This document is the supplementary material to the manuscript Synergistic estimation of atmospheric boundary layer height from multi-source observations over the Arctic Ocean during MOSAiC. It contains Tables S1–S4, Figs. S1–S4 and Sects. S1–S5, which provide supporting methods, evaluation results and sensitivity analyses for the main text.

1 Input features and training details

S1.1 Input blocks

Each sample combines four observational blocks, one auxiliary block and two time encodings. All continuous variables are standardised with a RobustScaler fitted on the training set (median and interquartile range).

The following table summarises the observation blocks and auxiliary inputs used in the reference configuration described in Sect. 3.1 of the main text, together with their dimensions and data sources.

Input blocks, their dimensions and sources.

Block	Dimension	Source and description
HATPRO brightness temperature	14	7 K-band channels (22.24–31.4 GHz) and 7 V-band channels (51.26–58 GHz)
MiRAC-P brightness temperature	8	6 channels in the 183.31 GHz absorption band and 2 window channels at 243 and 340 GHz
AERI radiance	50	Infrared channels selected by mRMR (Table S1)
Backscatter profile	200	CL31 ceilometer, 10 m vertical spacing, 10–2000 m
Wind profile	2 × 100	INTERPSONDE u and v components, 20 m spacing, 10–1990 m
CBH	2	Cloud-base height (zero when cloud-free) and a validity flag
Time encoding	3	Sine and cosine of month, sine of hour

S1.2 Derived physical features (34 dimensions)

(1) Layer-wise wind statistics (8 dimensions)

Let the wind vector at level k be (u_k, v_k) with $K = 100$ levels, divided into thirds: lower $k < 0.35K$, middle $0.35K \le k < 0.75K$, upper $k \ge 0.75K$.

The following features summarise wind speed and shear over different parts of the profile, providing the derived wind descriptors introduced in Sect. 3.1 of the main text.

Definition of the eight layer-wise wind statistics.

Feature	Definition
Layer-mean wind speed	Mean of $\sqrt{(u_k^2 + v_k^2)}$ within each layer (3 dimensions)
Layer-mean wind-speed shear	Mean of $\lvert\Delta V\rvert$ within each layer, $\Delta V_k = V_k - V_{k-1}$ (3 dimensions)
Mean vector shear	Mean of $\sqrt{(\Delta u_k^2 + \Delta v_k^2)}$ over the whole profile (1 dimension)

Feature	Definition
Mean wind-direction shear	Mean of $\lvert\Delta\theta_k\rvert$, with the direction unwrapped before differencing (1 dimension)

(2) Backscatter structure (7 dimensions)

Let Bk (k = 1…200) be the log-compressed backscatter profile after percentile clipping of adjacent-level differences and cumulative reconstruction, as described in Sect. 3.1. The gradient and curvature features use central differences in level-index coordinates: Gk = (Bk+1 − Bk−1)/2 and Ck = (Gk+1 − Gk−1)/2 at interior levels, with one-sided differences at the endpoints. These are profile-structure features per level-index step, not derivatives per metre.

The following features describe the vertical structure and characteristic transitions of the processed backscatter profile. They supplement the preprocessing and feature definitions in Sect. 3.1 of the main text.

Definition of the seven backscatter structure features.

Feature	Definition
Low-level mean backscatter	Mean backscatter over the lowest 40 levels (about 10–400 m)
Low-level gradient energy	Mean of $\lvert G_k\rvert$ over the lowest 40 levels
Profile gradient energy	Mean of $\lvert G_k\rvert$ over all 200 levels
Curvature energy	Mean of $\lvert C_k\rvert$ over all 200 levels
Backscatter peak height	Level index of $\mathrm{argmax}_k\, B_k$
Maximum-gradient height	Level index of $\mathrm{argmax}_k \lvert G_k\rvert$, i.e. the height of the sharpest backscatter transition, which marks interfaces near inversions and cloud base
Top–bottom contrast	Difference between the mean backscatter of the top 40 and the bottom 40 levels

(3) Microwave brightness-temperature structure (9 dimensions)

The following features summarise brightness-temperature levels and contrasts within and between the microwave bands, complementing the spectral inputs described in Sect. 3.1 of the main text.

Definition of the nine microwave brightness-temperature features.

Feature	Definition
HATPRO K-band mean	Mean brightness temperature of the first 7 channels
HATPRO V-band mean	Mean brightness temperature of the last 7 channels
K-band spectral slope	Difference between the first and last channel of the band
V-band spectral slope	Difference between the first and last channel of the band
K–V contrast	K-band mean minus V-band mean
MiRAC-P 183 GHz mean	Mean of the 6 absorption-band channels
MiRAC-P window mean	Mean of the 243 and 340 GHz channels
183 GHz spectral slope	Difference between the first and last absorption-band channel
Window channel difference	340 GHz minus 243 GHz

(4) AERI spectral statistics (5 dimensions)

Mean, standard deviation, maximum, minimum and range of the radiances across the 50 channels.

(5) Auxiliary and time (5 dimensions)

Cloud-base height, cloud-base validity flag, sine and cosine of month, and sine of hour.

S1.3 AERI channel selection

Table S1 identifies the 50 AERI channels retained by mRMR, specifying the infrared inputs described in Sects. 2.3 and 3.1 of the main text.

Table S1. AERI channels selected by mRMR

Channels are listed in decreasing order of importance.

No.	Wavenumber (cm^{-1})	No.	Wavenumber (cm^{-1})
1	547.23	26	539.04
2	1339.88	27	576.16
3	1255.99	28	1255.02
4	680.79	29	639.80
5	1291.67	30	1255.51
6	550.12	31	1337.95
7	1250.68	32	546.75
8	1297.4	33	558.32
9	553.49	34	1260.81
10	617.63	35	1281.54
11	1327.35	36	650.41
12	1253.10	37	1257.04
13	540.96	38	545.30
14	1289.74	39	1250.20
15	617.14	40	1340.36
16	1279.13	41	539.52
17	1342.78	42	1304.69
18	540.48	43	1262.74
19	683.68	44	675.97
20	1252.13	45	569.89
21	563.14	46	561.21
22	1294.08	47	1338.92
23	674.52	48	1254.06
24	1257.92	49	1268.04
25	1299.86	50	658.61

S1.4 Training settings

The following table documents the optimisation settings and loss-function parameters used for model training, supporting the training procedure described in Sect. 3.4 of the main text.

Training settings shared by all configurations.

Item	Setting
Optimiser	AdamW ($\beta = 0.9/0.999$)
Learning-rate schedule	OneCycleLR: linear warm-up over the first 10 % of steps, then cosine decay to 10^{-4} of the initial value
Mixed precision	Enabled (CUDA AMP)
Batch size	48
Maximum epochs	320

Item	Setting
Initial learning rate	2×10^{-4}
Weight decay	6×10^{-4}
Early-stopping patience	40 epochs
Data split	Random split, 70 % training and 30 % validation
Loss function	Asymmetric Huber + MSE + relative error + condition balancing + bias/spread constraints
Huber δ	1.0
MSE weight	0.20
Relative-error weight	0.12
Bias weight	0.18
Asymmetry coefficient	0.65 (underestimation penalised more strongly)
Condition-balance weight	0.30
Low-layer threshold / weight	250 m / 0.42
High-layer threshold / weight	500 m / 0.24
Weight for cloudy high-layer underestimation	0.16
Early-stopping criterion	Validation loss; select the checkpoint with the lowest validation loss and stop after 40 consecutive epochs without a decrease

S2 Supplementary evaluation results

S2.1 Ablation and sampling diagnostics

Figure S1 provides the absolute error metrics underlying the source-removal comparisons in Sect. 4.1 and Fig. 7 of the main text. It allows changes in RMSE, MAE and bias to be examined separately for each sky condition.

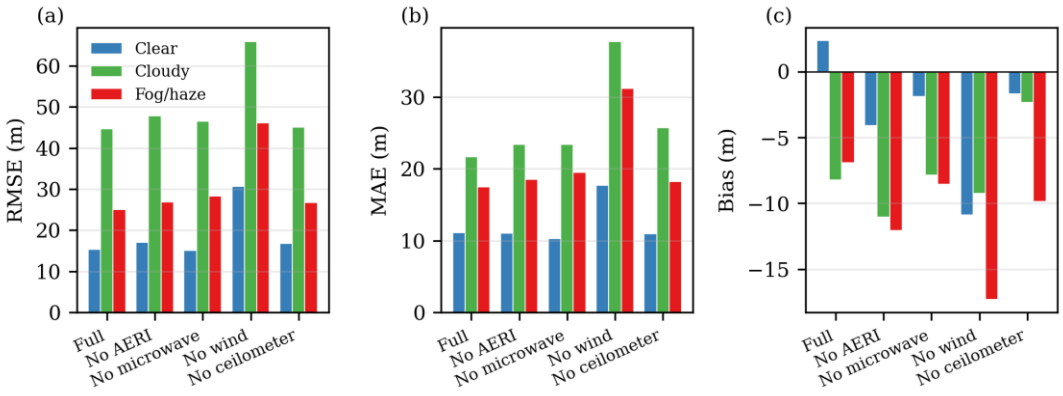


Fig. S1. Absolute RMSE, MAE and bias for the full-input model and the four single-source removal experiments, shown in panels (a), (b) and (c), respectively. All configurations use the same 834 launch-time records: 244 clear-sky, 393 cloudy and 197 fog/haze records. Cloudy combines ceilometer detection statuses 1–3. Bias is prediction minus reference. These are the same configuration-specific metrics used to calculate the differences in Figure 7 of the main text and the MOSAiC ablation rows in Fig. S4.

Figure S2 shows how sample availability and uncertainty in the error estimates vary across months, sky conditions, seasons and reference-height ranges. It supports the sampling limitations and depth-dependent performance discussed in Sect. 4.1 of the main text.

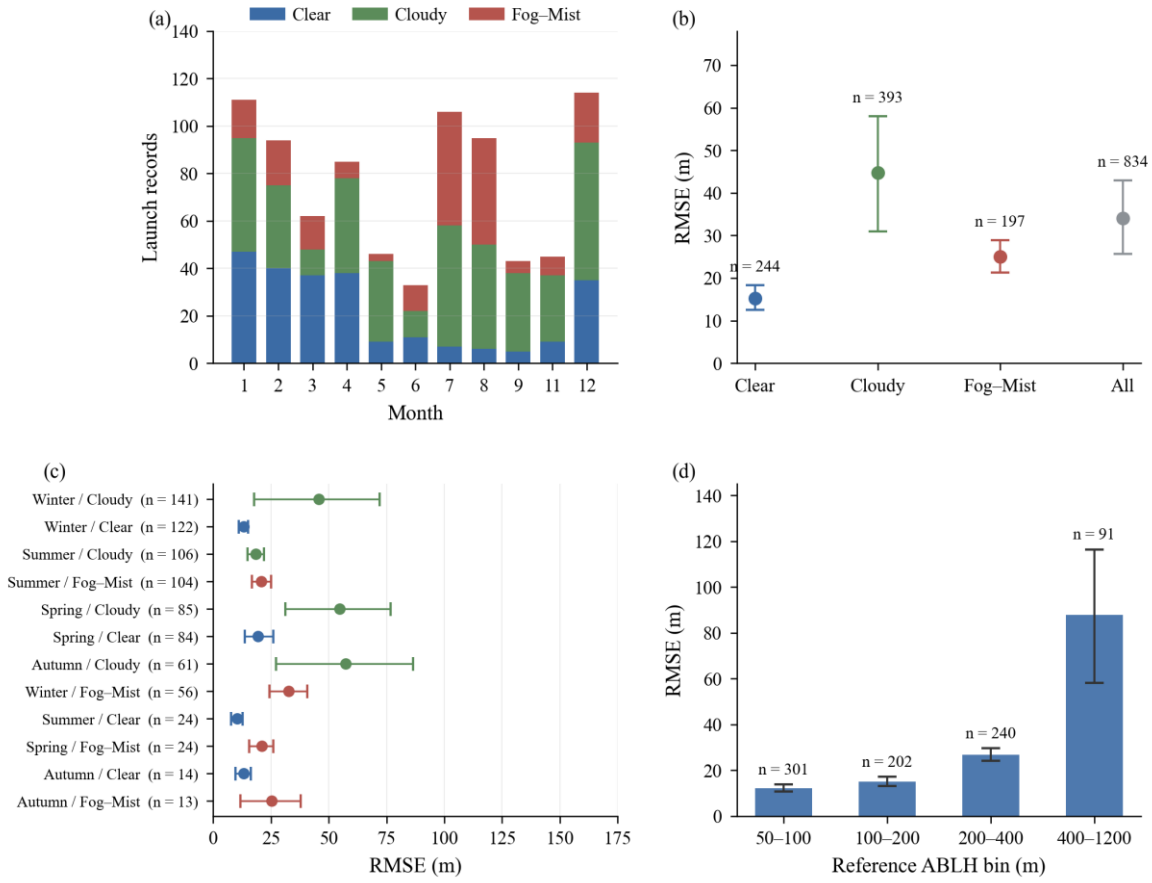


Fig. S2. Sample size and uncertainty of the evaluation on the 834 MOSAiC launch-time records. (a) Number of launch records per month, separated by sky condition. These launch-time counts differ from the counts of 10-min estimates used for the monthly distributions in Fig. 9 of the main text. (b) RMSE with 95 % bootstrap confidence intervals for each sky condition and for all records. (c) RMSE with 95 % bootstrap confidence intervals for each meteorological season and sky condition, sorted by sample size. Seasons are defined by calendar month: winter, December–February; spring, March–May; summer, June–August; and autumn, September–November, pooling available records across years. This grouping is used to examine sample availability and uncertainty in the evaluation; it differs from the five energy-budget-based periods used for the diurnal analysis in Fig. 10 of the main text. (d) RMSE with 95 % bootstrap confidence intervals as a function of the reference-ABLH bin. Confidence intervals are obtained from 2,000 percentile-bootstrap resamples.

S2.2 Depth-stratified comparison

Figure S3 visualises the depth-stratified RMSE comparison reported in Table S2, extending the diagnostic-reference comparison in Sect. 3.2 of the main text. The comparison concerns different

estimates derived from the same sounding system rather than accuracy against an independently observed ABLH.

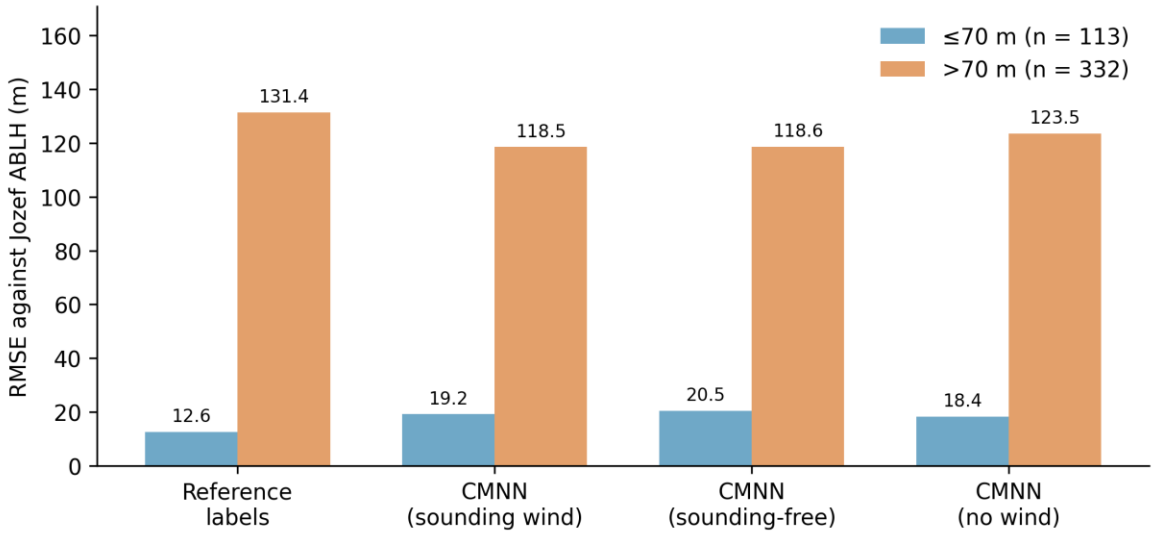


Fig. S3. RMSE (m) of the reference labels and three CMNN configurations against the PANGAEA ABLH diagnostic product for the same 445 common launches used in Table S2. Depth classes are defined by the study Ri-based reference ABLH: ≤70 m (n = 113) and >70 m (n = 332). The sounding-free configuration uses tower and Doppler-lidar winds.

Table S2 provides the sample sizes, RMSE, MAE and bias for the depth-stratified comparison, complementing Fig. S3 and the limitations discussed in the main-text conclusions. It shows that lower overall RMSE against the PANGAEA product does not imply improved agreement in the shallowest class.

On the 445 launches for which the PANGAEA diagnosis, study reference labels and all three model predictions are available, the estimates are compared with the ABLH of the MOSAiC lower-atmospheric-properties product (PANGAEA.957760). Depth classes are defined by the study Ri-based reference ABLH. Bias is the listed estimate minus the PANGAEA diagnosis, so a positive value means that the listed estimate is higher. This subset differs from the 560-launch six-diagnosis comparison in Figure 3 of the main text. The PANGAEA product applies a different diagnostic algorithm to the same soundings, providing a comparison across methods.

Table S2. Depth-stratified comparison with an alternative diagnostic product

Range	Estimate	n	RMSE (m)	MAE (m)	bias (m)
All	Reference labels	445	113.6	57.3	+48.6
All	CMNN (sounding wind)	445	102.8	56.6	+44.6
All	CMNN (sounding-free)	445	103.0	57.8	+44.2

Range	Estimate	n	RMSE (m)	MAE (m)	bias (m)
All	CMNN (no wind)	445	107.1	57.6	+40.0
≤70 m	Reference labels	113	12.6	9.5	+6.2
≤70 m	CMNN (sounding wind)	113	19.2	17.7	+15.7
≤70 m	CMNN (sounding-free)	113	20.5	17.7	+15.5
≤70 m	CMNN (no wind)	113	18.4	13.7	+10.9
>70 m	Reference labels	332	131.4	73.6	+63.0
>70 m	CMNN (sounding wind)	332	118.5	69.9	+54.4
>70 m	CMNN (sounding-free)	332	118.6	71.5	+53.9
>70 m	CMNN (no wind)	332	123.5	72.5	+49.9

Note: the 445-launch sample is the intersection of the 560 common launches in the six-diagnosis comparison and the 661 launch times with paired sounding-wind and Doppler-lidar predictions. All rows use this same subset; shallow and deeper classes contain 113 and 332 records, respectively.

Note: for reference ABLH of 70 m or less, the reference labels are closer to the PANGAEA diagnosis than the three model configurations (RMSE 12.6 m versus 18.4–20.5 m). Thus, lower overall RMSE against the PANGAEA product does not imply improved agreement in the shallowest class, consistent with the shallow-layer limitations discussed in Sect. 4.1.

S3 Relationships between estimated ABLH and near-surface conditions

S3.1 Near-surface variables

Table S3 extends the selected correlations shown in Fig. 12 of the main text to the full set of near-surface variables examined in Sect. 4.3. Pearson and Spearman coefficients describe linear and monotonic associations, respectively; the valid sample size varies among variables.

Samples are matched 10-minute Met City observations and CMNN-estimated ABLH. Pearson and Spearman coefficients are both given; n is the number of valid pairs.

Table S3. Correlation between CMNN-estimated ABLH and near-surface variables

Variable	Description	n	Pearson r	Spearman r
wspd_vec_mean_10m	average wind speed at 10 m	20,307	0.732	0.795
bulk_tau	bulk wind stress	15,931	0.727	0.818
wspd_vec_mean_2m	average wind speed at 2 m	20,302	0.732	0.813

Variable	Description	n	Pearson r	Spearman r
ustar_10m	friction velocity at 10 m	20,278	0.713	0.816
vapor_pressure_deficit_2m_pa	saturation vapor pressure at 2 m minus vapor_pressure_2m	20,073	0.466	0.430
wind_shear_10m_2m_mps	Near-surface wind-speed difference (10 m minus 2 m; m s^{-1})	20,297	0.456	0.433
down_long_hemisp	downwelling longwave flux	16,334	0.409	0.437
epsilon_10m	turbulent kinetic energy dissipation rate at 10 m	20,278	0.428	0.779
skin_temp_surface	surface radiometric skin temperature	16,289	0.378	0.445
Hl	latent heat flux	10,606	0.398	0.475
surface_kinematic_moisture_flux_m_s	Kinematic moisture flux (m s^{-1})	19,291	0.361	0.279
surface_air_temp_diff_c	skin_temp_surface - temp_2m; surface-air temperature difference	15,977	0.372	0.454
atmos_pressure_2m	air pressure at 2 m	20,078	-0.293	-0.344
temp_gradient_10m_2m_c	Near-surface temperature difference (10 m minus 2 m; K)	20,078	-0.332	-0.485
Hs_10m	sensible heat flux at 10 m	20,278	0.301	0.279
net_longwave_wm2	down_long_hemisp - up_long_hemisp	16,289	0.302	0.283
surface_kinematic_sensible_heat_flux_K_m_s	Kinematic sensible heat flux (K m s^{-1})	20,038	0.297	0.270
buoyancy_flux_10m_m2_s3	surface buoyancy flux derived from sensible and latent heat fluxes	19,290	0.296	0.272
up_long_hemisp	upwelling longwave flux	18,074	0.267	0.369
temp_2m	air temperature at 2 m	20,078	0.264	0.331
temp_6m	air temperature at 6 m	20,283	0.261	0.330

Variable	Description	n	Pearson r	Spearman r
temp_10m	air temperature at 10 m	20,283	0.256	0.323
dew_point_2m	dew point temperature at 2 m	20,073	0.254	0.326
net_radiation_wm2	net_shortwave_wm2 + net_longwave_wm2	15,875	0.198	0.265
rh_2m	relative humidity with respect to water at 2 m	20,073	0.160	0.234
rh_6m	relative humidity with respect to water at 6 m	20,276	0.151	0.234
bulk_Hl_10m	bulk latent heat flux at 10 m	17,991	0.162	0.057
rh_10m	relative humidity with respect to water at 10 m	20,278	0.128	0.212
dewpoint_depression_2m_c	temp_2m - dew_point_2m; near-surface dew point depression	20,073	-0.121	-0.190
mixing_ratio_2m	mixing ratio derived using T/P/RH at 2 m	20,073	0.120	0.331
vapor_pressure_2m	water vapor pressure at 2 m	20,073	0.117	0.326
mixing_ratio_10m	mixing ratio derived using T/P/RH at 10 m	20,278	0.114	0.323
net_shortwave_wm2	down_short_hemisp - up_short_hemisp	18,470	-0.067	0.112
bulk_Hs_10m	bulk sensible heat flux at 10 m	17,991	0.081	0.074
down_short_hemisp	downwelling shortwave flux	18,476	-0.034	0.110
up_short_hemisp	upwelling shortwave flux	18,792	-0.023	0.103
zeta_level_n_10m	Monin-Obukhov stability parameter z/L at 10 m	20,278	-0.004	-0.404
bowen_ratio_10m	Hs_10m / Hl; Bowen ratio at 10 m	19,541	-0.003	-0.110

S3.2 Low-level jet control analysis

Table S4 supports the LLJ analysis in Sect. 4.3 and Fig. 13 of the main text by combining group statistics, wind correlations, height-dependent occurrence and the below-core wind control. These

diagnostics describe associations and potential confounding rather than an independent causal effect of LLJs.

Table S4. Below-core minimum control of the LLJ–CMNN-estimated ABLH relationship (n = 24,864). Correlations involving core speed, core height or the below-core minimum are computed for LLJ cases only (n = 10,846); the LLJ-flag correlation uses all records.

Part (a) summarises LLJ occurrence and the mean estimated ABLH in LLJ and non-LLJ cases, providing the group comparison reported in the main text.

(a) Overall statistics

Item	Value
Total samples	24,864
LLJ samples	10,846
LLJ frequency	43.6 %
Mean CMNN-estimated ABLH, LLJ samples	219.1 m
Mean CMNN-estimated ABLH, non-LLJ samples	151.1 m
Mean difference	68.0 m

Part (b) shows the correlations among estimated ABLH and the LLJ descriptors, including the strong association between core speed and the below-core minimum wind speed.

(b) Key correlations

Variable pair	Pearson r
LLJ core speed vs below-core minimum wind speed	0.788
ABLH vs below-core minimum wind speed	0.534
ABLH vs LLJ core speed	0.684
ABLH vs LLJ core height	-0.004
ABLH vs LLJ flag (Spearman)	0.286

The below-core minimum is the lowest wind speed at levels below the LLJ core in each LLJ profile. It is used as a lower-layer reference, not as an independent background: it is strongly correlated with core speed and partly defines the jet amplitude.

Part (c) shows how LLJ occurrence varies across estimated-ABLH classes, helping to assess the sampling differences underlying the LLJ versus non-LLJ comparison.

(c) LLJ frequency as a function of CMNN-estimated ABLH

CMNN-estimated ABLH bin (m)	Samples	LLJ samples	LLJ frequency	Mean CMNN-estimated ABLH (m)
≤100	9,071	2,393	0.264	64.7
100–200	7,194	3,361	0.467	146.9
200–300	4,789	2,714	0.567	241.1
300–400	1,702	998	0.586	343.0
400–500	1,007	698	0.693	445.8
500–700	955	581	0.608	577.7
>700	146	101	0.692	764.7

Part (d) compares the raw core-speed correlation with the correlation after removing the linear association of both variables with the below-core minimum. This minimum is a lower-layer reference rather than an independent measure of background flow.

(d) Background-wind control

Analysis	Correlation with ABLH
LLJ core speed, raw	r = 0.684
LLJ core speed, after removing below-core background from both variables	r = 0.505

LLJ core speed and the below-core minimum are strongly collinear (r = 0.788). After removing the below-core minimum from both core speed and CMNN-estimated ABLH, their correlation is 0.505, compared with 0.684 before control. The below-core minimum correlates with CMNN-estimated ABLH (r = 0.534) and contributes to the jet amplitude.

S4 Sensitivity analyses

S4.1 Temporal separation around radiosonde launches

The supplementary no-buffer experiment used 20,199 training and 8,658 validation records. Performance was compared on the same 834 launch-time test records as in the manuscript. In the reference split, development records were separated from the launch-time test records by at least 30 min on each side of every launch-time test record on the 10 min time grid, a 1 h total gap. For these 834 test records, the median distance on the 10 min time grid to the nearest development record was 30 min in the reference split and 10 min in the no-buffer split. Overall RMSE decreases from 34.04 to 23.35 m, MAE from 17.60 to 12.98 m, and R^2 increases from 0.9537 to 0.9782. Clear-sky RMSE decreases from 15.27 to 12.40 m, cloudy RMSE from 44.71 to 28.87 m and fog/haze RMSE from 25.06 to 21.33 m. For the 91 records above 400 m, RMSE decreases from 87.95 to 59.01 m. The contrast documents the sensitivity of evaluation to nearby development observations and supports temporal separation around the launch-time test records.

S4.2 Joint sensitivity to the diagnostic grid, Ri threshold and persistence setting

The primary diagnosis uses a 20 m grid, $Ri_c = 0.35$, a 30 m reference height and two consecutive levels. For a joint alternative setting, the same Ri profiles are interpolated to a 5 m diagnostic grid, Ri_c is set to 0.50, and four consecutive levels are required. The sensitivity search begins at 35 m, following the lower usable height in Jozef et al. (2024). Among 24,730 records with both diagnoses, mean ABLH is 178.5 m under the primary setting and 194.4 m under the alternative setting. The alternative-minus-primary difference has a mean bias of 15.8 m, MAE of 25.0 m, RMSE of 38.0 m and median of 13.6 m; the diagnoses correlate at r = 0.981. After excluding the 21.4 % of alternative diagnoses at the 35 m lower boundary, the mean difference is 26.0 m, MAE is 26.0 m, RMSE is 38.6 m and r = 0.985. The alternative joint setting shifts diagnosed heights upward while preserving their broad variability.

S4.3 Backscatter gradient clipping and reconstruction

The retained preprocessing sets negative backscatter values to zero, applies log1p compression, clips adjacent-level differences at their global 0.5th and 99.5th percentiles, and reconstructs the profile cumulatively from the lowest range gate. Its structural effect is evaluated on 834 MOSAiC launch-time profiles by comparing the log-compressed profile before clipping with the reconstructed profile. Within the lowest 20 range gates, 0.107 % of adjacent-level differences are clipped; the mean absolute reconstruction change is 0.0080 in log1p units, or 0.20 % of the mean pre-clipping feature. Within the lowest 48 gates, the corresponding values are 0.064 %, 0.0118

and 0.39 %. The median absolute change is zero in both regions. Profiles with clipped gradients inside these regions number 16 and 24, respectively, out of 834. The small aggregate changes support retention of the dominant lower-profile structure while limiting extreme gradients. Cumulative reconstruction proceeds upward, so clipping at an upper range gate affects subsequent higher gates

S5 Comparison of configurations across sky conditions

Figure S4 brings together the condition-specific RMSE results discussed in Sect. 4.1 and the NSA evaluation in Fig. 14 of the main text. It allows the input configurations and sampling strategies to be compared within each site; differences between sites also reflect their distinct temporal sampling and reference-height distributions.

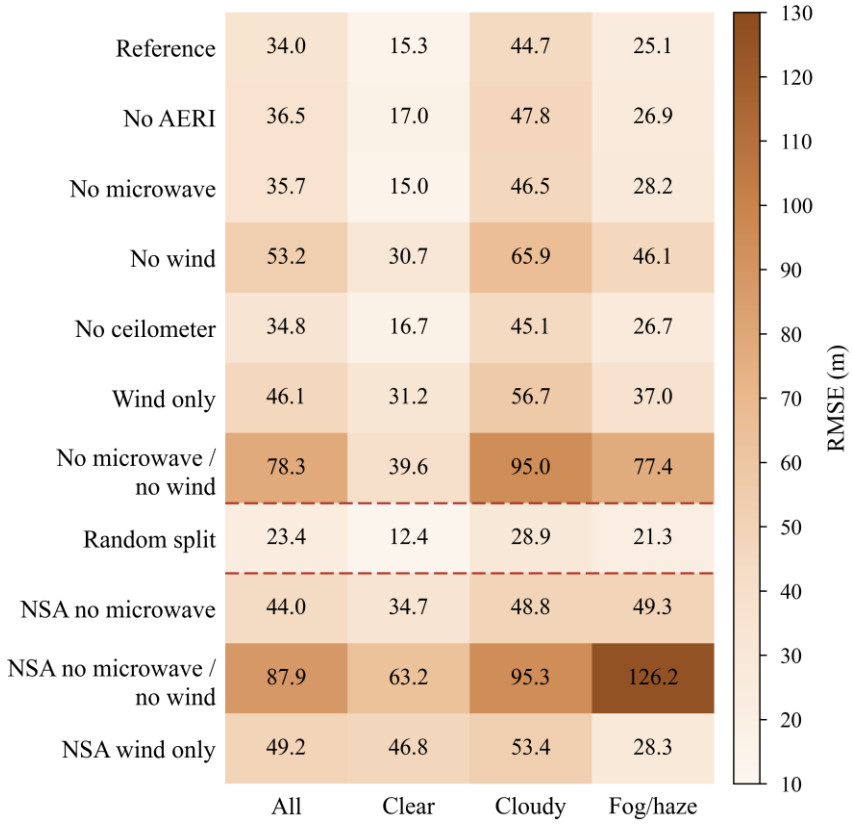


Fig. S4. RMSE of all trained configurations, separated by sky condition. The upper block shows MOSAiC estimates for the 834 launch-time records; the lower block shows the North Slope of Alaska (NSA) estimates for the 449 launch-time records. For the reported no-microwave transfer results at NSA, the model retains radiosonde-interpolated (INTERPSONDE) wind inputs, and the Ri-based reference ABLH is diagnosed from the same interpolated profiles (Sect. 4.4 of the main text). Rows differ in the observations supplied as input: the reference configuration uses all inputs, the ablation rows each remove one observation type, and the random-split row is the configuration without the launch-time buffer. Colours encode RMSE in metres. The thin red dashed line separates the random-split configuration from the ablation experiments, and the thicker red dashed line separates the MOSAiC and NSA datasets.