

Forecasting Undercasts on Mount Washington

Lucas Ehinger

Massachusetts Institute of Technology

<https://lucasehinger.github.io/weather/>

October 2026

Abstract

An *undercast* is a cloud deck lying below a summit, leaving the summit in clear air above it. This document describes the undercast forecast published at lucasehinger.github.io/weather/ for the 6,288 ft summit of Mount Washington, New Hampshire, together with the analysis behind it. Hand-labelled webcam images give only 22 undercast days, so labels are instead derived from the remarks Mount Washington Observatory observers append to the hourly KMWV METAR. Five cuts on those remarks reach precision 0.70 and recall 0.73 against the webcam labels and yield 6,852 undercast hours among 253,317 observations from 1997–2026. A simple physics rule based on a modelled inversion reaches only precision 0.09 at a 3% base rate. The forecast in production is a single gradient-boosted tree model fed by six numerical weather prediction (NWP) sources (HRRR, RAP, NAM, GFS, ECMWF, NBM). On a held-out year with a 2.5% base rate, it reaches ROC-AUC 0.941 and precision 0.74 one hour ahead, falling to ROC-AUC 0.764 and precision 0.08 six days ahead. The limiting factor is the quality of the labels, not the classifier.

Contents

1	Introduction	2
2	The website	2
3	Labels: when was it actually undercast?	3
3.1	Webcam labels	3
3.2	Observer remarks	3
3.3	The cuts	3
3.4	The 29-year record	4
4	A physics-based baseline	6
5	Machine-learning model	7
5.1	Inputs	7
5.2	Algorithms	7
5.3	Validation design	8
5.4	Results by source and lead	8
5.5	Choice of algorithm	9
5.6	Overfitting checks	10
5.7	Production model	11

6	Limitations and next steps	15
A	Website controls and links	16
B	Per-source and per-algorithm results	16
C	Input parameters	18

1 Introduction

For hikers, undercasts are a treat and are often unexpected. The project began with the assumption that a cloud-base forecast (e.g. from Windy) would be enough to predict them. It is not. Undercasts require a temperature inversion that traps cool, moist air below the summit [1, 2], but a modelled inversion on its own is a weak predictor (section 4). Academic literature on undercasts specifically is sparse; [3, 4, 5] are useful starting points.

The website (section 2) publishes an hourly undercast forecast for Mount Washington up to six days ahead. The rest of this document covers how its training labels are built (section 3), why a physics rule fails (section 4), how the machine-learning (ML) model is built and validated (section 5), and its limitations (section 6). Interactive views and per-source breakdowns from the website are collected in the appendices.

2 The website

The forecast page, `/weather/`, has a summit selector (Mt. Washington, Mt. Lafayette, Mt. Moosilauke, Mt. Monadnock), a set of quick links to external forecasts for the selected summit (section A), a collapsible overview of the NWP sources (table 1), and two panels:

Undercast forecast. An hourly strip covering the coming forecast window. Each bar’s height is the model’s score for that hour divided by the decision threshold for that hour’s forecast lead, with a dashed line marking the threshold. Hours at or above the line are called undercast, and the panel names the first undercast window. Hours already in the past are dropped. If a required NWP source did not publish in time, the panel says *Forecast unavailable* rather than showing a quiet forecast. This is deliberate: a missing forecast must not read as “no undercast”.

Model performance. A collapsible panel with the held-out-year confusion matrix of the production model at each forecast lead (1, 24, 48, 72, 96, 120 and 144 h), using the same thresholds as the forecast (table 9).

The undercast model is trained on Mount Washington observations only (section 3) and does not transfer to the other summits. When another summit is selected, the page says the prediction is only available for Mt. Washington. A GitHub Actions job refreshes the forecast every six hours.

Four supporting pages document the analysis: *Model Details* (`/weather/details/`, the main write-up), *What the Cuts Keep and Throw Away* (`/weather/cuts/`, a webcam gallery of the label cuts), *Observation Viewer* (`/weather/observations/`, 42 METAR reports beside their webcam frames) and *What the Models Actually Show* (`/weather/historical-conditions/`, the inversion analysis plus a day-by-day viewer of the NWP profiles and the noon webcam image).

Table 1: NWP sources as summarised on the forecast page and the details page. Range is the forecast range used on the site.

Model	Range (h)	Resolution	Notes
HRRR	0–48	3 km, CONUS	High-Resolution Rapid Refresh; hourly, short range; cloud fields
RAP	0–48	13 km, regional	HRRR’s parent model; cloud fields
NAM	0–60	12 km, regional	North American Mesoscale; cloud fields
GFS	0–120	~0.25°, global	Long range, coarse; cloud fields
ECMWF	0–144	~0.25°, global	IFS open data; strong dynamics but no cloud fields; 3-hourly steps
NBM	0–144	~2.5 km, CONUS	National Blend of Models; calibrated, surface fields only, incl. probabilistic ceiling and visibility

3 Labels: when was it actually undercast?

3.1 Webcam labels

The Mount Washington Observatory (MWOBS) archives roughly two years of daily timelapses from two summit webcams (the *tower* and *observatory* cameras). Every noon frame was labelled by hand. That gives 589 days from 25 May 2024 to 3 January 2026, of which 22 are undercast. Labelling was generous: partial undercasts (visible on one camera only, or a broken deck) count as undercast. Because the timelapses fade from one day into the next, some frames blend two days. A clean frame with a matching timestamp is used wherever one exists. Twenty-two positives is too few to train a classifier, but enough to validate a labelling rule.

3.2 Observer remarks

KMWN is one of the few remaining staffed weather stations in the U.S. Every hour, observers add free-text remarks to the automated METAR. The key remark is **TPS LWR** (“tops lower”), which gives the coverage and the top height (feet above sea level) of a cloud deck below the summit. For example:

```
KMWN 211559Z 06010KT 70SM SCT/// OVC070 09/M06 RMK TPS LWR SCT040 SUN DMLY VSBL BINOV
```

Here the visibility is 70 SM (observers estimate it from distant landmarks; automated stations stop at 10). **SCT///** is a scattered layer of unknown height, and **OVC070** is an overcast lid 7,000 ft above the station. **TPS LWR SCT040** is a scattered deck topping out at 4,000 ft, 2,288 ft below the observer. **SUN DMLY VSBL** means the sun is dimly visible through the layer above, and **BINOV** means breaks in the overcast. Heights in the METAR body are above the station; **TPS LWR** heights are above sea level.

3.3 The cuts

Using the webcam days as truth, five conditions on the remarks were chosen (table 2). An hour is *undercast* if it passes all five. It is *ambiguous*, and excluded from training, if it fails at least one but still passes every looser bound in table 2. Otherwise it is *not undercast*.

Figure 1 shows the cuts added one at a time, starting from “the observer reported a deck below the summit” (21 undercasts caught, 92 false alarms) and ending at 16 caught and 7 false alarms. Nearly all the recall is present at the start; the cuts mainly strip out false alarms. Table 3 drops

Table 2: Label cuts on the KMWN remarks. Heights are relative to the 6,288 ft summit.

Test	Undercast	Ambiguous	Reason
TPS LWR coverage	BKN or OVC	also SCT	at least 5/8 of the sky below the summit is covered
Visibility	> 40 SM	> 25 SM	poor visibility means the summit is at least partly in cloud
Deck top below summit	≥ 500 ft	≥ 250 ft	a deck just below the summit means you are mostly in it
Lowest layer overhead, any cover	$\geq 1,000$ ft	≥ 500 ft	e.g. FEW002 touches the summit and passes the ceiling test
Lowest BKN/OVC ceiling overhead	$\geq 5,000$ ft	$\geq 3,500$ ft	a low lid makes a cloud “sandwich”, not an undercast

one cut at a time. Coverage is the only real trade-off: it removes 25 false alarms but costs four genuine undercasts, all decks the observer called SCT. Scored against all 589 webcam days, the full set of cuts has **precision 0.70 and recall 0.73**.

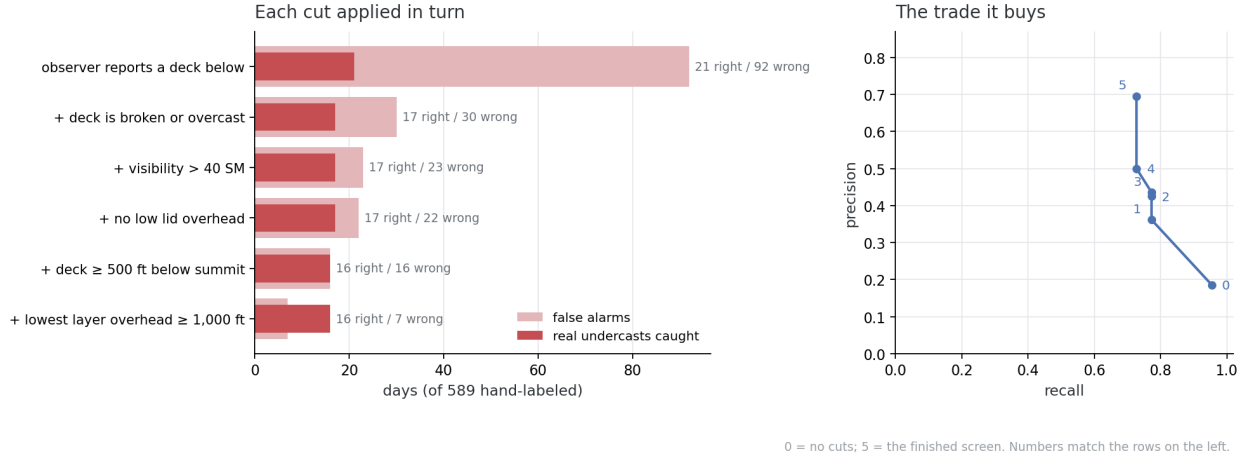


Figure 1: Cuts applied cumulatively on the webcam-labelled days. Left: undercasts caught against false alarms at each stage. Right: the same stages as precision against recall.

The cuts page shows every labelled day on which the observer reported a deck below the summit (113 days), grouped by outcome. 23 pass every cut, 16 of them true undercasts. 29 fail only on coverage (4 undercast). 1 fails only on visibility, 1 only on the lid, 3 only on depth and 9 only on the lowest overhead layer, none of them undercast. 47 fail two or more cuts (1 undercast). The gallery also shows why precision tops out near 0.70. The remark records how much of the sky a deck covers, not how much of the horizon it fills. BKN is written the same way for an unbroken floor of cloud and for scattered cumulus with open valleys, so no threshold on these numbers can tell the two apart (fig. 2).

3.4 The 29-year record

Applied to all 253,317 hourly KMWN observations from 1997 to 2026, the cuts pass 6,852 hours as undercast. That is over 300 times more positive examples than the webcam labels provide (fig. 3a). The share of passing observations rises from 1.2% in 1997 to 5.4% in 2024 (fig. 3b). This tracks a

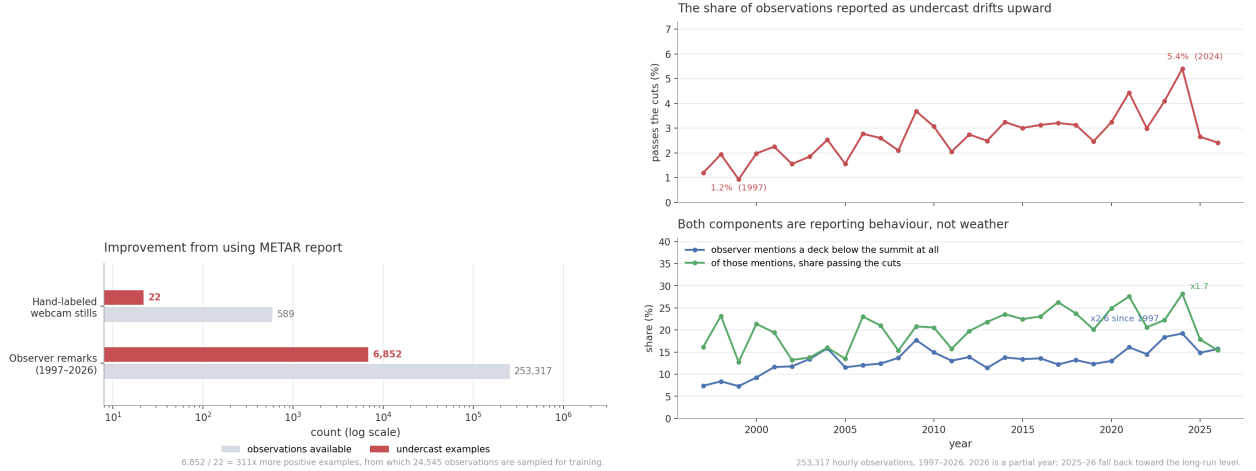
Table 3: Effect of removing each cut from the full set, on the webcam days.

Cut dropped	Extra false alarms	Extra undercasts caught	Precision
BKN or OVC coverage	+25	+4	0.38
Lowest layer overhead $\geq 1,000$ ft	+9	0	0.50
Deck ≥ 500 ft below summit	+3	0	0.62
Visibility > 40 SM	+1	0	0.67
No low lid overhead	+1	0	0.67



Figure 2: Noon webcam frames. Top: a day that passes every cut. Bottom: a hand-labelled undercast rejected only because the observer coded the deck SCT.

rise in how often observers mention a deck at all, so it is most likely a change in reporting practice, not in climate. To keep a model from learning the calendar, every undercast observation is matched with non-undercast observations from the same year, month and hour of day.



(a) Labelled data from each source (log scale). (b) Pass rate by year (top) and its components (bottom).

Figure 3: The remark-derived undercast record.

4 A physics-based baseline

The first attempt used hand-written rules on the NWP output (table 4), with thresholds fitted on the training split.

Table 4: Physics rules (top) and their best performance on the held-out year at a $\sim 3\%$ base rate (bottom). The random baseline fires as often as HRRR’s rule D.

A: inversion	$T_{850} > T_{925}$ by more than a fitted margin: a capping inversion in the layer just below the summit			
B: summit above deck	no modelled ceiling, or one above 1,917 m			
C	A and B			
D	C and modelled surface visibility above a threshold			

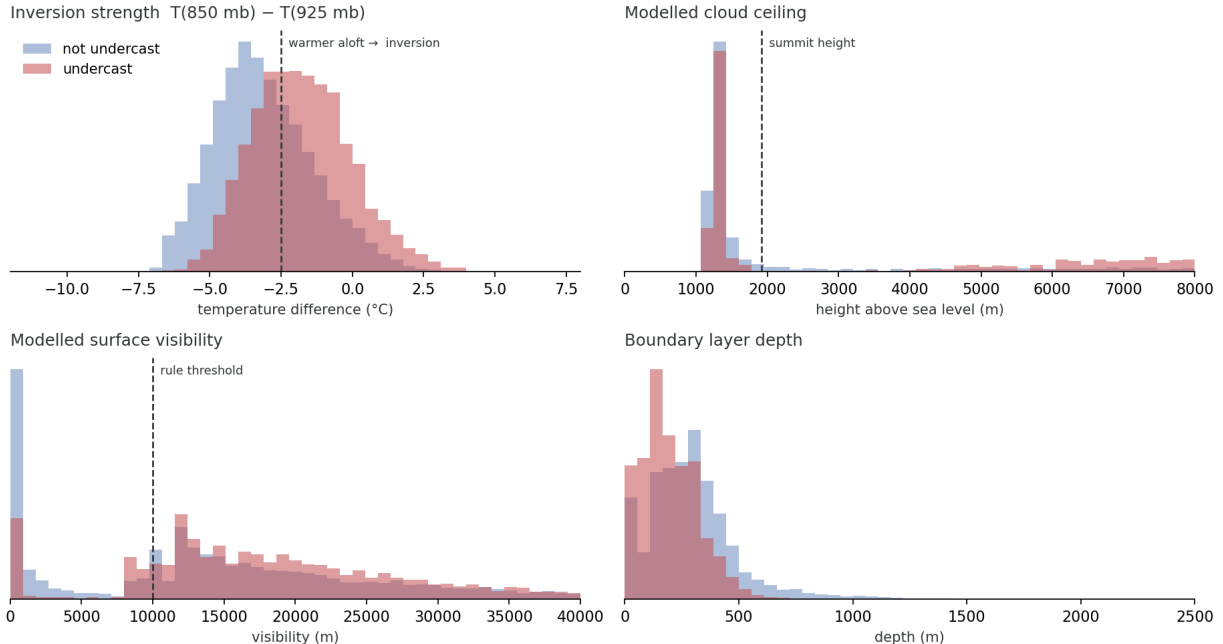
Source	Best rule	Precision	Recall	False alarms
Random	—	0.03	0.21	1,751
HRRR	D	0.09	0.60	1,651
NAM	D	0.07	0.53	1,879
RAP	D	0.07	0.48	1,743
GFS	C	0.07	0.44	1,604
ECMWF	A (no ceiling field)	0.06	0.68	1,975

The best rule is about three times better than chance, which still makes it useless in practice. Two rules that seemed obvious made things worse. Requiring a modelled cloud top below the summit *selects against* undercast: HRRR rows with a cloud top below 1,917 m have a 9.2% undercast rate, compared with 16.8% overall and 21.5% when HRRR reports no cloud. Adding a low-cloud-fraction term lowers the training F1 of every rule combination, from 0.41 to 0.09. Both failures come from terrain resolution. HRRR’s 3 km grid flattens the summit cell to roughly 1,200 m, so there is no model atmosphere between its ground and the observer. Fields on pressure surfaces are less sensitive to this.

Figure 4 compares the distributions directly for 24,315 HRRR near-analysis hours (3,597 undercast). The 850–925 mb temperature difference is the single most informative quantity in the

dataset. Undercast hours are about 1.5K more stable in that layer, yet 36% of them fall on the “no inversion” side of the best threshold. Every panel shows a real shift, and none shows a clean split. That is why this is a machine-learning problem.

HRRR fields on undercast hours against everything else



Near-analysis (~1 h lead), 2014–2026. Densities, so the two groups are comparable despite very different counts.

Figure 4: HRRR fields for undercast hours against all other hours: inversion strength, modelled ceiling, surface visibility and boundary-layer depth. The spike at zero visibility is the summit in cloud.

5 Machine-learning model

5.1 Inputs

About two dozen fields are taken from every NWP source that publishes them (full list in section C): cloud cover by layer, cloud base, top and ceiling, temperature and relative humidity on pressure levels, 2 m temperature and humidity, boundary-layer height, freezing level, visibility, wind, precipitation, and derived differences such as summit-minus-925 mb temperature and 925–850 mb humidity lapse. Missing values are filled with the constant `-9999` and get missingness indicators. Cloud-geometry columns also carry a `_no_cloud` flag, because a missing value there means the model reports clear sky, not that data is missing. A separate model is trained for each NWP source, and a *combined* model uses all sources’ columns together (213 features).

5.2 Algorithms

Three tree ensembles are trained for each source: Random Forest, Gradient Boosting and XGBoost. Undercast hours make up about 3% of the data, and the imbalance is handled with class weights (balanced weights for the forest and for gradient boosting, `scale_pos_weight` for XGBoost) instead of synthetic samples. Given how noisy the labels are, anything more elaborate would mostly fit labelling error.

5.3 Validation design

Neighbouring hours are not independent, so the splits respect time:

- Cross-validation folds are grouped by week (5 folds), so a multi-day inversion cannot straddle the training and test sets.
- Two holdouts are kept apart by date. Any training row that shares a date with a holdout, or falls within one day of it, is dropped (about 17% of training rows). The *held-out year* keeps the true base rate: 16,690 hours, 413 of them undercast (2.5%). The *webcam holdout* is the 589 hand-labelled days at all seven leads (4,122 forecast rows, 154 undercast).
- Negatives are subsampled 5:1 in training, which keeps every positive. Training-set precision is therefore inflated, so each decision threshold is fitted on the unsampled held-out year: one F1-optimal threshold per forecast lead.

The combined model trains on 31,009 rows (5,572 positive) spanning 104 weeks, 2023–2026. The ECMWF open-data archive starts in 2023, which limits the combined model to those years. Figure 5 shows why this design matters. A first version split rows at random, so the other ~ 23 hours of the same day sat in the training set. That gave precision above 0.9, which was memorisation, not forecast skill. With date-grouped splits, the same model, data and algorithms reach about 0.2.

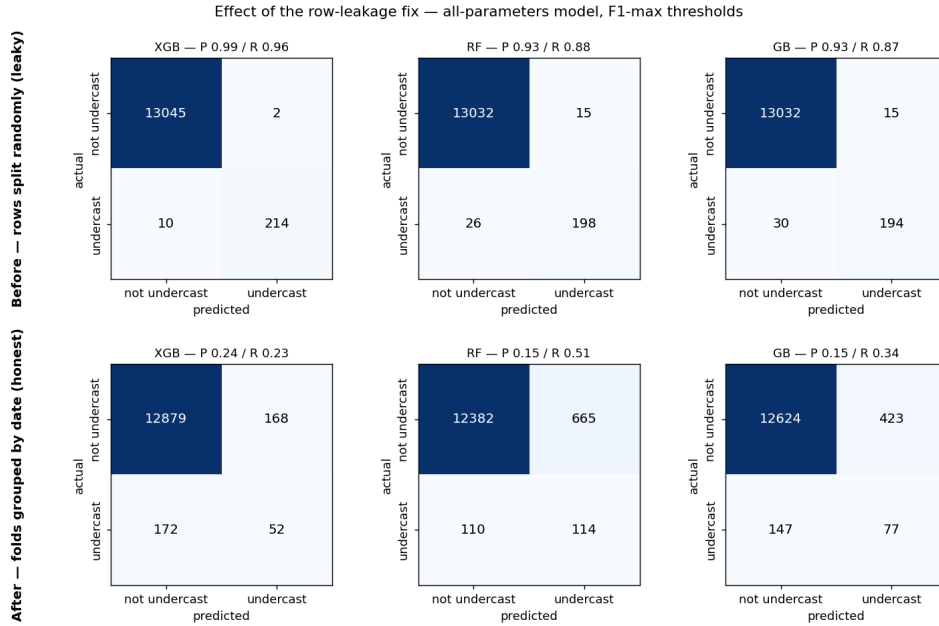


Figure 5: First-pass all-sources model, out-of-fold, with a row-wise split (top) and a date-grouped split (bottom). This is the only figure from the first pass and is not comparable with the other numbers here.

5.4 Results by source and lead

Each test row is a forecast issued 1, 24, 48, 72, 96, 120 or 144 h before its valid time. One model serves all leads, and each lead gets its own threshold. The combined model beats every individual source at every lead, and the margin grows with lead time: within 0.05 ROC-AUC of the best single source at 1 h, 0.09 ahead at 120 h (table 5 and fig. 6). ROC-AUC declines gradually with

lead, while precision drops quickly. Figure 7 summarises each source across algorithms, fig. 8 shows the curves of the combined model and fig. 9 its most important features. Feature importances are hard to interpret, because many columns are strongly correlated (section 5.6). Even so, ECMWF moisture structure (925–850 mb RH difference) ranks first under every algorithm.

Table 5: ROC-AUC on the held-out year by forecast lead, Gradient Boosting. A dash means the source does not reach that lead.

Source	1 h	24 h	48 h	72 h	96 h	120 h	144 h
Combined	0.941	0.922	0.880	0.857	0.824	0.775	0.764
RAP	0.890	0.859	0.804	—	—	—	—
HRRR	0.889	0.870	0.837	—	—	—	—
ECMWF	0.879	0.872	0.842	0.847	0.804	0.739	0.754
NAM	0.861	0.830	0.818	—	—	—	—
NBM	0.844	0.769	0.762	0.734	0.690	0.675	0.651
GFS	0.842	0.818	0.790	0.747	0.745	0.682	—

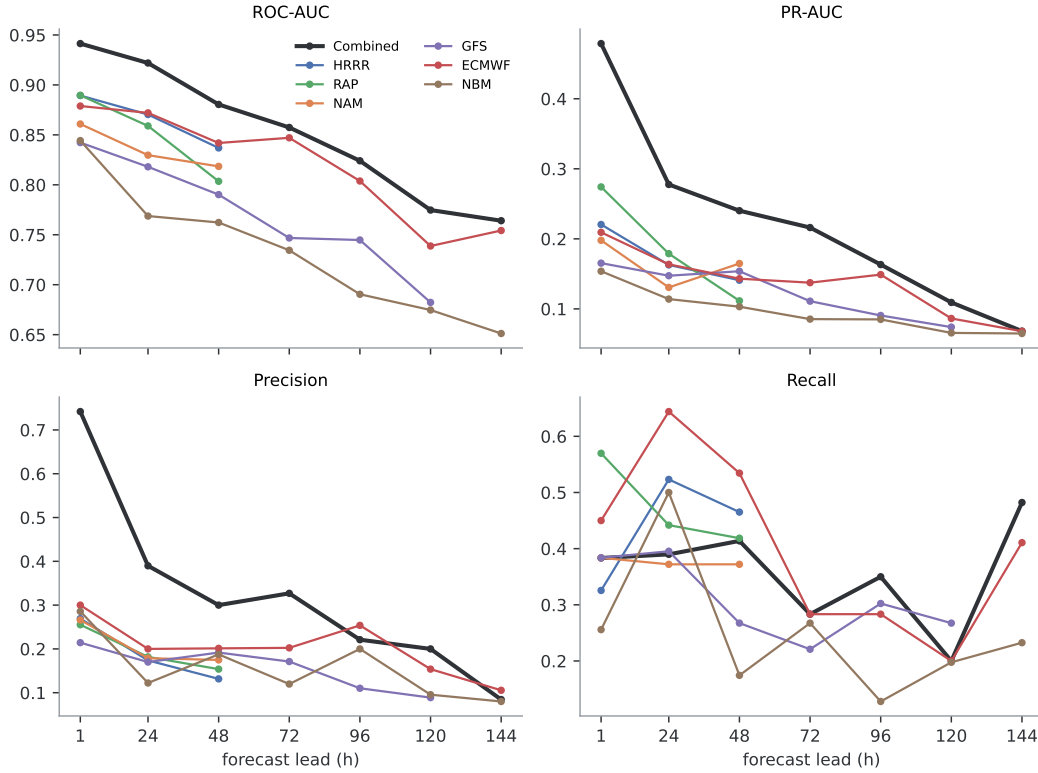


Figure 6: Skill against forecast lead on the held-out year, Gradient Boosting, each lead at its own threshold. The combined model is the heavy line.

5.5 Choice of algorithm

An ensemble only helps when its members make different mistakes. On one NWP source, the three algorithms’ predictions have rank correlations of 0.86–0.95. The six NWP sources under one algorithm correlate at only 0.34–0.58 (fig. 10). Accordingly, voting across algorithms gains nothing

Per-source undercast skill on the held-out year, by algorithm

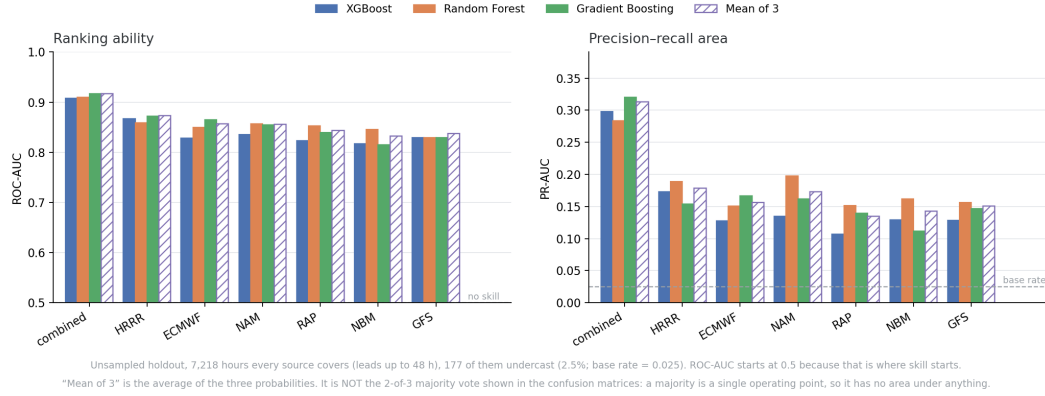


Figure 7: ROC-AUC and PR-AUC for each source on the held-out year. “Mean of 3” averages the three algorithms’ probabilities.

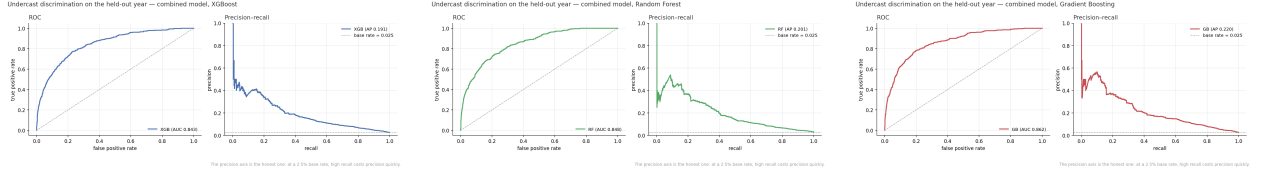


Figure 8: ROC and precision–recall curves of the combined model on the held-out year: XGBoost, Random Forest, Gradient Boosting (left to right). The three finish within 0.03 ROC-AUC of each other.

(table 6), and neither do four alternative algorithms (table 7). The disagreement worth exploiting is between NWP sources: the combined model reaches ROC-AUC 0.908 [0.860, 0.951], +0.013 [+0.003, +0.026] above the average of the six single-source models (95% intervals from a bootstrap over days; fig. 10, right). The production model is therefore a single combined-source Gradient Boosting model with per-lead thresholds.

Table 6: Combined model, each algorithm at its own F1-optimal threshold.

Predictor	Held-out year			Webcam days		
	P	R	F1	P	R	F1
XGBoost	0.26	0.29	0.27	0.35	0.18	0.24
Random Forest	0.28	0.31	0.30	0.25	0.06	0.10
Gradient Boosting	0.28	0.32	0.30	0.41	0.25	0.31
Any of the three	0.22	0.39	0.28	0.36	0.28	0.32
2 of 3 (vote)	0.31	0.30	0.30	0.38	0.17	0.23
Unanimous	0.37	0.24	0.29	0.28	0.05	0.09
Mean probability	0.28	0.31	0.29	0.31	0.14	0.19

5.6 Overfitting checks

The combined model has 213 feature columns, and its training rows are far from independent. Each undercast hour appears once per lead, and consecutive hours belong to the same event (table 8). There are more feature columns than distinct undercast days.

The features are just as redundant. The 213 columns represent 78 physical quantities plus four

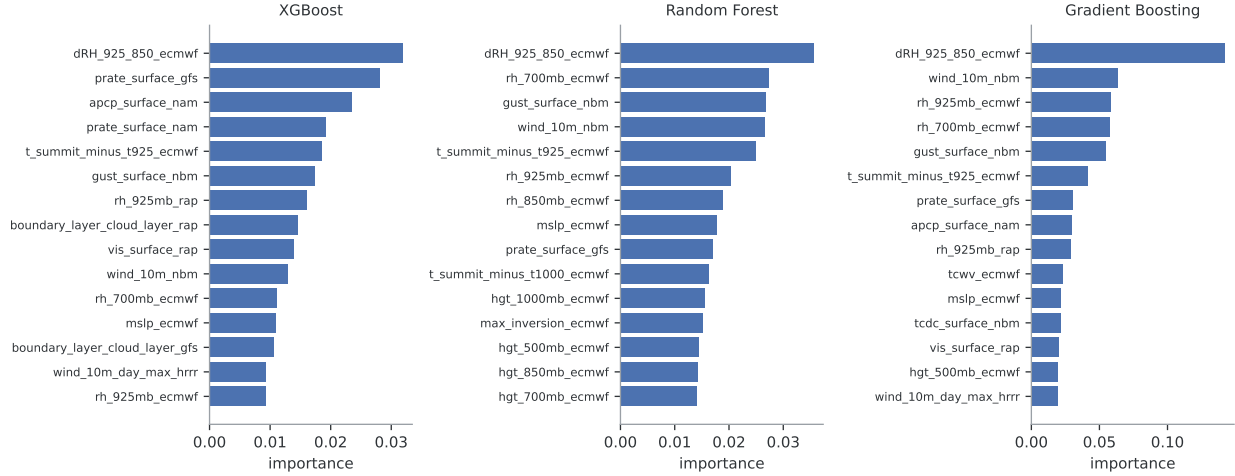


Figure 9: Top 15 feature importances of the combined model for each algorithm. Per-source importances are in section B.

Table 7: ROC-AUC on the held-out year for other algorithms. Histogram gradient boosting gets the raw columns, since it splits on missing values natively. Logistic regression gets median imputation and standardisation.

Algorithm	Combined	HRRR
Gradient Boosting	0.862	0.866
Random Forest, deeper trees	0.849	0.855
Random Forest	0.848	0.862
XGBoost	0.843	0.863
Logistic regression	0.819	0.861
Histogram gradient boosting	0.819	0.854
Extra Trees	0.799	0.850

calendar encodings (82 in total). Of the physical quantities, 41 come from more than one source and 37 from only one. The median $|\text{Spearman } r|$ between the same quantity from different sources is 0.872, against 0.181 for random pairs of different quantities (fig. 11). A principal component analysis of the 11,620 training rows where every source reports, over the 98 columns populated more than half the time, needs 21 components for 90% of the variance and 29 for 95%. The participation ratio is 8.7.

Directly reducing model capacity does not help (fig. 12). Keeping only the top k features matches the full model at $k \geq 70$ and loses up to 0.05 ROC-AUC below that. A maximum tree depth of 3 is as good as any other within the ± 0.005 reseed noise.

5.7 Production model

The production forecast is a Gradient Boosting model with maximum depth 3, trained on all 213 combined features, with a separate threshold for each lead. Table 9 gives the held-out-year results that the website’s Model performance panel shows, and fig. 13 gives the confusion matrices. At 1 h, three of every four undercast calls are right. At 24 h it is about two in five, and at 144 h about one in twelve.

Ensembling helps across weather models, not across algorithms

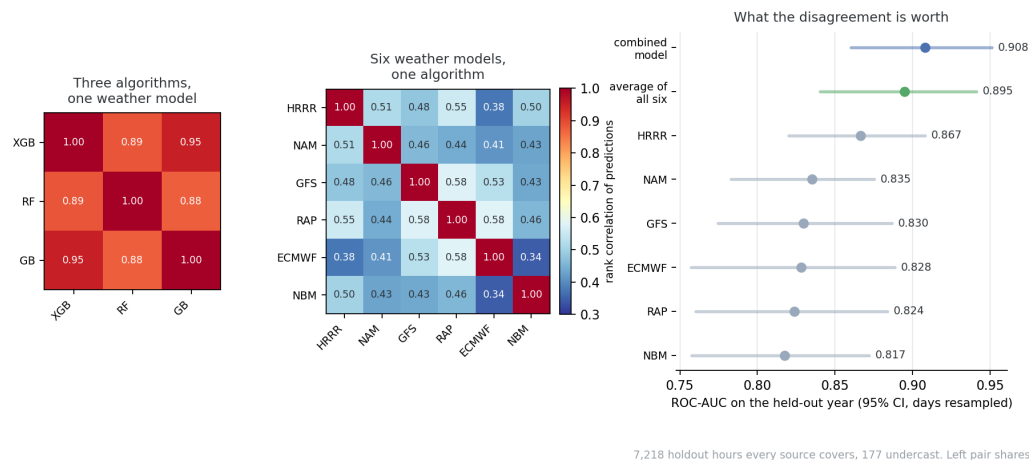


Figure 10: Rank correlation between algorithms (left) and between NWP sources (middle), on a shared colour scale; ROC-AUC with bootstrap 95% intervals (right).

Table 8: Effective size of the combined model's training set.

Counting unit	<i>n</i>	Per feature
Positive rows	5,572	26.2
Distinct positive valid hours (leads de-duplicated)	797	3.7
Distinct undercast episodes (gaps ≤ 2 h bridged)	182	0.85
Distinct undercast days	175	0.82

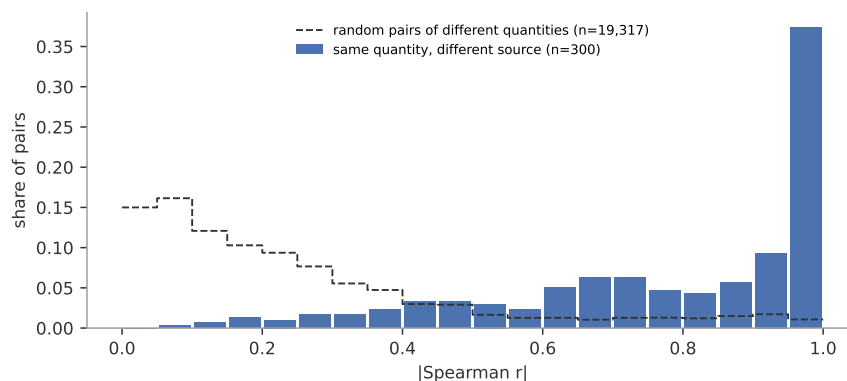


Figure 11: |Spearman r | for all 300 same-quantity, different-source pairs (bars) against 19,317 random pairs of different quantities (dashed), each as a share of its own total.

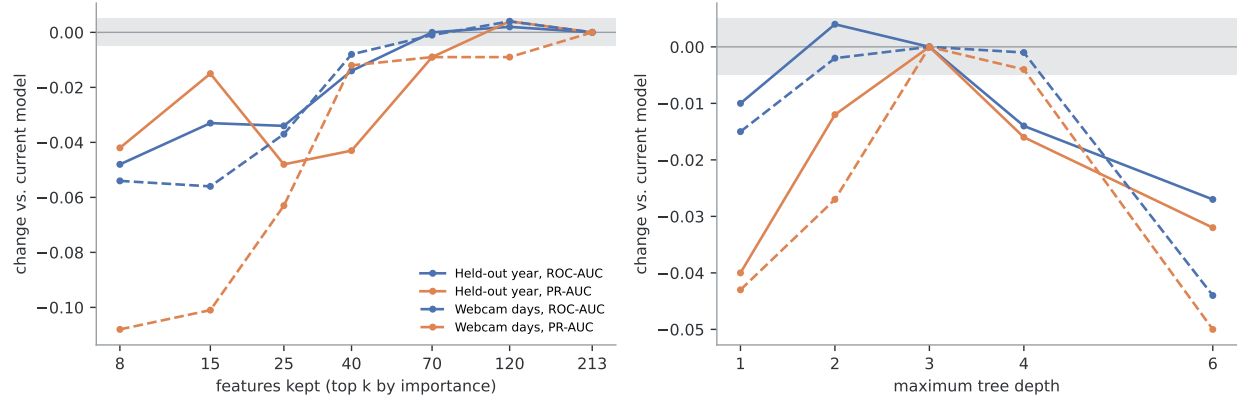
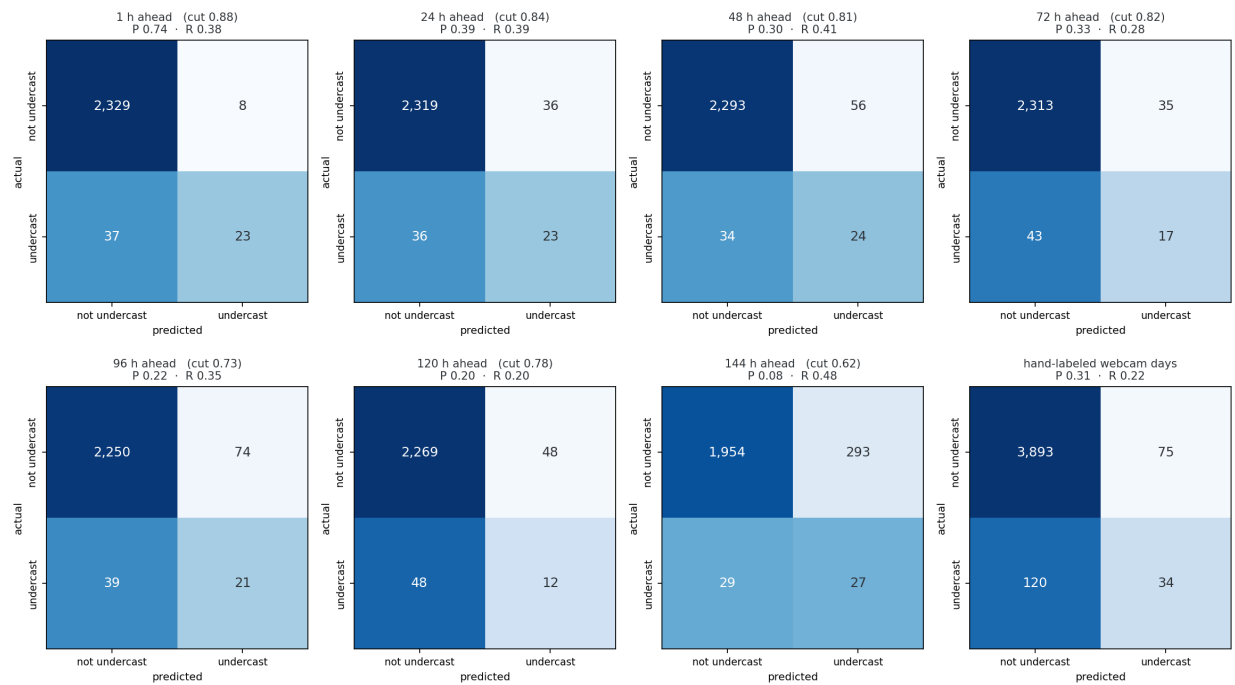


Figure 12: Change in ROC-AUC and PR-AUC against the production configuration when keeping only the top k features (left) or changing the maximum tree depth (right), on both holdouts. The grey band is the ± 0.005 spread over five seeds.

Table 9: Production model on the held-out year (true base rate), by forecast lead.

Lead (h)	Threshold	Hours	Undercast	TP	FP	FN	Precision	Recall	ROC-AUC
1	0.880	2,397	60	23	8	37	0.74	0.38	0.941
24	0.835	2,414	59	23	36	36	0.39	0.39	0.922
48	0.805	2,407	58	24	56	34	0.30	0.41	0.880
72	0.815	2,408	60	17	35	43	0.33	0.28	0.857
96	0.735	2,384	60	21	74	39	0.22	0.35	0.824
120	0.780	2,377	60	12	48	48	0.20	0.20	0.775
144	0.620	2,303	56	27	293	29	0.08	0.48	0.764

The forecast the site publishes — combined model, Gradient Boosting



All but the last panel: the held-out year at its true 2.5% base rate, at the per-lead thresholds the live page uses. Last panel: every hand-labeled day, scored against human webcam labels.

Figure 13: Production model confusion matrices at each lead on the held-out year, and on the webcam days.

6 Limitations and next steps

- **Labels set the ceiling.** The remark cuts are about 70% precise, so roughly three in ten training positives are not real undercasts. No classifier can be much cleaner than its target. The webcam set behind that estimate has only 22 positive days.
- **Reporting drift.** The pass rate quadruples over 1997–2026, most likely because of changing observer practice. Matched sampling removes the calendar signal, but the label definition is not stationary.
- **Vertical resolution.** The pressure levels used are 1000, 925, 850, 700 and 500 mb. The 1,564 m gap between 850 and 700 mb straddles the 1,917 m summit, so the top of the inversion can only be interpolated. Fetching 800 and 750 mb fields would bracket it.
- **Observed profiles.** The MWOBS mesonet includes an Auto Road Vertical Profile: six stations from 1,600 to 5,300 ft plus the summit tower, with 5-minute temperature, humidity, wind and solar radiation [6]. It would measure the inversion directly. Synoptic Data holds it only from 18 February 2026, so earlier history must be requested from MWOBS.
- **Not yet explored.** The effect of the 5:1 negative subsampling ratio, and which physical factors drive the per-source differences.

Data, code and AI use

All scripts are in the repository <https://github.com/lucasehinger/lucasehinger.github.io> (`scripts/`). Model metadata and metrics are in `files/weather/models/obs/`. Figures in this document come from the website or from `scripts/build_weather_report_figures.py`. For data access, contact ehingerl@mit.edu. NWP data is from NOAA and ECMWF, accessed with the Herbie Python package. Webcam imagery and METAR remarks are from the Mount Washington Observatory. Most of the code was written with AI assistance. The author made the modelling decisions and reviewed every analysis step, and wrote most of the website text.

References

- [1] Mount Washington Observatory, “Undercast,” <https://mountwashington.org/undercast/>.
- [2] Green Mountain Club, “Temperature inversion & undercast,” <https://www.greenmountainclub.org/temperature-inversion-undercast-obsession/>.
- [3] *Atmospheric Research*, <https://www.sciencedirect.com/science/article/abs/pii/S016980951930941X>.
- [4] *Boundary-Layer Meteorology*, <https://doi.org/10.1007/s10546-009-9407-1>.
- [5] EatSleepWild, “What is a cloud inversion?” <https://eatsleepwild.com/what-is-a-cloud-inversion/>.
- [6] Fowler *et al.*, *J. Atmos. Oceanic Technol.* **40**, 439 (2023): the Mount Washington Observatory regional mesonet.

A Website controls and links

The summit selector on the forecast page loads quick links for the selected summit from `files/weather/links/<summit>.json`. For Mount Washington these are the NWS point forecast, the NWS hourly graph, the NWS Gray, ME area forecast discussion, Windy, the MWOBS Higher Summits Forecast and the Mount Washington Avalanche Center. The other summits get the same six links, with the NWS and Windy links pointed at that summit. The details page also has interactive pickers: an algorithm and source picker for feature importances (section B), a metric picker for skill against lead (fig. 6), and a lead and source picker for confusion matrices (table 10). The historical-conditions page has a date, model and unit picker for the day-by-day NWP viewer.

B Per-source and per-algorithm results

Table 10: Single-source and combined models, Gradient Boosting. Out-of-fold (OOF) AUCs are from week-grouped 5-fold cross-validation on the 5:1 subsampled training set. Held-out precision and recall use the single all-lead threshold.

Source	Training			OOF		Held-out year		
	Years	Rows	Features	ROC	PR	Thr.	P	R
Combined	2023–26	31,009	213	0.781	0.485	0.78	0.28	0.32
HRRR	2014–26	34,829	46	0.855	0.549	0.67	0.16	0.48
RAP	2021–26	16,707	39	0.800	0.447	0.70	0.19	0.43
NAM	2020–26	15,607	38	0.820	0.475	0.73	0.19	0.39
GFS	2021–26	33,990	39	0.736	0.372	0.65	0.13	0.35
ECMWF	2023–26	31,013	40	0.722	0.374	0.66	0.16	0.49
NBM	2020–26	50,097	31	0.687	0.320	0.70	0.13	0.20

Table 11: Held-out-year precision/recall by lead for each source, Gradient Boosting, each lead at its own threshold.

Source	1 h	24 h	48 h	72 h	96 h	120 h	144 h
Combined	.74/.38	.39/.39	.30/.41	.33/.28	.22/.35	.20/.20	.08/.48
HRRR	.27/.33	.17/.52	.13/.47	—	—	—	—
RAP	.26/.57	.18/.44	.15/.42	—	—	—	—
NAM	.27/.38	.18/.37	.17/.37	—	—	—	—
GFS	.21/.38	.17/.40	.19/.27	.17/.22	.11/.30	.09/.27	—
ECMWF	.30/.45	.20/.64	.20/.53	.20/.28	.25/.28	.15/.20	.11/.41
NBM	.29/.26	.12/.50	.19/.17	.12/.27	.20/.13	.10/.20	.08/.23

Table 12: Combined model by algorithm and lead on the held-out year: ROC-AUC, precision / recall.

Algorithm	1 h	24 h	48 h	72 h	96 h	120 h	144 h
XGBoost	0.940 .60/.50	0.907 .32/.49	0.871 .24/.52	0.836 .21/.35	0.789 .20/.18	0.753 .10/.32	0.724 .09/.21
Random Forest	0.938 .54/.57	0.911 .33/.51	0.876 .28/.47	0.839 .30/.25	0.797 .22/.35	0.774 .24/.17	0.726 .07/.39
Gradient Boosting	0.941 .74/.38	0.922 .39/.39	0.880 .30/.41	0.857 .33/.28	0.824 .22/.35	0.775 .20/.20	0.764 .08/.48

Table 13: Production model on the 589 webcam days, by lead, using the held-out-year thresholds. With 22 positives, these figures are noisy.

	1 h	24 h	48 h	72 h	96 h	120 h	144 h
TP / FP / FN	8/11/14	9/11/13	10/16/12	4/7/18	1/6/21	6/5/16	1/0/21
Precision	0.42	0.45	0.38	0.36	0.14	0.55	1.00
Recall	0.36	0.41	0.45	0.18	0.05	0.27	0.05

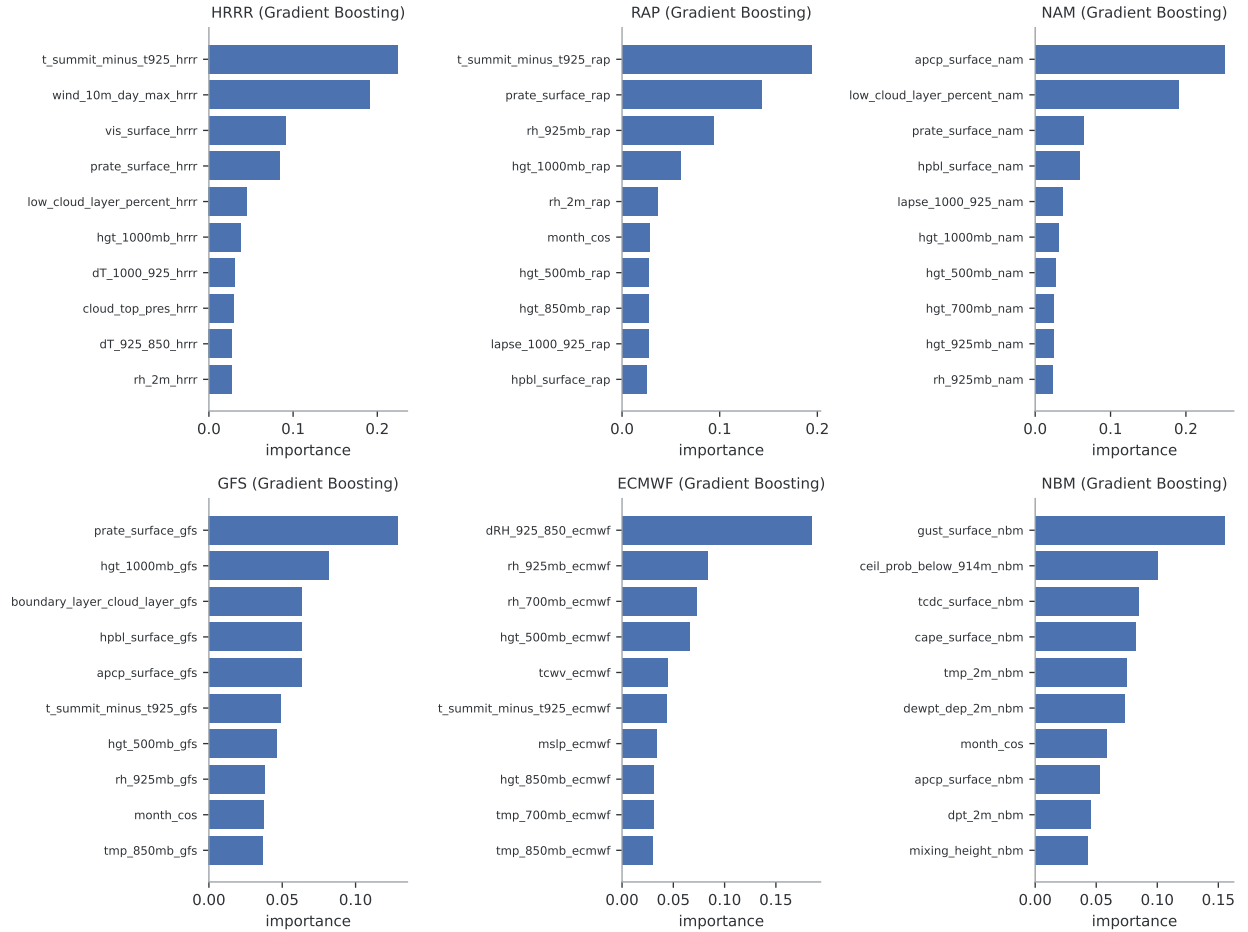


Figure 14: Top 10 feature importances of each single-source Gradient Boosting model.

C Input parameters

Each field is pulled from every NWP source that publishes it. Derived quantities are added: layer temperature differences and lapse rates (`dT_1000_925`, `dT_925_850`, `dT_850_700`, `dT_1000_850`, `lapse_*`), `dRH_925_850`, the strongest inversion in the column (`max_inversion`), summit-minus-level temperatures (`t_summit_minus_t925`, `t_summit_minus_t1000`), and calendar encodings (month and hour as sine/cosine).

Key	Description
<code>cloud_top</code> , <code>cloud_top_hgt</code> , <code>cloud_top_pres</code>	Cloud-top height (AGL), geopotential height, pressure
<code>cloud_base_m</code> , <code>cloud_base_pres</code>	Cloud-base height (AGL) and pressure
<code>cloud_ceiling_m</code>	Height of the lowest BKN/OVC layer (AGL)
<code>low/middle/high_cloud_layer_percent</code>	Cloud cover below ~ 2 km, 2–6 km, above 6 km
<code>tcdc_surface</code> , <code>tcdc_high_cloud</code> , <code>cdcb_high_cloud</code>	Total and high cloud cover
<code>boundary_layer_cloud_layer</code>	Cloud fraction within the planetary boundary layer
<code>tmp_{1000,925,850,700,500}mb</code>	Temperature on pressure levels (~ 100 , 750, 1500, 3000, 5500 m)
<code>rh_{1000,925,850,700}mb</code>	Relative humidity on pressure levels
<code>vvel_{925,850,700}mb</code>	Vertical velocity on pressure levels
<code>hgt_{1000,925,850,700,500}mb</code>	Geopotential height of pressure levels
<code>tmp_2m</code> , <code>rh_2m</code> , <code>dpt_2m</code> , <code>dewpt_dep_2m</code>	Screen-level temperature, humidity, dewpoint and dewpoint depression
<code>t_summit</code>	Temperature interpolated to the 1,917 m summit
<code>hpbl_surface</code> , <code>mixing_height</code>	Planetary boundary-layer and mixing height
<code>cape</code> , <code>cape_surface</code>	Convective available potential energy
<code>hgt_0C_iso</code>	Freezing level
<code>vis_surface</code>	Surface visibility
<code>ceil_prob_below_*</code> , <code>vis_prob_below_*</code>	NBM probability of ceiling (152–2,012 m) or visibility (1–5 mi) below a threshold
<code>wind_10m</code> , <code>wind_10m_day_max</code> , <code>gust_surface</code>	10 m wind, daily maximum, gust
<code>prate_surface</code> , <code>apcp_surface</code>	Precipitation rate and accumulation
<code>mslp</code> , <code>sp_surface</code> , <code>tcwv</code>	Sea-level and surface pressure, column water vapour