AN OPEN NATIONAL BENCHMARK FOR FORECASTING REGISTERED WILDFIRE OCCURRENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-learning benchmarks for wildfire forecasting are built almost entirely on satellite fire observations, yet where those products have been compared with agency records they detect a minority of the smaller fires those registers hold, and detection improves steeply with fire size. We release an open benchmark for forecasting where forest fires will be *registered*: 83 regions of Russia, selected by the continuity of their fire records, 64,620 hexagonal cells of about 253 km^2, daily over the April to October seasons of 2000 to 2024, labelled with 420,995 registered fires; 0.11% of cell-days carry at least one. Every input is openly licensed and redistributable, with weather sampled on a fixed hierarchical grid that keeps every cell within 68 km of its weather, so the dataset can be published in full and served in real time from the same sources. The protocol is built for extreme imbalance: a temporal split, evaluation on the complete national grid rather than on sampled negatives, and metrics a dispatcher can act on, above all the share of the day's fire cells caught inside a fixed alarm budget. Across 807 days of four fire seasons, a gradient-boosted baseline catches 80.5% of the day's fire cells while flagging 15.2% of the grid, against 63.7% for climatology, 53.6% for the Canadian Fire Weather Index and 47.9% for the Russian statutory index at the same budget; it beats climatology on 97% of individual days and leads at every budget we scored, from 1% to 25% of the grid. Capture varies little across fire size, so the model finds small fires as well as large ones. Matched-representation experiments then show that the hand-designed fire-weather summaries this field relies on, including the statutory index, are replaceable: the same model given the raw 30-day weather series in their place matches or beats them, while a recurrent model and a graph over weather channels show no consistent advantage over it.

1 INTRODUCTION

A fire service answers for the fires in its register. A register entry starts a duty to respond, it determines what is reported upward, and at the end of the season it is the record against which the service is judged. Machine-learning benchmarks for wildfire forecasting, however, are built almost entirely on satellite observations: all ten public benchmarks whose labels we could trace take their targets wholly or partly from satellite active-fire or burned-area products, and none from agency records of occurrence (Table 8).

These two records are not the same. Where satellite products have been compared against agency databases, the MODIS active-fire product reached a 50% probability of detection only for fires whose final size was 10 ha in the eastern United States to 78 ha in the western United States, and then only under ideal conditions, while the agency records in turn held only 23.5 to 48% of the fires MODIS observed (Fusco et al., 2019); against Landsat-mapped perimeters, the same product detected 82% of fires of at least 18 ha, with 50% detection at 105 ha (Hawbaker et al., 2008). In a registry dominated by small fires the shortfall is severe: in Turkey, where 78.5% of recorded fires are smaller than one hectare, MODIS detected 0.1 to 0.3% of them (Coskuner, 2022). The register we model is not of that kind: over the whole period its median fire burns 53 ha. In the years we evaluate on, however, the median is 22 ha (Section 3.1), above the eastern threshold and well below the western one and the Landsat-based estimate. Those figures are MODIS in North America, and no comparable measurement exists for the Russian register; establishing one is the natural next experiment.

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.

We therefore build a benchmark whose target is the administrative event itself. The task is to forecast, for each cell of a national grid and each day of the fire season, whether a forest fire will be entered in the federal register.

The contribution is fourfold. A national benchmark on registered labels: 64,620 cells of about 253 km^2 covering 83 regions of Russia selected by the continuity of their fire records, daily across the April to October seasons of 2000 to 2024, with 420,995 registered forest fires and a positive rate of 0.11% of cell-days, rising to 0.133% in the test years. Inputs that may be redistributed, so the dataset can be published whole rather than as a recipe, and the same sources can serve the model in real time. A decision protocol instead of a panel score: evaluation on the complete national grid at a fixed alarm budget, with rare-event verification scores and a comparison against two operational fire-danger indices and two climatologies. And a representation study: on identical rows and paired seeds we find no evidence that the hand-designed fire-weather summaries are necessary.

2 Related work

Forecasting on agency records is not new. Occurrence prediction from agency fire records has a long statistical literature, beginning with logistic models of daily people-caused occurrence in Ontario (Martell et al., 1987) and continuing through calibrated statistical and machine-learning comparisons on provincial records (Phelps & Woolford, 2021a), with guidance on how such models should be evaluated (Phelps & Woolford, 2021b). At national scale, the United States has operational products built on historical fire occurrence: the USGS WFPI-based Large Fire Probability issues daily 1 km forecasts across the conterminous United States for fires above 500 acres, calibrated on 2001 to 2017 occurrence data (U.S. Geological Survey, 2026), and earlier work forecasts distributions of large-fire counts on federal lands (Preisler et al., 2009). Closest to this work, Earnest et al. (2024) forecast fire occurrence across the conterminous United States from agency records at 40 km with lead times to ten days. Provincial systems are operational elsewhere: Ontario runs a human-caused occurrence prediction system in daily use (Woolford et al., 2021). What we could not find, and what this paper provides, is an *open benchmark* on registered labels: a released dataset, a fixed split, a decision protocol on the complete national grid and a harness, for a country where none existed. The novelty is the artifact and the protocol rather than the idea of forecasting registered fires, and we make no claim of priority on the latter.

Satellite targets dominate the benchmarks. Table 8 lists the public wildfire datasets whose label provenance we could establish. Every one takes its target wholly or partly from a satellite product: active-fire detections from MODIS or VIIRS, or burned-area maps that are themselves satellite-derived; the Canadian Fire Spread Dataset combines agency and Landsat perimeters with satellite hotspots. None is built on agency point records of occurrence, which is the target here. Several of these datasets address spread or detection rather than occurrence, where a satellite target is the natural choice. Operationally, the ECMWF probability-of-fire system predicts the probability of at least one MODIS active-fire detection in a cell of about 9 km (McNorton et al., 2024), which is explicitly a different quantity from the event a fire service registers. A second consequence of the satellite habit is visible in the reported class balance: these datasets state positive rates after resampling, so the natural prevalence of the task is usually not recoverable from the paper.

The choice of label matters empirically. Bakke et al. (2023) report that replacing a satellite-derived target with a national fire record raised test performance considerably in Norway, which they attribute to the small fires satellites miss. Their primary target was satellite-derived and the national record entered a separate analysis, so this is consistent with the premise of this benchmark rather than proof of it: a higher score on a different label with a different base rate shows the labels differ, not which is right. Russian machine-learning work on fire has also used satellite-derived targets (Shadrin et al., 2024), while remote-sensing work has matched satellite fire areas to register records (Kotelnikov & Loupian, 2022).

Designed fire-weather summaries. Operational fire-danger rating and much of the occurrence literature rest on hand-designed summaries of the weather series, above all the Canadian Fire Weather Index System (Van Wagner, 1987) and, in Russia, the statutory Nesterov-type index (Nesterov, 1949; Federal Forestry Agency of Russia, 2011). Deep-learning work has begun to learn from daily se-

ries directly: Kondylatos et al. (2022) compare recurrent and convolutional-recurrent models against tree ensembles and the Fire Weather Index for next-day danger in Greece. In that comparison the tree ensembles received the previous day's values together with a ten-day mean, while the sequence models received the series itself, so the variables and the window were shared but the representation was not. We test the narrower question of what happens when every arm receives an identical representation.

Rare-event verification. We import the verification practice of operational meteorology rather than inventing scores: the Peirce skill score (Peirce, 1884), the symmetric extremal dependence index, which is not degenerate as the base rate falls (Ferro & Stephenson, 2011), the fractions skill score for neighbourhood credit (Roberts & Lean, 2008), and relative economic value under a cost-loss model (Richardson, 2000). Capture inside a fixed alarm budget, our headline metric, is one point of the Molchan error diagram used in earthquake forecasting (Molchan, 1990; Zechar & Jordan, 2008).

3 THE BENCHMARK

3.1 LABELS

Labels come from the federal register of natural fires kept in the Russian remote-monitoring information system, as republished under CC BY 4.0. The 83 regions hold 476,762 recorded forest fires; 423,240 begin in the April to October season, and the 420,995 of those whose coordinates fall inside the grid are the labels. Our snapshot ends on 12 September 2024, mid-season; no day beyond it is scored, because absence of data is not absence of fire. The agency keeps the register continuously, but we found no public live feed of per-fire records, so the benchmark is built on this fixed snapshot. Six region labels in the register are excluded because it holds only one to six years of records for each, against at least nine for every included region and all twenty-five for 73 of them. The criterion is record continuity alone; the excluded labels are listed, as they appear in the source data, in the data card.

Two properties of the register matter for anyone using this benchmark, and we state both.

It is a mixed-source record, not a ground-truth record. Russian procedure allows fires in remote and control zones to be entered from satellite observation alone, without a site visit (Kotelnikov et al., 2019). Of the labels, 47.0% begin in aviation-monitored districts, 18.3% in ground-monitored zones, 9.4% under space monitoring and 4.8% in control zones where suppression may be suspended, the rest on other land categories. About one label in seven therefore begins where observation is space-based or suppression optional. That is a share of zones, not of detection methods: the register does not record how each fire was first seen, and a fire in any zone may first have been seen from space.

What counts as a registered fire changed in 2017. Over the whole period the median fire burns 53 ha and 0.1% are under one hectare, but that median hides a break. From 2000 to 2016 the seasonal median ranges from 48 to 93 ha and at most 0.8% of fires are under ten hectares; from 2018 it ranges from 20 to 25 ha and 30 to 39% are under ten hectares, with 2017 between the two (Table 7 in the appendix). The pattern is what one expects when small fires start being recorded, and the system's multi-year archive is documented as inhomogeneous (Kotelnikov & Loupian, 2022); we found no regulation that dates the change, so we report the break rather than explain it.

Three consequences matter for anyone using this benchmark. The evaluated years have a median fire of 22 ha, not 53 ha. The base rate is 0.105% of cell-days before 2021 and 0.133% in the test years, and the change in recording is the obvious candidate for that difference. And the released baseline, fitted through 2020, sees both regimes while its training years are dominated by the earlier one. Two things limit the damage: capture in the test years is nearly flat across fire size, including 79.8% on fires of 1 to 10 ha that barely existed in the early training years (Table 4), and per-season capture is reported rather than averaged away (Table 7). What the paper cannot do is measure skill across the break, because every test year lies on one side of it.

How a fire becomes a label. A fire is assigned to the cell containing its recorded coordinates and to the start date recorded in the register. The occurrence label is binary, so several fires in one cell-day count once; where an area is needed, it is the largest fire in that cell-day, never the sum.

Table 1: Every input, its source and its licence. No input carries a share-alike or non-commercial term, which is what allows the dataset to be redistributed whole rather than as a download recipe.

Input	Source	Licence
Fire labels	Federal fire register	CC BY 4.0
Weather	Open-Meteo archive	CC BY 4.0
Terrain	Copernicus DEM	Free, notice required
Land cover and its change	MODIS MCD12Q1	No restrictions
Peat	Northern peatland maps	CC BY 4.0
Surface water	JRC Global Surface Water	CC BY 4.0
Population and built-up	GHSL	CC BY 4.0
Administrative boundaries	Natural Earth	Public domain

3.2 Grid, weather and features

The grid is a hexagonal tessellation at a mean cell area of 252.9 km^2 (Uber Technologies, 2018), giving 64,620 cells over the included regions. Weather is sampled at 1,581 fixed points, the centres of the cells two resolutions coarser, so about 41 cells share each point, every cell lies within 68 km of its weather, and the sampling geometry never changes with time. That resolution is a limit on every arm that reads the weather, including both fire-danger indices, and Section 7 returns to what it means for the comparison. Weather comes from an open archive of reanalysis and operational model output (Hersbach et al., 2020; Muñoz-Sabater et al., 2021); Section 7 describes that mixture. Daily variables cover temperature, dew point, humidity, precipitation, snow, wind, and a drying-demand block of evapotranspiration, vapour-pressure deficit, radiation and sunshine; the statutory index and the Canadian Fire Weather Index codes are recomputed from those same values, the latter following the published equations (Van Wagner, 1987; Van Wagner & Pickett, 1985), with an implementation that agrees with an independent one to a correlation of 0.9999 across 20 points and every calendar day of the period (Bourgault et al., 2023). That is a check on our code, not on the index: both implementations approximate noon values with daily maxima, and both use standard day-length tables at latitudes they were not written for.

The released baseline is a gradient-boosted decision tree (Ke et al., 2017) fitted as three separate models on the same features: a ranking head, whose score orders cells and produces every capture number here; a rate head, whose Poisson output is a fire count per cell-day; and a large-fire head, trained on cell-days with a fire of at least 100 ha. Decisions are quantiles of a trailing 30-day window of the model's own scores, so they need no same-day labels and survive drift in the level of the score.

Each cell-day carries 106 features: terrain, peat, surface water, land cover and its change, population and settlement distance, the cell's own fire history and that of its neighbours, calendar position, and the daily weather with its rolling summaries, of which 25 are weather and 81 are not. The data card and dictionary in the release document every column, its units and its source.

3.3 Splits and what is released

The modelling panel keeps every fire cell-day and samples ten non-fire cell-days for each, so panel ROC and precision-recall in Table 6 and Table 5 are computed at that ratio and are not comparable with rates on the full grid; every decision metric is computed on the complete grid instead. The split is temporal and fixed (Figure 2 in the appendix): training 2000 to 2017, validation 2018 to 2020, test 2021 to 2024, with no shuffling. Released baselines are fitted on training and validation years together, as an operational model would be, and no test-year row is used in fitting.

The release is a frozen, checksummed version 1: the cell table with static features, the daily weather at the 1,581 sampling points, the label records with their coordinates, dates, areas and protection zones, the split assignment, and the ready modelling panel, together with the scripts that rebuild each of them from public sources. The frozen copy rather than the rebuild path is the benchmark, because access to the upstream portals is not assured everywhere. Every number in this paper is computed from version 1, and any later version will be published alongside it rather than in place of it. It also contains the evaluation harness, which takes a file of cell, date and score and returns the metrics used here. Its contract is fixed so that a competing method is scored exactly as ours

Table 2: The three operating points in this paper. They answer different questions and are not interchangeable.

Operating point	Rule	Where used
15.2% of the grid	the day's highest-ranked cells	capture, Table 3, Figure 1
4.0% of cell-days	trailing 30-day quantile, class at least 4	rare-event scores
0.99% of cell-days	the index's own class-5 share, per region	Section 5.1

was: capture is pooled over fire cell-days rather than averaged over days; days with no registered fire are dropped; ties are broken at random from a fixed seed, which matters for arms with many equal scores; the alarm budget is a share of the national grid, with per-region budgets available as an option; and rare-event scores are computed at a threshold the submitter supplies or, by default, at the budget itself. Submissions must respect the information set in Section 4: no within-season fire history and no static layer from a later epoch than the forecast year. A data card and dictionary document every column, the licence of every input, and the label biases described above.

4 Evaluation protocol

Panel metrics on sampled negatives are reported for completeness, but every decision metric is computed on the complete national grid: 807 days, every day of the April to October seasons of the four test years, 69,607 fire cell-days in total. Each evaluated day is preceded by a warm-up period that fills the trailing decision window, so all arms share one operating point.

The information set. A forecast for day D may use that day's weather, the static layers as of the latest epoch not later than D's year, and fire history built from fires registered strictly before D's year. Within-season fire history is therefore never available to the model. This is a property of the benchmark rather than of the agency: the register is kept continuously, but no live public feed of it exists, and a competitor working from the released files has the same information we did. Rows older than a layer's first epoch take that epoch, and the rule admits a same-year layer in 2000, 2001, 2005, 2010, 2015, 2020 and 2023; only 2023 is inside the test period. A land-cover composite spanning 2023 could in principle encode that season's burns. We see no sign of it, since 2023 is the second weakest test season, but that is weak evidence and a rescore against the 2020 composite is owed. Using the analysis-day weather makes this a same-day task, and a deployed system substitutes a forecast for it, which costs skill we do not measure here.

The budget is the one the deployed service uses, set by a cost-loss rule rather than tuned on these results. A fixed daily share has a known consequence: it tests which *cells* to watch on a given day and never which days matter, because the same share is flagged in October as at the July peak. The trailing-quantile rule in Table 2 is the one that varies with the season, and its scores are reported alongside.

The headline metric is **capture at a fixed alarm budget**: rank every cell for a given day, flag the top 15.2%, and report the share of that day's fire cells inside the flagged set, which answers an operational question rather than a statistical one. We report alongside it the number of alarms per fire cell caught, top-K recall for the K highest-ranked cells nationally, and rare-event scores at the shipped alarm threshold. Three decision rules appear in this paper (Table 2). Every capture at the headline budget ranks the day's cells and flags the top share; the rare-event scores use the deployed rule, a trailing 30-day quantile of the model's own scores, which needs no same-day labels; and Section 5.1 gives the model, region by region, as many cell-days as the statutory index puts in class five. Ties are broken at random with a fixed seed, which matters for the indices, where about 41 cells share one weather point and therefore one value. Confidence intervals are bootstrapped over *days*; days are treated as the independent unit, which understates dependence between neighbouring days within a synoptic regime.

5 Results

Table 3 gives the main comparison at the operating point and Figure 1 the whole curve. The model's interval is disjoint from every baseline's, although the two climatologies overlap each other. It costs

Table 3: Capture of registered fire cells at an equal alarm budget of 15.23% of the grid, on all 807 season days of the test period (69,607 fire cell-days). Intervals are 95% bootstrap over days. The last column is the share of the day's fire cells inside the 100 highest-ranked cells nationally, averaged over days.

Cells flagged by	Capture (%)	95% CI	Alarms / fire cell	Top-100 (%)
This benchmark's baseline	80.5	[79.1, 81.8]	142	5.7
Climatology, season	63.7	[61.7, 65.7]	179	2.0
Climatology, day window	62.5	[60.2, 64.5]	183	3.4
Canadian FWI	53.6	[51.9, 55.1]	213	0.4
Nesterov index (statutory)	47.9	[46.4, 49.3]	238	0.3
Random	15.1	[14.9, 15.4]	754	0.2

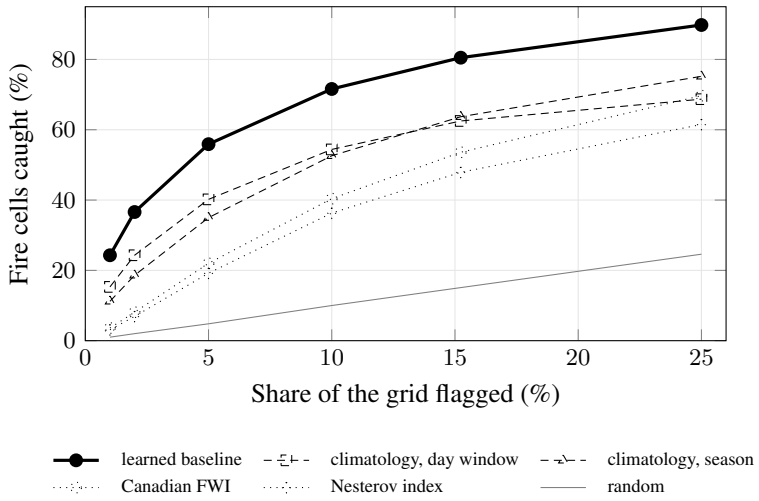

Figure 1: Capture against alarm budget on all 807 test-season days, one point of which is the operating point used in Table 3. The learned baseline leads at every budget, and by the widest relative margin at the tightest ones.

142 alarms per fire cell caught, and it is better than climatology on 97% of individual days and worse on 2%. The learned baseline leads at every budget we scored: at 1% of the grid it catches 24.3% against 15.3% for the day-window climatology, 3.6% for the Canadian index and 3.1% for the statutory one, and at 25% it reaches 89.8%. The two climatologies swap places along the curve, so the ordering that is stable is the model above both climatologies above both indices.

The two climatologies are defined as follows. The seasonal one scores a cell by how often it burned anywhere in the evaluated part of the year across 2003 to 2020, so it never sees a test year; the day-window one scores it by fires within a fortnight of the same calendar date in every earlier year, so for later test years it does see earlier ones. Neither is smoothed. The blunter arm captures more at the operating point and less at the sharp end, and their intervals overlap, so we draw no conclusion from the difference between them.

Two observations deserve emphasis. First, **climatology, not the fire-danger indices, is the competitor to beat**: at this budget, a static map of where fires have occurred in this season historically outscores both indices by ten points or more. Any claim of skill in this setting should be made against climatology. Second, and this limits what the comparison can claim, both indices are computed on the 1,581-point weather grid, so about 41 cells share one value and ties are broken at random. They cannot express terrain, land cover, population or fire history at cell level, and part of the gap is therefore resolution rather than fire science. The comparison answers whether an operational index used as a ranking beats a learned model on this grid, not whether the index is a poor index. The same limit applies at the sharp end. Watching the 100 highest-ranked cells in the country, about 0.15% of the grid, the learned baseline finds 5.7% of the day's fire cells, against 0.3 to 0.4% for either index.

Table 4: Capture by size, where size is the largest fire registered in that cell-day, at the same national alarm budget on the full grid.

Fire size	Fire cell-days	Capture (%)
<1 ha	4	75.0
1-10 ha	23,097	79.8
10-25 ha	12,190	77.1
25-100 ha	17,931	81.5
100-200 ha	2,869	83.8
200-1000 ha	8,743	84.2
≥1000 ha	4,773	79.8

Is the gap weather, or place and season? A model that reads terrain, land cover, population and a cell's own fire history could in principle rank cells well without using the weather at all, in which case the comparison against fire-danger indices would be unfair in a way that flatters us. We therefore refit the identical model on the identical rows with all 25 weather features removed, leaving the 81 that describe place, history and calendar. Averaged over three seeds, ROC falls from 0.9198 to 0.8801 and precision-recall from 0.6108 to 0.5049, a loss of 0.0398 ROC and 0.1059 precision-recall. The daily weather therefore carries a large part of the skill, more than an order of magnitude more than the differences between representations in Section 6, and the task is not a static risk map in disguise.

At the shipped alarm threshold, the rare-event scores are SEDI +0.503 and Peirce skill +0.277 on a base rate of 0.133%, with recall 31.6% while alarming 4.0% of cell-days. The frequency-matched fractions skill score rises from 0.052 at the cell to 0.160 over a two-ring neighbourhood, so part of what the cell-level scores count as a miss is a fire one or two cells away; both values are far below the level usually called useful in that literature (Roberts & Lean, 2008). Relative economic value at this threshold is positive only for cost-loss ratios between 0.00095 and 0.0105 and, as for any single threshold, peaks at the Peirce score where the ratio equals the base rate, so the warning pays for users whose cost of acting is small against their loss. No baseline is scored on either measure here, so they describe this model rather than rank it.

The second head emits a rate rather than an ordering, and it is reported here because a probability invites a different test. It ranks almost as well as the ordering head (ROC 0.9093), but its Brier score of 0.001381 gives a skill of -0.036 against climatology: at this base rate the score is dominated by uncertainty and its decomposition is unstable to the choice of bins, so we report the score and not a calibration claim. Every decision in this paper uses the ordering rather than the probability.

Capture is nearly flat across three orders of magnitude of fire size (Table 4): fires of 1 to 10 ha are caught at 79.8%, and fires above 1000 ha at 79.8%. The benchmark is therefore not a large-fire task with extra steps. Across seasons capture ranges from 75.6% to 84.5%, so the average hides no collapsed year (Appendix B). Across the 83 regions the median is 81.4% and the best is 97.7%. The weakest regions with at least a hundred fire cell-days are Komi at 34.5%, Rostov at 40.8%, Karelia at 42.6%, Tuva at 45.3% and Chukotka at 47.7%, none of them among the largest fire regions. A national budget is the likely reason, since a region loses on days when its fires are real but modest against the national ranking, but we have not run the per-region alternative and so cannot attribute the difference. The harness supports per-region budgets and we report this as an open question.

5.1 Against the statutory danger classes

Comparing a model against an index at a budget we choose is fair only if the budget does not flatter us, so we repeat the comparison on the index's own terms, on the full grid. Within each region, the statutory index reaches class five, the federal boundary of 10,000 (Federal Forestry Agency of Russia, 2011), on some of the test period's cell-days, and the model flags exactly as many there, its highest-ranked ones; together these are 0.99% of the cell-days. Held to that tempo, the share of fire cell-days flagged rises from 2.74% for the index to 16.71%, a factor of 6.1; the model is ahead in 67 of the 71 regions scored, level in 3 and behind in 1. No label sets any threshold, but the index's

Table 5: Heads trained at three size thresholds, each scored at all three as ROC / PR on the large-fire label, and in the last column as ROC on any fire. Panel metrics on the test split, with non-fire cell-days sampled at ten per fire cell-day.

Head	≥25 ha	≥100 ha	≥200 ha	Any fire
Large-fire head, 25 ha	0.9194 / 0.3985	0.9230 / 0.2252	0.9222 / 0.1862	0.9086
Large-fire head, 100 ha	0.9132 / 0.3665	0.9287 / 0.2380	0.9300 / 0.2047	0.8831
Large-fire head, 200 ha	0.9091 / 0.3447	0.9279 / 0.2282	0.9299 / 0.1979	0.8748
Occurrence head	0.9108 / 0.3611	0.9054 / 0.1785	0.9032 / 0.1445	0.9197

Table 6: Learned against designed representations: identical rows, one million training examples, three paired seeds, evaluated on the test split as panel metrics with non-fire cell-days sampled at ten per fire cell-day. The left pair fits on the training years only, the right pair fits on training and validation years, as the released baseline does.

Arm	ROC	PR	ROC	PR
Gradient boosting, raw 30-day window	0.9151	0.5897	0.9212	0.6101
Recurrent model, raw 30-day window	0.9155	0.5945	0.9186	0.6036
Gradient boosting, designed summaries	0.9130	0.5851	0.9184	0.6065
Graph over weather channels	0.9089	0.5714	0.9129	0.5873

tempo over a season is known only afterwards. Regions with fewer than 50 fire cell-days are not scored, nor are the 6 where the index never reaches class five.

The mechanism is structural rather than statistical. The statutory index accumulates day by day and resets on rain, so on this grid it peaks in August, while April is the biggest fire month: 33.8% of the labels and 31.3% of those in the test years begin in April, when the index is still small. A month-aware or region-aware recalibration of the index might recover part of that, and we have not tested one.

5.2 Large fires

Russian rules define a large forest fire as 25 ha or more in ground-protection districts and 200 ha or more in aviation districts and control zones (Ministry of Natural Resources and Environment of Russia, 2022). We train a dedicated head at a single national threshold of 100 ha and report all three (Table 5). The dedicated head beats the occurrence head on large fires at every threshold, by +0.0233 ROC and +0.0595 precision-recall at 100 ha, and it is correspondingly worse at ranking any fire, which is why both heads are kept. Training at the rarer 200 ha target does not help: that head has 70,865 positives to learn from against 114,312, and it loses to the 100 ha head at its own threshold. At 25 ha, matching does pay, by 0.032 precision-recall.

6 Learned against designed representations

Does the field need its hand-designed fire-weather summaries? We compare four arms on identical rows, one million training examples, three paired seeds: a gradient-boosted tree on the designed feature set, the same learner on the raw 30-day weather window flattened, a recurrent model over that window, and a graph model whose nodes are the weather variables rather than places, with a learned adjacency between them, each arm receiving the same static features. Table 6 reports both fits: on the training years alone, and on training plus validation as the released baseline is fitted.

We find no evidence in this benchmark that the designed summaries are necessary. They include rolling precipitation, dry-spell length, days since snowmelt and the statutory index with its accumulator state, and the same learner given the raw series instead leads by 0.0021 ROC when fitted on the training years and by 0.0028 when fitted like the released model, in three of three seeds each. The differences are small, three seeds cannot establish much on their own, and these are panel metrics on sampled negatives rather than full-grid capture, so the claim is an absence of evidence for the

summaries rather than evidence against them. The two neural arms do not settle the matter: the recurrent model leads the flat tree when both are fitted on the training years and falls behind it when both are fitted as the released model is, so what we can claim is no consistent difference between them rather than an improvement in either direction. The channel graph is behind in both fits. The full raw window also carries eight provider variables the designed set never uses, so it is not a pure representation comparison. Restricted to the twelve variables the designed features are computed from, and fitted on the training years, the flat tree scores 0.9146 against 0.9130 for the designed set, so the conclusion does not depend on the extra variables.

7 LIMITATIONS

Two seams sit on the split boundary. The weather product changes in 2017 and the registration practice changes in 2017, while the training split ends in 2017. Every test number therefore crosses both, and this paper cannot separate generalisation from either shift. The pinned re-pull will settle the first; the second can only be addressed by evaluating within one regime, which we leave to users of the benchmark.

The comparison against the indices is limited by resolution. Both indices are computed on the 1,581-point weather grid, so roughly 41 cells share each value while the model reads cell-level terrain, land cover, population and fire history. Part of the gap is therefore spatial resolution rather than fire science, and a resolution-matched arm is the right next comparison.

The weather record is not a single product. The archive's default selection combines reanalysis with operational model output from 2017 onward. Measured at six of our grid points, 2014 values are identical to a pure reanalysis request while 2022 values differ on most days. Every arm in every comparison reads the same weather, and the released baseline is fitted through 2020 so it has seen both regimes, but absolute skill is measured across a product seam. A re-pull with the model pinned is in progress and will be reported.

Skill is measured on analysis-day weather. Decay with forecast lead time is not measured. Archived forecasts at fixed lead times exist for these points only from 2024, which covers part of one season.

Absolute probabilities cannot be recalibrated live, because no live public feed of the register exists. Ordering, the alarm and the danger class are scale-invariant and unaffected. The label-shift literature (Lipton et al., 2018) assumes a structure that does not hold for weather-driven fire, so the exposure here is concept shift in registration practice, which cannot be corrected without labels.

The labels are administrative. They inherit the register's own biases: a mixed observation basis (Kotelnikov et al., 2019), almost no fires under one hectare, and fire-area distributions that differ with the level of fire protection (Kotelnikov & Loupian, 2022). Satellite estimates of burned area exceed official statistics by three to five times in Russia (Krylov et al., 2014), which is an area comparison and not a count comparison.

8 CONCLUSION

The register a fire service answers for turns out to be forecastable at national scale from inputs that can be redistributed in full. At a budget of 15.2% of the grid, a plain gradient-boosted baseline finds 80.5% of the day's fire cells, ahead of climatology and of our reconstructions of the two fire-danger indices, with capture that varies little across fire size. In this benchmark the hand-designed summaries behave as a convenient encoding rather than a necessary one, although the margins are small and the test is one dataset. We release the dataset, the splits, the evaluation harness and the baselines so that both findings can be contested.

AI USE STATEMENT

In this work, we used generative AI tools for research execution: to implement methods and run experiments, to clean and reformat the dataset, to retrieve, translate and summarise related work, and to summarise results. We also used them to draft sections of the paper and to polish its writing. We did not use them to generate synthetic data or to formulate or prove mathematical claims.

We have reviewed all AI-assisted work. The tables were generated from the stored result files by a script, and the numbers in the text were checked against those files; claims attributed to other studies were checked against the full text where we could obtain it and otherwise against the abstract; and bibliographic records were checked against Crossref and OpenAlex. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

The dataset contains no personal data. Labels are administrative records of forest fires published under an open licence, and the remaining inputs are openly licensed environmental and demographic rasters; the release carries the attribution notices those licences require. The benchmark describes where fires are registered, which is public information already published by the responsible agency.

Two risks deserve statement. First, a forecast of where fires will be registered is not a forecast of where fires will occur, and using it as though it were could misdirect resources away from areas that are poorly monitored; the label biases in Section 3.1 are reported for exactly this reason. Second, any system that ranks places for attention can be used to justify withdrawing attention from the rest. The alarm budget is a policy choice, not a property of the model, and we report cost as alarms per fire cell caught so that the trade is visible rather than hidden inside a threshold.

Reproducibility statement

The release described in Section 3.3 is available at the anonymous link `https://osf.io/5ksx9/overview?view_only=6536660696cc4755aa2154affb220769`, with its data card, licence notices and checksums. It contains the data files and the fixed split, the scripts that rebuild them from public sources, the training and evaluation code with a pinned environment, the evaluation harness, and the result files behind every table together with the script that regenerates the tables from them. Section 4 states the protocol in full, including the alarm budget, the warm-up period and the bootstrap unit. The limitations in Section 7 name the one input whose provenance is not reproducible from a single pinned request, and the pinned re-pull that addresses it.

References

Sigrid Jørgensen Bakke, Niko Wanders, Karin van der Wiel, and Lena Merete Tallaksen. A data-driven model for Fennoscandian wildfire danger. *Natural Hazards and Earth System Sciences*, 23 (1):65–89, 2023. doi: 10.5194/nhess-23-65-2023.

Quinn E. Barber, Piyush Jain, Ellen Whitman, Dan K. Thompson, Luc Guindon, Sean A. Parks, Xianli Wang, Matthew G. Hethcoat, and Marc-André Parisien. The Canadian fire spread dataset. *Scientific Data*, 11(1):764, 2024. doi: 10.1038/s41597-024-03436-4.

Pascal Bourgault, David Huard, Travis James Smith, et al. xclim: xarray-based climate data analytics. *Journal of Open Source Software*, 8(85):5415, 2023. doi: 10.21105/joss.05415.

Kadir Alperen Coskuner. Assessing the performance of MODIS and VIIRS active fire products in the monitoring of wildfires: A case study in Turkey. *iForest - Biogeosciences and Forestry*, 15 (2):85–94, 2022. doi: 10.3832/ifor3754-015.

Jeffrey Earnest, Amy McGovern, Christopher Karstens, and Israel Jirak. Part I: Improving wildfire occurrence prediction for CONUS using deep learning and fire weather variables. *Artificial Intelligence for the Earth Systems*, 3(3), 2024. doi: 10.1175/AIES-D-23-0057.1.

Federal Forestry Agency of Russia. Order No. 287 of 5 July 2011 on the Classification of Fire Danger in Forests According to Weather Conditions, Appendix 2, 2011.

Christopher A. T. Ferro and David B. Stephenson. Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26(5): 699–713, 2011. doi: 10.1175/WAF-D-10-05030.1.

Emily J. Fusco, John T. Finn, John T. Abatzoglou, Jennifer K. Balch, Sepideh Dadashi, and Bethany A. Bradley. Detection rates and biases of fire observations from MODIS and agency

reports in the conterminous United States. *Remote Sensing of Environment*, 220:30–40, 2019. doi: 10.1016/j.rse.2018.10.028.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. WildfireSpreadTS: A dataset of multi-modal time series for wildfire spread prediction. In *Advances in Neural Information Processing Systems 36*, 2023.

Todd J. Hawbaker, Volker C. Radeloff, Alexandra D. Syphard, Zhiliang Zhu, and Susan I. Stewart. Detection rates of the MODIS active fire product in the United States. *Remote Sensing of Environment*, 112(5):2656–2664, 2008. doi: 10.1016/j.rse.2007.12.008.

Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Fantine Huot, R. Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yifan Chen. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. doi: 10.1109/TGRS.2022.3192974.

Ilektra Karasante, Lázaro Alonso, Ioannis Prapas, Akanksha Ahuja, Nuno Carvalhais, and Ioannis Papoutsis. SeasFire cube: A multivariate dataset for global wildfire modeling. *Scientific Data*, 12 (1):368, 2025. doi: 10.1038/s41597-025-04546-3.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pp. 3146–3154, 2017.

Spyridon Kondylatos, Ioannis Prapas, Gustau Camps-Valls, and Ioannis Papoutsis. Mesogeos: A multi-purpose dataset for data-driven wildfire modeling in the Mediterranean. In *Advances in Neural Information Processing Systems 36*, 2023.

Spyros Kondylatos, Ioannis Prapas, Michele Ronco, Ioannis Papoutsis, Gustau Camps-Valls, Maria Piles, Miguel-Ángel Fernández-Torres, and Nuno Carvalhais. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17):e2022GL099368, 2022. doi: 10.1029/2022GL099368.

Roman V. Kotelnikov and Evgeny A. Loupian. Features of remotely estimated distributions of forest fire areas for territories with different levels of fire protection. *Sovremennye Problemy Distantsionnogo Zondirovaniya Zemli iz Kosmosa*, 19(4):75–87, 2022. doi: 10.21046/2070-7401-2022-19-4-75-87.

Roman V. Kotelnikov, Evgeny A. Loupian, Sergey A. Bartalev, and Dmitry V. Ershov. Space monitoring of forest fires: History of the creation and development of ISDM-Rosleskhoz. *Lesovedenie, no. 5*, pp. 399–409, 2019. doi: 10.1134/S0024114819050048.

Alexander Krylov, Jessica L. McCarty, Peter Potapov, Tatiana Loboda, Alexandra Tyukavina, Svetlana Turubanova, and Matthew C. Hansen. Remote sensing estimates of stand-replacement fires in Russia, 2002–2011. *Environmental Research Letters*, 9(10):105007, 2014. doi: 10.1088/1748-9326/9/10/105007.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. *Proceedings of Machine Learning Research*, 80:3122–3130, 2018.

David L. Martell, Simon Otukol, and Brian J. Stocks. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Canadian Journal of Forest Research*, 17(5):394–401, 1987. doi: 10.1139/x87-068.

Joe R. McNorton, Francesca Di Giuseppe, Ewan Pinnington, Matthew Chantry, and Christopher Barnard. A global probability-of-fire (PoF) forecast. *Geophysical Research Letters*, 51(12): e2023GL107929, 2024. doi: 10.1029/2023GL107929.

Ministry of Natural Resources and Environment of Russia. Order No. 244 of 1 April 2022 Approving the Rules for Extinguishing Forest Fires, 2022.

G. M. Molchan. Strategies in strong earthquake prediction. *Physics of the Earth and Planetary Interiors*, 61(1-2):84–98, 1990. doi: 10.1016/0031-9201(90)90097-H.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

V. G. Nesterov. *Gorimost' lesa i metody ee opredeleniya [Forest Flammability and Methods for Its Determination]*. Goslesbumizdat, Moscow, 1949.

Charles S. Peirce. The numerical measure of the success of predictions. *Science*, ns-4(93):453–454, 1884. doi: 10.1126/science.ns-4.93.453.b.

Nathan Phelps and Douglas G. Woolford. Comparing calibrated statistical and machine learning methods for wildland fire occurrence prediction: A case study of human-caused fires in Lac La Biche, Alberta, Canada. *International Journal of Wildland Fire*, 30(11):850–870, 2021a. doi: 10.1071/WF20139.

Nathan Phelps and Douglas G. Woolford. Guidelines for effective evaluation and comparison of wildland fire occurrence prediction models. *International Journal of Wildland Fire*, 30(4):225–240, 2021b. doi: 10.1071/WF20134.

Haiganoush K. Preisler, Robert E. Burgan, Jeffery C. Eidenshink, Jacqueline M. Klaver, and Robert W. Klaver. Forecasting distributions of large federal-lands fires utilizing satellite and gridded weather information. *International Journal of Wildland Fire*, 18(5):508–516, 2009. doi: 10.1071/WF08032.

David S. Richardson. Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563):649–667, 2000. doi: 10.1002/qj.49712656313.

Nigel M. Roberts and Humphrey W. Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, 2008. doi: 10.1175/2007MWR2123.1.

Dmitrii Shadrin, Svetlana Illarionova, Fedor Gubanov, Ksenia Evteeva, Maksim Mironenko, Ivan Levchunets, Roman Belousov, and Evgeny Burnaev. Wildfire spreading prediction using multimodal data and deep neural network approach. *Scientific Reports*, 14, 2024. doi: 10.1038/s41598-024-52821-x.

Uber Technologies. H3: A hexagonal hierarchical geospatial indexing system. `https://h3geo.org`, 2018.

U.S. Geological Survey. WFPI-based large fire probability (WLFP). `https://www.usgs.gov/fire-danger-forecast/wfpi-based-large-fire-probability-wlfp`, 2026.

C. E. Van Wagner. Development and structure of the Canadian Forest Fire Weather Index system. Forestry Technical Report 35, Canadian Forestry Service, Ottawa, 1987.

C. E. Van Wagner and T. L. Pickett. Equations and FORTRAN program for the Canadian Forest Fire Weather Index system. Forestry Technical Report 33, Canadian Forestry Service, Ottawa, 1985.

Douglas G. Woolford, David L. Martell, Colin B. McFayden, Jordan Evens, Aaron Stacey, B. Michael Wotton, and Dennis Boychuk. The development and implementation of a human-caused wildland fire occurrence prediction system for the province of Ontario, Canada. *Canadian Journal of Forest Research*, 51(2):303–325, 2021. doi: 10.1139/cjfr-2020-0313.

Zhengsen Xu, Sibo Cheng, Lanying Wang, Hongjie He, Wentao Sun, Jie Li, and Linlin Xu. BCWildfire: A long-term multi-factor dataset and deep learning benchmark for boreal wildfire risk prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. doi: 10.1609/aaai.v40i46.41299.

J. Douglas Zechar and Thomas H. Jordan. Testing alarm-based earthquake predictions. *Geophysical Journal International*, 172(2):715–724, 2008. doi: 10.1111/j.1365-246X.2007.03676.x.

Yu Zhao, Sebastian Gerard, and Yifang Ban. TS-SatFire: A multi-task satellite image time-series dataset for wildfire detection and prediction. *Scientific Data*, 12(1):1817, 2025. doi: 10.1038/s41597-025-06271-3.

A How the benchmark is built

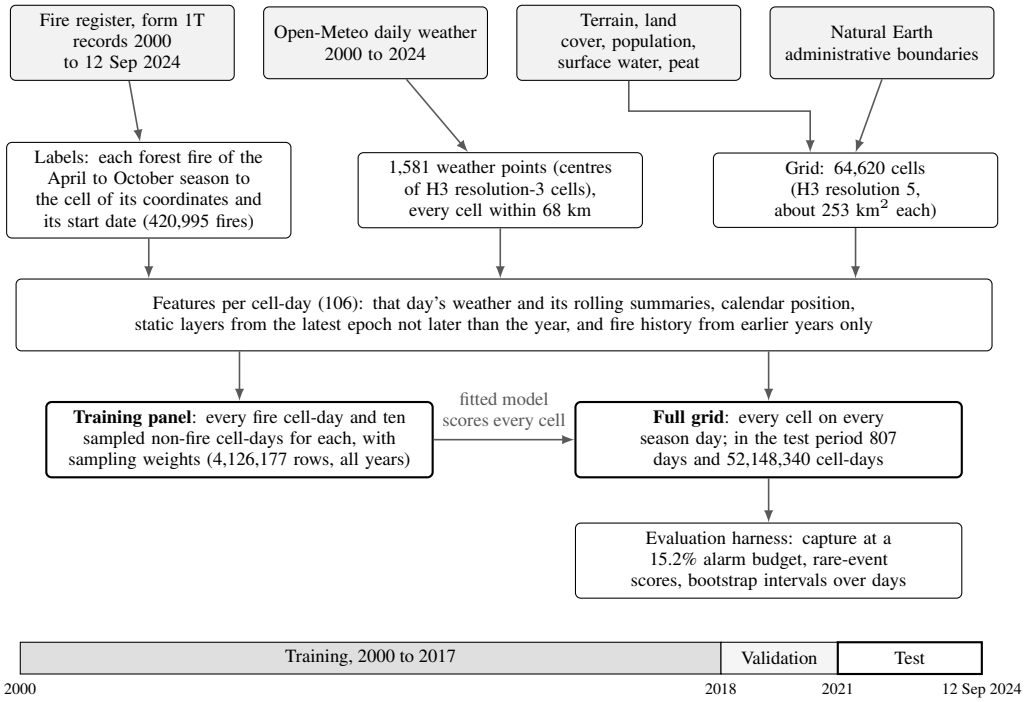

Figure 2: How the benchmark is built and split. Grey boxes are the public sources. Seasons run April to October and the split is temporal, with no shuffling. The released baseline is fitted on the sampled panel of the training and validation years and scored on the complete grid, where every decision metric is computed; the test years are only scored.

B ADDITIONAL TABLES

Table 7: Registered forest fires by season: the 420,995 labels, that is the fires of the April to October season whose coordinates fall inside the grid. The median size and the share of small fires break between 2016 and 2018 while the count of fires of 25 ha or more does not collapse, the pattern expected when small fires begin to be recorded rather than when the area field changes meaning.

Season	Fires	Median (ha)	$<$10 ha (%)	$\geq$100 ha (%)	$\geq$25 ha (n)	Capture (%)
2000	8,351	80	0.5	32.7	8,283	–
2001	4,900	48	0.5	28.5	4,871	–
2002	12,450	70	0.5	33.1	12,320	–
2003	26,846	69	0.4	37.8	26,647	–
2004	12,239	70	0.7	38.0	12,061	–
2005	11,194	70	0.5	38.8	11,079	–
2006	20,819	70	0.8	38.3	20,507	–
2007	13,392	70	0.6	38.4	13,239	–
2008	21,941	80	0.5	43.3	21,734	–
2009	19,668	66	0.5	36.4	19,444	–
2010	16,157	69	0.5	37.9	15,984	–
2011	18,691	71	0.4	39.5	18,552	–
2012	17,606	93	0.4	48.3	17,459	–
2013	9,850	74	0.4	41.1	9,772	–
2014	14,231	77	0.6	42.3	14,036	–
2015	11,174	79	0.6	43.3	11,046	–
2016	10,355	77	0.5	43.2	10,248	–
2017	14,874	47	12.0	31.4	10,226	–
2018	24,838	24	34.1	20.6	12,298	–
2019	28,456	24	34.1	22.5	14,182	–
2020	26,658	23	35.3	22.7	12,983	–
2021	24,650	20	39.2	20.9	11,109	84.5
2022	17,589	23	34.2	22.9	8,471	84.1
2023	20,357	23	32.8	22.4	9,909	76.1
2024	13,709	25	30.4	24.8	6,912	75.6

Table 8: Public machine-learning wildfire datasets and the provenance of their labels. All take their target wholly or partly from satellite products, and none from agency records of occurrence. Label sources were traced from the dataset papers, except for the three entries marked with an asterisk, which rest on secondary descriptions.

Dataset	Label source	Product	Region
Next Day Wildfire Spread (Huot et al., 2022)	Active fire	MODIS	US
WildfireSpreadTS (Gerard et al., 2023)	Active fire	VIIRS 375 m	US
WSTS+*	Active fire	VIIRS 375 m	US
TS-SatFire* (Zhao et al., 2025)	Active fire	Image series	US
SeasFire (Karasante et al., 2025)	Burned area	GWIS, GFED	Global
Mesogeos (Kondylatos et al., 2023)	Burned area	EFFIS	Mediterr.
FireCube (Kondylatos et al., 2022)	Burned area	EFFIS and MODIS	Greece
WildfireDB*	Active fire	VIIRS 375 m	US
Canadian Fire Spread (Barber et al., 2024)	Perimeters	NBAC, hotspots	Canada
BCWildfire (Xu et al., 2026)	Active fire	MODIS	Canada (BC)
This benchmark	**Agency register**	**Federal fire register**	**Russia**