-response category error rises between 256 and 1,024 labels, from 13.2% to 15.2% and from 10.0% to 11.3%. Both metrics improve on the sequential FFMC control. The apparent trade-off depends on the evaluation distribution.

Table 2: Recall of registered-fire cell-days at 5% alerts on BC 2023, in percent. Each absolute cell gives the mean single-context value followed by the five-context value. Differences in the last two rows compare five-context FMs with the five-context logistic ensemble selected by single-context validation AP; brackets are 35-day simultaneous intervals. Bagged tree values above are descriptive comparisons without intervals.

Model (single → five contexts)	**N=64**	**N=256**	**N=1,024**	**N=4,096**
TabPFN	9.3 → 11.5	10.1 → 10.7	12.3 → 13.8	15.6 → 15.8
TabFM	9.5 → 11.6	11.2 → 12.0	12.5 → 14.6	15.3 → 16.0
CatBoost	9.2 → 11.6	10.5 → 12.0	10.8 → 11.2	12.7 → 13.8
LightGBM	9.0 → 11.6	10.6 → 11.6	10.4 → 11.3	13.2 → 14.2
Logistic regression	9.9 → 9.6	10.1 → 9.6	10.1 → 10.4	9.4 → 9.4
Five-context TabPFN - classical ensemble	+1.8 [+0.8, +2.8]	+1.0 [-0.5, +2.5]	+3.3 [+1.4, +5.2]	+6.3 [+4.0, +8.6]
Five-context TabFM - classical ensemble	+2.0 [+0.04, +3.9]	+2.4 [+0.4, +4.4]	+4.1 [+2.5, +5.7]	+6.5 [+3.5, +9.5]

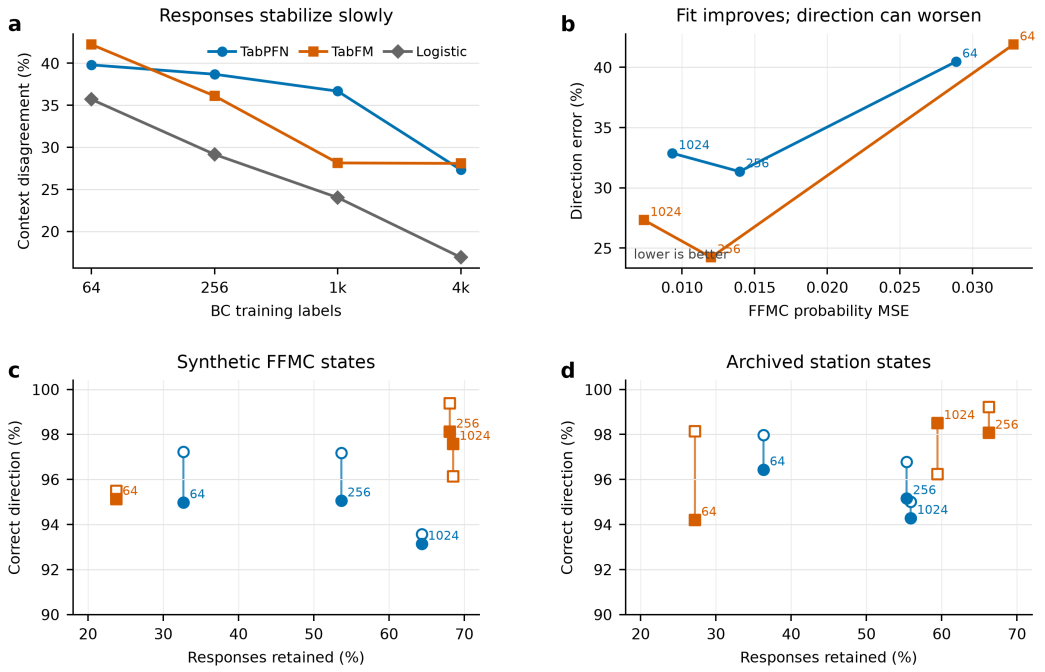

Figure 3: Prediction, context stability and FFMC-equation direction differ. (a) Mean category disagreement across four conditional weather perturbations and ten context pairs on BC 2023. (b) On archived station weather, probability MSE improves while FFMC-teacher direction error falls to 256 labels and then rises; lower-left is better. (c,d) Filled points show CDC fidelity and open points same-coverage magnitude selection on FFMC teachers. CDC wins in one of six settings per control. These tests do not establish causal fire response.

On archived station states, probability MSE falls from 0.029 to 0.009 for TabPFN and from 0.033 to 0.007 for TabFM across 64–1,024 labels (Figure 3b). Direction-category error falls to 256 labels, then rises to 1,024: from 31.4% to 32.9% for TabPFN and from 24.2% to 27.3% for TabFM.

The five-context agreement rule retains 23.7–68.5% of nonzero synthetic FFMC responses and gets 93.1–98.1% of their directions right (Figure 3c,d). At the same coverage, choosing responses by their magnitude does as well or better in five of six FM settings; only TabFM at 1,024 labels favours

agreement. The station control gives the same six decisions. High fidelity here comes mainly from abstention, and agreement cannot detect a wrong direction shared by all five contexts.

5.3 Capacity and weather-input controls

At N=1,024, official TabPFN fine-tuning changes integrated AP by 0.04 points [-0.05, +0.14] and a 32-estimator TabFM by 0.01 points [-0.05, +0.07]; warning gains are unresolved and neither intervention improves all response-stability metrics.

TimesFM 2.5 forecasts of temperature and humidity yield no wholly positive AP interval in 128 matched contrasts with GFS inputs. A matched AIFS/GFS comparison is unresolved in 29 of 32 AP contrasts and all 64 warning contrasts. Removing coordinates repairs one failing Russian transfer route but improves none of 20 local comparisons.

Exploratory experiments frozen after BC 2023 was inspected add forecast-derived fire-weather state to every learner. A fixed blend of model and FFMC rankings hurts ranking; using the state as a feature improves 7 of 20 same-model contrasts, including FM and classical models, but none of eight FM contrasts against the best classical learner. On BC 2024 only 1 of 20 feature gains is positive. These results do not support an FM-specific benefit from physical features.

5.4 Out-of-year replication on BC 2024

We repeated adaptation on BC 2024 under a protocol frozen before any 2024 incident was joined to weather: contexts from 2021–2022, selection on 2023, and the same models, budgets, seeds and decision rules. The season was quieter (1,077 fire cell-days, prevalence 1.2%) and easier to rank: FFMC-85 recalls 12.1% at 5% alerts and every bagged model recalls about 14–22%.

On BC 2024, five-context averaging exceeds the mean single context in all 16 point estimates (Figure 2a). Five simultaneous intervals are positive, including both FMs at 64 labels on AP and recall. TabPFN gains 0.9 AP and 2.1 recall points; TabFM gains 1.8 and 4.4 points. CatBoost improves from 14.9% to 18.1% recall at 64 labels, while logistic regression moves from 13.3% to 13.8%.

The prespecified full-replication rule for an FM advantage fails (Figure 2b). Only two of sixteen simultaneous intervals are positive, both at 1,024 labels (TabPFN +0.47 AP points; TabFM +1.8 recall points); fourteen are unresolved and none adverse. Fifteen of sixteen point estimates are positive, but their signs alone do not establish an effect. At 64 labels the intervals are wide (TabFM +3.5 recall points, [-1.2, +8.2]), with fewer positive cases per issue block in this quieter season. The comparator itself moved: validation selects logistic regression at 64 and 256 labels but CatBoost at 1,024 and 4,096, and bagged CatBoost recalls more fires than either FM at 64 labels (18.1% versus 15.7% and 17.3%), whereas the FMs have higher recall point estimates from 1,024 labels (20.7–22.1% versus 18.9–20.3%). The physical-state and oracle findings do not replicate as gains (1 of 20 and 1 of 8 positive contrasts), and response disagreement falls from 41–44% at 64 labels to 23–29% at 1,024 for both FMs while the tree ensembles stay at 37–52%.

6 Discussion

The practical order of comparison matters. Before collecting local labels, FFMC-85 is a stronger 5%-alert baseline than either foundation model transferred from Russia. After local labels arrive, averaging contexts improves trees as well as FMs. A fair selection procedure should therefore average every candidate before comparing them and give forecast-derived fire-weather state to all candidates. The BC results do not identify a pretraining effect: bagged trees match FM recall at 64–256 labels, while FM point estimates lead from roughly 1,024 labels without a direct inferential test against bagged trees.

Prediction scores leave the weather-response question open. More labels make FM responses more reproducible, but they remain sensitive to context. On FFMC-equation controls, lower probability error can coincide with worse direction error. Agreement across contexts does not select correct directions better than magnitude at matched coverage. None of these tests validates a causal response of real fires to weather.

The evaluation extends to other rare-event tasks where a mechanistic model supplies a reference response. Wildfire is demanding because geographic transfer, a severe class imbalance and weather-dependent explanations occur together. The FFMC equation makes one part of the explanation test measurable; it does not establish the true response of ignition risk.

The study has two complete Russian journals, one external jurisdiction and one external test year. It predicts registered occurrence rather than every ignition or burned area; historical forecast dissemination times and some fire-date semantics remain uncertain. The 2024 season confirms gains from averaging but not the prespecified small-budget FM advantage. Wider geographic and temporal replication is needed before treating the observed 1,024-label FM lead as general.

7 Conclusion

The frozen 2024 test confirms a gain from averaging contexts, not a small-budget advantage specific to tabular foundation models. Zero-label transfer remains unsupported, and FFMC-85 is the stronger label-free warning baseline. Model selection for fire forecasts should compare averaged trees and foundation models at the same label budget, then audit weather-response stability separately from ranking.

Reproducibility Statement

The accompanying anonymous artifact records model versions, supervision identities, derived results and independent metric checks. The short supplement states the evaluation contract and principal sensitivities. The checks reconstruct saved metrics and intervals from predictions, but do not refit every model.

AI Use Statement

Generative AI assisted literature search, experimental design, code development, result checks, manuscript review and text editing. The authors verified the analyses and claims and take responsibility for the paper.

Ethics Statement

The target is registered fire occurrence and may reflect reporting and monitoring practices. Forecast errors could misallocate scarce resources; the study defines no autonomous warning policy or deployment threshold. Locations are aggregated to coarse grid cells. Operational use requires local validation, human oversight and analysis of unequal error costs.

References

[1] Hollmann et al. (2025). Accurate predictions on small data with a tabular foundation model. Nature. https://www.nature.com/articles/s41586-024-08328-6

[2] Srock et al. (2018). The Hot-Dry-Windy Index: A New Fire Weather Index. https://research.fs.usda.gov/treesearch/56562

[3] Abdullah (2026). Understanding Context Sampling in TabPFN on Small Tabular Datasets. Preprint. https://arxiv.org/html/2607.26628v1

[4] Becker, Natel and Nowack (2026). Assessing and explaining temporal deep learning models for wildfire danger prediction. https://doi.org/10.1088/3049-4753/ae5aa0

[5] Laberge, Pequignot, Mathieu, Khomh and Marchand (2023). Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set. JMLR 24(364):1–50. https://www.jmlr.org/beta/papers/v24/23-0149.html

[6] NOAA Global Forecast System archive. https://registry.opendata.aws/noaa-gfs-bdp-pds/

[7] Jäger et al. (2026). TabPFN-3.5: Technical Report. https://arxiv.org/abs/2609.17895v1

[8] Google Research (2026). TabFM: A Zero-Shot Foundation Model for Tabular Data. Official release card and citation: https://huggingface.co/google/tabfm-1.0.0-pytorch

[9] Das, Kong, Sen and Zhou (2024). A decoder-only foundation model for time-series forecasting. ICML, PMLR 235:10148–10167. https://proceedings.mlr.press/v235/das24c.html ; TimesFM 2.5 checkpoint card: https://huggingface.co/google/timesfm-2.5-200m-pytorch

[10] CFFDRS Python implementation, pinned commit 0f57fcca. https://github.com/cffdrs/cffdrs_py

[11] CFFDRS R implementation and input documentation, pinned commit 4d20a303. https://github.com/cffdrs/cffdrs_r

[12] Xu, Dai, Chang, Wang and Dong (2026). Does Your Wildfire Prediction Model Actually Work, or Just Score Well? Preprint, v2. https://arxiv.org/abs/2605.18911v2

[13] Li, Barnard and Deng (2026). Practical and Efficient Rashomon Set Sampling for Model Interpretability. AISTATS, PMLR 300:262–270. https://proceedings.mlr.press/v300/li26a.html

[14] British Columbia Wildfire Service. BC Wildfire Fire Incident Locations – Historical. BC Data Catalogue record e2dadc60-292f-4d98-b42b-56ca9e4fe694. https://catalogue.data.gov.bc.ca/dataset/e2dadc60-292f-4d98-b42b-56ca9e4fe694

[15] Province of British Columbia. Open Government Licence – British Columbia, version 2.0. https://www2.gov.bc.ca/gov/content/data/policy-standards/data-policies/open-data/open-government-licence-bc

[16] Natural Earth. 1:10m Land physical vector and Terms of Use. https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-land/ ; https://www.naturalearthdata.com/about/terms-of-use/

[17] D'Amour et al. (2022). Underspecification Presents Challenges for Credibility in Modern Machine Learning. JMLR 23(226):1–61. https://jmlr.org/papers/v23/20-1335.html

[18] Marx, Calmon and Ustun (2020). Predictive Multiplicity in Classification. ICML, PMLR 119:6765–6774. https://proceedings.mlr.press/v119/marx20a.html

[19] Pawelczyk, Broelemann and Kasneci (2020). On Counterfactual Explanations under Predictive Multiplicity. UAI, PMLR 124:809–818. https://proceedings.mlr.press/v124/pawelczyk20a.html

[20] Kondylatos et al. (2022). Wildfire Danger Prediction and Understanding With Deep Learning. Geophysical Research Letters 49(17):e2022GL099368. https://doi.org/10.1029/2022GL099368

[21] Li et al. (2024). Projecting Large Fires in the Western US With an Interpretable and Accurate Hybrid Machine Learning Method. Earth's Future 12(9):e2024EF004588. https://doi.org/10.1029/2024EF004588

[22] Ahsen, Vogel and Stolovitzky (2019). Unsupervised Evaluation and Weighted Aggregation of Ranked Classification Predictions. JMLR 20(166):1–40. https://jmlr.org/papers/v20/18-094.html

[23] Goibert, Calauzenes, Irurozki and Clemencon (2023). Robust Consensus in Ranking Data Analysis: Definitions, Properties and Computational Issues. ICML, PMLR 202:11584–11597. https://proceedings.mlr.press/v202/goibert23a.html

[24] Sena and Azevedo (2026). Real-Time Explanations for Tabular Foundation Models. ICLR 2026 workshop; arXiv:2603.29946. https://arxiv.org/abs/2603.29946

[25] Deng, Li and Wang (2026). Do foundation models work for geospatial tabular data? An investigation of TabPFN and a proposed enhancement based on geospatial sparse attention. International Journal of Geographical Information Science:1–38. https://doi.org/10.1080/13658816.2026.2691066

[26] Ledingham et al. (2026). Evaluating Machine Learning Models for Post-Wildfire Debris-Flow Prediction. arXiv:2608.05265. https://arxiv.org/abs/2608.05265

[27] Van Wagner (1987). Development and Structure of the Canadian Forest Fire Weather Index System. Canadian Forestry Service, Forestry Technical Report 35. https://ostrnrcan-dostrncan.canada.ca/handle/1845/228434

[28] Purucker et al. (2026). Beyond IID: How General Are Tabular Foundation Models, Really? arXiv:2606.30410. https://arxiv.org/abs/2606.30410

[29] Loza et al. (2026). Empirical Evaluation of Out-Of-Distribution Performance of Tabular Foundation Models. arXiv:2607.26000. https://arxiv.org/abs/2607.26000

[30] Xu et al. (2026). Trustworthy Data-Driven Wildfire Risk Prediction and Understanding in Western Canada. arXiv:2601.01677. https://arxiv.org/abs/2601.01677

[31] Hamman et al. (2025). Quantifying Prediction Consistency Under Fine-tuning Multiplicity in Tabular LLMs. ICML, PMLR 267:21662–21684. https://proceedings.mlr.press/v267/hamman25a.html

[32] Thomas et al. (2024). Retrieval & Fine-Tuning for In-Context Tabular Models. NeurIPS 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/c40daf14d7a6469e65116507c21faeb7-Abstract-Conference.html

[33] Feuer et al. (2024). TuneTables: Context Optimization for Scalable Prior-Data Fitted Networks. NeurIPS 37. https://proceedings.neurips.cc/paper_files/paper/2024/hash/97dc07f1253ab33ee514f395a82fa7cc-Abstract-Conference.html

[34] Li, Hsu and Wang (2026). Adapting Video Foundation Models for Spatiotemporal Wildfire Forecasting via Cross-Modal Progressive Fine-Tuning. IEEE Transactions on Geoscience and Remote Sensing 64:3000218. https://ieeexplore.ieee.org/document/11343839/

[35] Treerath and Pittorino (2026). Active In-Context Learning for Tabular Foundation Models. arXiv:2603.27385. https://arxiv.org/abs/2603.27385

[36] Illarionova, Shadrin, Gubanov, Shutov, Tasuev, Evteeva, Mironenko and Burnaev (2025). Exploration of geo-spatial data and machine learning algorithms for robust wildfire occurrence prediction. Scientific Reports 15:10712. https://doi.org/10.1038/s41598-025-94002-4

[37] Grinsztajn, Oyallon and Varoquaux (2022). Why do tree-based models still outperform deep learning on typical tabular data? NeurIPS 35, Datasets and Benchmarks Track. https://arxiv.org/abs/2207.08815

[38] Erickson, Purucker, Tschalzev, Holzmüller et al. (2025). TabArena: A Living Benchmark for Machine Learning on Tabular Data. NeurIPS 38, Datasets and Benchmarks Track. https://arxiv.org/abs/2506.16791

[39] Qu, Holzmüller, Varoquaux and Le Morvan (2025). TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. ICML, PMLR 267. https://arxiv.org/abs/2502.05564

[40] Di Giuseppe, McNorton, Lombardi and Wetterhall (2025). Global data-driven prediction of fire activity. Nature Communications 16:2918. https://doi.org/10.1038/s41467-025-58097-7

[41] Chen, Jain, Ramsey, Chen and Flannigan (2025). Comparative Analysis of Ensemble and Deterministic Models for Fire Weather Index (FWI) System Forecasting. Weather and Forecasting 40(12):2579–2600. https://doi.org/10.1175/WAF-D-25-0069.1