Response to Reviewers 2

Title: Synergistic estimation of atmospheric boundary layer height from multi-source observations over the Arctic Ocean during MOSAiC

Author(s): Yuliang Liu et al.

MS No.: egusphere-2026-4501

MS type: Research article

Response:

Thank you for your suggestions. All points have been addressed in the revised manuscript. The point-by-point responses below follow the order of the report, with the reviewer's text quoted verbatim.

Several changes concern the whole paper, not single sentences: the acronym ABL is used consistently after its definition, the verb "retrieve" has been replaced by "infer" or "estimate" throughout, the reference labels are referred to as a Richardson-number-based reference, reflecting their definition dependence, the figure numbers have been shifted by one after the insertion of the new comparison figure, and the persistence criterion of the reference labels has been reduced from four to two consecutive levels on the 20 m grid; a sensitivity comparison with the setting used by Jozef et al. (2024) is reported in Sect. S4.2.

Response to Comment #1: Major Comment 1

Reviewer Comment:

While the introduction gives a good general background on Arctic ABL structure and dynamics based on previous literature, it fails to give sufficient background on ABL studies for MOSAiC specifically. There have been numerous papers that have analyzed the ABL structure, stability, and forcings throughout the MOSAiC year that should be discussed as background information to set the stage for this paper. For example, the Jozef et al. 2023 and 2024 papers did a deep analysis of the seasonal variability of ABL stability and structure, and the key drivers, during MOSAiC. The current paper provides a very similar analysis, however gives no credit/discussion to what was already shared in the Jozef et al. papers other than surface-level citations for general background on Arctic ABL features and methods. The current paper needs to add discussion on how the type of results here differ from those shared in the Jozef et al. and other MOSAiC-specific ABL-focused papers, and also give credit to these papers when similar findings are shared.

Response:

Thank you for this comment. The original Introduction did not adequately discuss MOSAiC-specific ABL studies or distinguish the contribution of the present work from previous findings. We have expanded the literature review to include the studies of Jozef et al. (2023a, 2024), Peng et al. (2023), and Akansu et al. (2023), and added comparisons with their findings in the Results and Discussion.

The revised manuscript acknowledges that seasonal changes in ABL stability, cloud–ABL relationships, and associations between LLJs and ABL structure have already been investigated during MOSAiC. Jozef et al. (2023a) examined thermodynamic and kinematic controls on stability, while Jozef et al. (2024) used self-organizing maps to characterize recurring vertical structures and their relationships with clouds and LLJs. We now cite these studies where our results show corresponding seasonal, cloud-relative and LLJ-related patterns, and describe these results as consistent with earlier findings.

We also clarify the methodological contribution of the present study. The supervised CMNN combines microwave, AERI, ceilometer, wind-profile and cloud observations to estimate a Richardson-number-based reference ABLH at 10 min intervals. We evaluate its performance

across sky conditions, assess model dependence on each observation source, and examine the effect of replacing sounding-derived winds with tower and Doppler-lidar observations. The seasonal, cloud-base-relative and LLJ analyses characterize the CMNN-estimated ABLH and provide physical context for the model estimates. The contribution therefore lies in the development and evaluation of the multi-source estimator, including its sounding-input dependence, not in a new explanation of MOSAiC ABL variability.

The specific revisions are provided below.

Item 1 (lines 38–41).

Although observations from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) and other field campaigns have substantially advanced knowledge of Arctic ABL structure, challenges remain in representing its variability in numerical weather prediction at high latitudes, as reflected in documented forecast and reanalysis biases (Jozef et al., 2023a, 2024; Xi et al., 2024; Wesslén et al., 2014; Tjernström et al., 2012).

Item 2 (lines 49–50).

In the near-neutral regime, an elevated capping inversion is commonly present, with its base at approximately 200–300 m and its top extending to 1–2 km (Jozef et al., 2024).

Item 3 (lines 101–104, 108–110).

ML has already been applied to MOSAiC ABL observations: Jozef et al. (2024) used self-organizing maps to classify recurring vertical structures. Complementary analyses by Jozef et al. (2023a) examined the thermodynamic and kinematic drivers of ABL stability from multiple observations.

Building on these studies, we investigate whether condition-adaptive fusion of microwave, AERI, ceilometer and wind-profile observations can provide continuous estimates of a Richardson-number-based reference ABLH.

Item 4 (lines 428–429, 438–439).

Sections 4.2 and 4.3 examine seasonal and regime-dependent patterns in CMNN-estimated ABLH and their consistency with previous MOSAiC studies.

These seasonal patterns in the CMNN estimates are consistent with previous MOSAiC findings (Jozef et al., 2023a, 2024; Peng et al., 2023).

Item 5 (lines 582–585).

Reference ABLH differs among cloud-base categories, with seasonal means of approximately 100–300 m in the coupled and CBH-above-ABLH categories, compared with 454 m in winter and 562 m in spring when ABLH exceeds CBH. These contrasts are consistent with the diverse cloud–ABL structures documented during MOSAiC (Jozef et al., 2024; Peng et al., 2023).

Item 6 (lines 515–516).

LLJs are detected in 43.6 % of the analysed 10-min records, compared with approximately 76 % of the 1,377 radiosonde profiles examined by Jozef et al. (2024).

Item 7 (lines 586–590).

Wind stress and friction velocity show positive correlations with the CMNN-estimated ABLH, consistent with a role for mechanical mixing. Associations with downward longwave radiation, surface–air temperature difference and latent heat flux indicate additional radiative and thermodynamic influences. Mean CMNN-estimated ABLH is also higher in LLJ cases than in non-LLJ cases (219.1 versus 151.1 m). These relationships agree with previous MOSAiC findings (Jozef et al., 2023a, 2024).

Response to Comment #2: Major Comment 2

Reviewer Comment:

Major Comment 2. Are you aware of the dataset at https://doi.pangaea.de/10.1594/PANGAEA.957760, which provides information about the atmospheric boundary layer (depth and stability), temperature features (near-surface temperature and temperature inversion characteristics), wind features (near-surface wind speed and low-level jet characteristics), moisture features (near-surface mixing ratio and cloud characteristics), and surface radiation budget (up- and downwelling longwave and shortwave radiative flux) at the time of each MOSAiC radiosonde launch? This dataset contains a lot of features you are also pulling together in your paper. It would strengthen your paper and add trust in your truth labeling if you used this dataset to support or validate your ABLH truth labels for the non-interpolated radiosonde launch times. Inclusion or discussion of this dataset also helps support that you have done sufficient analysis and consideration of previous works on this topic.

Response:

Thank you for your suggestions. We have now (i) cited and described the dataset in Sect. 2.1, (ii) added a systematic launch-time comparison of our reference labels with this dataset and with the ARM radiosonde ABLH product in Sect. 3.2 (Figure 1 in this response; Figure 3 in the main text), and (iii) clarified that the reference labels are definition-dependent and that the comparison quantifies differences between diagnostic methods.

The comparison uses the 834 launch-time records of our reference-label table anchored to actual radiosonde launches. It is restricted to radiosonde launch times and does not evaluate the temporally interpolated ABLH between launches. The estimates are matched within 5 min, and the ARM estimates are restricted to quality-control value 0. At these launch times, 825 records match the Jozef dataset and 568 match the ARM product; 560 launches have all six estimates and form the common sample shown in Figure 1 (Figure 3 in the main text).

On the 560 launches with all six diagnoses available, our reference labels have an RMSE of 106.8 m, an MAE of 54.1 m and a mean positive difference of 46.3 m relative to the Jozef dataset (study minus Jozef), with a Pearson correlation of 0.81. Of these launches, 74.6 % differ by no more than 50 m, and the median absolute difference is 22.5 m. Figure 1 reports bias as row minus column: its Jozef-row/Study-column value is therefore −46.3 m, the same difference expressed in the opposite direction. The depth-stratified comparison in Table S2 uses the smaller 445-launch subset with paired predictions from all configurations and consequently has different sample-specific statistics. These estimates apply different diagnostic methods to the same radiosonde observations; the comparisons quantify agreement and systematic differences between methods.

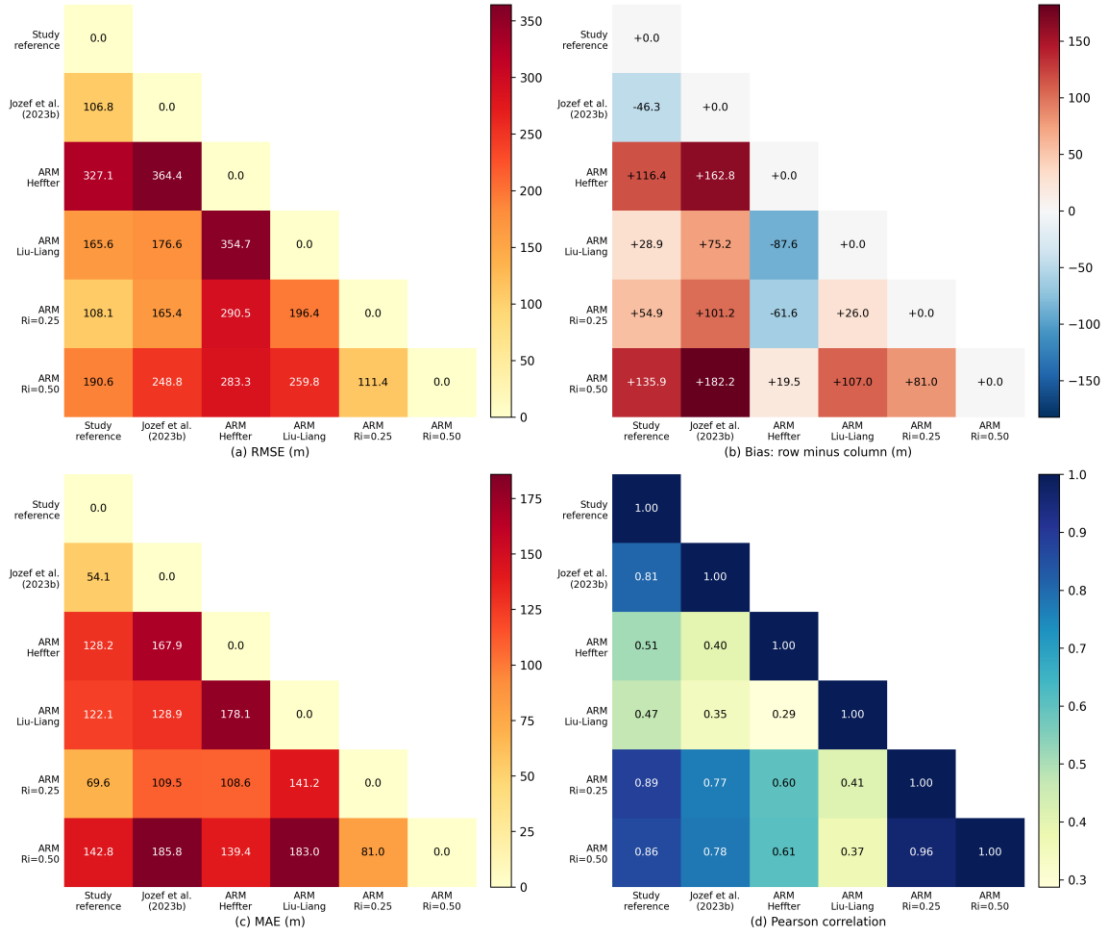


Figure 1 (Figure 3 in the main text). Pairwise comparison of six ABLH estimates for the same 560 radiosonde launches: the Ri-based reference labels of this study, the ABLH of the MOSAiC lower-atmospheric-properties dataset (Jozef et al., 2023b), and four estimates from the ARM radiosonde ABL-height product (Heffter, 1980; Liu and Liang, 2010; bulk Richardson number thresholds of 0.25 and 0.50). (a) RMSE, (b) bias (row minus column; positive when the row dataset is higher), (c) MAE, and (d) Pearson correlation. The lower triangle is shown, because RMSE, MAE and r are symmetric and bias is antisymmetric. In panel (b), the Jozef row and Study reference column show −46.3 m (Jozef minus Study), equivalent to the +46.3 m study-minus-Jozef difference reported in the text. All ARM estimates are restricted to quality-control value 0. These statistics quantify the consistency between diagnostic methods and do not identify any dataset as an error-free truth.

The specific revisions are provided below.

Item 1 (lines 138–140).

The MOSAiC lower-atmospheric-properties dataset (Jozef et al., 2023b) provides ABLH, stability and related atmospheric properties at each radiosonde launch and is used as an alternative diagnosis in Sect. 3.2.

Item 2 (lines 244–249).

To quantify label-definition uncertainty, the launch-time labels are compared with alternative diagnoses of the same radiosonde profiles: the ABLH of the Jozef dataset (Jozef et al., 2023b) and four estimates from the ARM ABL-height product, namely the Heffter (1980) and Liu and Liang (2010) methods and bulk Richardson thresholds of 0.25 and 0.50 (ARM, 2023). All six are matched to the same launches within 5 min, and the ARM estimates are restricted to quality-

control value 0. Of the 834 launch records, 825 match the Jozef dataset and 568 the ARM product; 560 launches have all six estimates and form the common sample of Figure 3.

Item 3.

The following references have been added to the revised manuscript.

Jozef, G. C., Klingel, R., Cassano, J. J., Maronga, B., de Boer, G., Dahlke, S., and Cox, C. J.: Lower atmospheric properties relating to temperature, wind, stability, moisture, and surface radiation budget over the central Arctic sea ice during MOSAiC [data set], PANGAEA, https://doi.org/10.1594/PANGAEA.957760, 2023b.

Jozef, G. C., Klingel, R., Cassano, J. J., Maronga, B., de Boer, G., Dahlke, S., and Cox, C. J.: Derivation and compilation of lower-atmospheric properties relating to temperature, wind, stability, moisture, and surface radiation budget over the central Arctic sea ice during MOSAiC, Earth System Science Data, 15, 4983–4995, https://doi.org/10.5194/essd-15-4983-2023, 2023c.

Heffter, J. L.: Transport layer depth calculations, in: Second Joint Conference on Applications of Air Pollution Meteorology, American Meteorological Society, 787–791, 1980.

Atmospheric Radiation Measurement (ARM) User Facility: Planetary Boundary Layer Height (PBLHT), ARM Mobile Facility (MOS) MOSAIC, ARM Data Center [data set], https://doi.org/10.5439/1991783, 2023.

Response to Comment #3: Major Comment 3

Reviewer Comment:

Line 133. I am not convinced of the validity of using samples within 1 hr before and after each radiosonde as validation samples only. The quality and accuracy of the data used for training/testing vs validation should be the same, otherwise the model won't train as well as it could and/or your performance evaluation results can be misleading. By explicitly reserving the data surrounding the radiosonde launches for validation, it seems to me that you are actually introducing an influence of sounding interpolation and temporal dependence, rather than reducing it. Your training/testing and validation sets should be a random (yet representative) selection of the total available samples, rather than a hand-picked subset.

Response:

Thank you for your suggestions. The model is developed on a random split of the continuous INTERPSONDE series, as the reviewer suggests. To limit temporal dependence, 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. The remaining model-development data are divided into training and validation subsets. The validation subset (7,459 records) is used for checkpoint selection and early stopping. The test set consists of 834 launch times that were not used for training or model selection.

The test records are defined by the launch times themselves, not selected by hand. They do not participate in training or model selection, and they provide an independent test set at withheld times.

To examine the role of temporal separation, 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. 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. On the same 834 launch-time test records used in the manuscript, the no-buffer experiment yielded an RMSE of 23.35 m and an R^2 of 0.9782, compared with 34.04 m and 0.9537 under the original split. Cloudy-condition RMSE was 28.87 m without the buffer and 44.71 m under the original split. The lower errors obtained when nearby development records are retained

show the sensitivity of the evaluation to temporal proximity. Together with the increased separation of development and test records, these results support the use of the buffer to reduce the influence of temporal dependence in the continuous observations and interpolated sounding profiles and to provide a more stringent performance assessment.

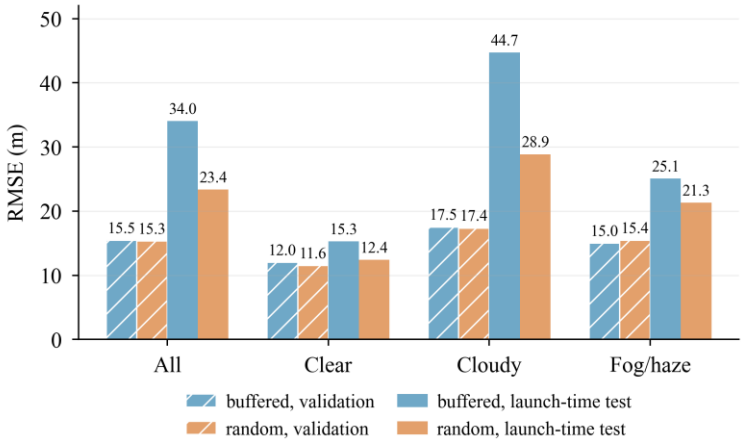


Figure R1. Validation and launch-time RMSE for the two splitting strategies. Bars show the RMSE over all records and for clear-sky, cloudy and fog/haze conditions; all values are in metres. Hatched bars denote each experiment's internal validation set (7,459 records for the buffered split and 8,658 for the random split), and solid bars denote the same 834 launch times used for testing in both experiments. Blue denotes the buffered split used in the manuscript and orange the random split without the buffer. The two validation sets are internal to their own experiments and are not a paired sample.

The specific revisions are provided below.

Item 1 (lines 319–327).

This section evaluates the performance of the condition-adaptive multi-source network (CMNN) against the Ri-based reference ABLH. To limit temporal dependence, 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. The remaining development data were divided into training and validation sets, with the validation set comprising 7,459 randomly selected records used for checkpoint selection and early stopping. Performance is reported separately for this internal validation set and for an evaluation set of 834 launch times withheld from model development. The launch-time evaluation also enables comparison with alternative sounding-based ABLH diagnoses. In a supplementary experiment without the temporal buffer, RMSE decreases from 34.04 to 23.35 m on the same 834 launch-time records (Sect. S4.1). This difference demonstrates the sensitivity of the evaluation to temporally adjacent development samples and supports retaining the buffer for the reported test performance.

Item 2 (lines 329–335).

In the validation set, the RMSE is 12.0 m under clear skies, 17.8 m under cloudy conditions and 15.0 m in fog/haze; in the test set (834 launch times), the corresponding values are 15.27, 44.71 and 25.06 m, respectively, and R^2 exceeds 0.9 in all three categories (Figure 5). Differences in sampling, temporal dependence and reference-label uncertainty may contribute to the validation–test gap. The model therefore reproduces the reference well under clear skies but only moderately

under cloudy conditions, where the RMSE is higher than under the other conditions and some cases show large discrepancies. The degradation under cloud is likely related to the reduced sensitivity of AERI above optically thick cloud and to the altered backscatter structure observed by the ceilometer.

Response to Comment #4: Major Comment 4

Reviewer Comment:

Major Comment 4. I am not convinced that using your ML model to infer ABLH adds value on top of simply using the radiosonde or INTERPSONDE data as a single source for ABLH detection. Since you are using the radiosonde data as a feature for your model, as well as for the truth label, then I don't see what the purpose is for having the rest of the multimodal sensors. If you were arguing that multimodal sensors and an ML model add value over radiosondes because they take continuous measurements and don't require manual deployment multiple times per day, then I would understand the value, however since the radiosonde data are still a feature in your model, this argument cannot be made. While the methodology is interesting, I don't see what the use case would be, because if you need radiosonde data for the ML model, why not just use the radiosonde-based ABLH retrieval alone? Radiosonde-based ABLH is much simpler to compute, and is already suggested to be well trusted, as it is used for the model truth labeling. One idea to improve the value of an ML model would be to remove the radiosonde data from the model features to create a radiosonde-independent ABLH estimation method, where you could then make the above argument. Another idea would be to hand label the "true" ABLH values using the multimodal observations, and then you could argue that the value of the ML model is to pull information out of all of the multimodal observations for a highly accurate ABLH estimate without having to hand label. I'd recommend you strengthen your methods one of these ways, or if there is something I am missing, please explain more thoroughly how using your current ML model adds value over traditional algorithmic ABLH detection methods.

Response:

Thank you for your suggestions. The original manuscript did not adequately distinguish two questions: whether the additional observations improve estimation beyond sounding wind alone, and whether the estimator can operate without sounding-derived inputs. We use the sounding-wind configuration as a performance reference, and address the practical application through a configuration in which tower and Doppler-lidar winds replace all sounding-derived inputs. Sounding temperature, humidity and pressure are used to diagnose the labels, not as predictive inputs.

To address the role of the other sensors, we compare five input configurations on the same 834 MOSAiC launch-time records, using a consistent evaluation procedure and each model's training-set normalisation parameters (Table R4). The full-input model reduces RMSE from 46.09 m for wind alone to 34.04 m, a reduction of 12.05 m (26.1 %). This provides evidence that the other observations and auxiliary features collectively add information beyond wind for estimating the Ri-defined reference. The no-wind model gives an RMSE of 53.17 m. Removing microwave inputs gives 35.67 m, only 1.63 m above the full-input result, indicating that their overall marginal benefit is modest. Without both microwave and wind inputs, the remaining AERI, ceilometer, cloud, cloud-base-height and time features give an RMSE of 78.25 m and an R^2 of 0.755. Thus, the remaining inputs have predictive skill, but their combination with wind performs substantially better. These comparisons quantify the collective contribution of the additional observations.

The strong wind-only result also requires qualification. Wind is used both as a predictor and in the Ri diagnosis, so the wind contribution combines atmospheric information with input–label

coupling. Neither the wind-only experiment nor the source-removal experiments separate these contributions. We therefore interpret the differences in Table R4 as changes in agreement with the specified reference, not as proof that the model detects the true ABLH more accurately than conventional diagnoses.

To address the suggested removal of sounding inputs, we evaluate the Doppler-lidar configuration using tower winds at 2, 6 and 10 m and Doppler-lidar winds at 16 gates between 72 and 502 m; the other inputs and training protocol are unchanged. On the same 661 launch times, replacing INTERPSONDE winds increases RMSE from 36.6 to 53.5 m and decreases R^2 from 0.95 to 0.89 (Table 1, reproduced in Item 3 below). The 10–72 m gap is interpolated between the tower and the first lidar gate without sounding data. Limited vertical coverage and reduced availability of lidar winds under cloud may contribute to the error increase. This configuration provides the practical use case: estimation from tower and remote-sensing observations at inference time, without requiring a concurrent sounding input. It is potentially applicable at sites with suitable observations but no routine sounding programme, although performance at a new site would require evaluation.

Table R4. Input-configuration experiments reported in this response, evaluated on the same 834 MOSAiC launch-time records as the manuscript. All five configurations are evaluated on the same 834 launch-time records, with inputs normalised using the respective training-set parameters. These records are withheld from model development. Timestamps, Ri reference labels and sky categories are identical across configurations. This sample differs from the 661 paired launches in Table 1. Bias is prediction minus reference; RMSE, MAE and bias are in metres.

Configuration	n	RMSE	MAE	Bias	R^2
Full inputs	834	34.04	17.60	-4.83	0.95
No wind	834	53.17	30.36	-11.64	0.89
No microwave	834	35.67	18.63	-6.27	0.95
Wind only	834	46.09	28.46	-2.42	0.92
No microwave / no wind	834	78.25	46.86	-14.85	0.76

Note: n is the number of launch-time records and R^2 is the coefficient of determination.

The added value demonstrated here is therefore a multi-source estimator that can retain useful agreement with the Ri-based reference after sounding inputs are replaced, together with a quantified contribution of the other inputs beyond sounding wind alone. Both the sounding-wind and Doppler-lidar configurations remain trained and evaluated against sounding-derived labels. The new input-restriction comparisons are reported here to clarify the contribution; the corresponding manuscript passages on inputs, Doppler substitution and source dependence are provided below.

The specific revisions are provided below.

Item 1 (lines 182–193, 200–205).

The input features fall into two groups: instrumental observations and auxiliary parameters. The instrumental observations comprise passive brightness temperatures from 22 microwave channels (HATPRO and MiRAC-P) and 50 AERI channels selected by mRMR; an active ceilometer backscatter profile with a vertical resolution of 10 m and 200 levels; and the INTERPSONDE wind profile on 100 levels with 20 m vertical spacing, spanning 10–1990 m. The thermodynamic information enters through the radiances instead. The auxiliary parameters include cloud state,

cloud-base height and related single-level descriptors. The model trained with these inputs is hereafter referred to as the reference configuration. INTERPSONDE wind profiles are used both as inputs to the reference configuration and in the diagnosis of the Ri-based reference ABLH, creating shared wind information between the model inputs and reference labels. This dependence may facilitate reproduction of the wind-related relationships embedded in the reference diagnosis, so part of the reported agreement may reflect access to information used to construct the labels. To assess and reduce this dependence, we use input ablation, alternative wind observations and independent-site evaluation to examine the incremental contribution of the other observations, the dependence on shared wind inputs and the transferability of the input–reference ABLH relationship. Second, we train a Doppler-lidar configuration, hereafter the Doppler configuration, in which INTERPSONDE winds are replaced by tower winds at 2, 6 and 10 m and Doppler-lidar winds at 16 range gates between 72 and 502 m. All other inputs, reference labels and the training protocol remain identical. The 10–72 m gap between the tower and lidar observations is filled by interpolation between the two observation sources. The two configurations are compared at the same launch times with paired predictions to examine whether the model can still reproduce variations in the reference ABLH without directly receiving the INTERPSONDE wind profiles used in its diagnosis (Sect. 3.1).

Item 2 (lines 394–405).

The Doppler-lidar experiment described in Sect. 3.1 assesses whether the model can reproduce the reference ABLH without directly receiving the INTERPSONDE winds used in its diagnosis. On the same 661 launch times with paired predictions, replacing INTERPSONDE winds with tower and Doppler-lidar observations increases RMSE from 36.6 to 53.5 m and decreases R^2 from 0.950 to 0.894 (Table 1). The RMSE increase is smaller under clear skies (4.9 m) than under cloudy conditions (22.2 m). Despite this reduction in performance, the Doppler configuration retains substantial agreement with the reference labels, demonstrating that direct access to the INTERPSONDE wind profiles is not required to reproduce much of the variation in the Ri-based reference ABLH. The performance reduction may reflect both the removal of shared INTERPSONDE wind information and differences between the wind observations. The Doppler-lidar winds extend only to 502 m and may therefore miss wind shear above this height, while reduced availability of upper-level lidar winds under cloudy conditions may further limit the input information. These effects cannot be separated by the present comparison. Nevertheless, the retained performance after replacing the shared wind input indicates that the agreement of the reference configuration is not solely attributable to the direct use of winds involved in label construction.

Item 3 (Table 1 caption).

Table 1. Launch-time performance of the sounding-wind and Doppler-lidar configurations on the same 661 paired launches. Here sonde denotes the reference configuration, in which the wind input is taken from INTERPSONDE; lidar denotes the Doppler-lidar configuration, in which the INTERPSONDE wind is replaced by tower and Doppler-lidar winds. All other inputs and the training protocol are identical. RMSE and bias are in metres.

Condition	**n**	**RMSE-sonde**	**RMSE-lidar**	**Bias-sonde**	**Bias-lidar**	**R^2-sonde**	**R^2-lidar**
All	661	36.6	53.5	−5.5	−6.0	0.95	0.89
Clear	193	15.8	20.7	+1.7	−5.4	0.97	0.96
Cloudy	322	48.3	70.5	−9.1	−6.0	0.93	0.86
Fog/haze	146	24.3	37.6	−7.2	−6.9	0.97	0.92

Note: n is the number of paired launch times, bias is prediction minus reference, and R^2 is the coefficient of determination.

Item 4 (lines 387–393).

This dependence reflects both the information carried by the wind profiles and their shared use in constructing the Ri-based reference labels. On the same 834 launch-time records, the full-input model reduces RMSE from 46.09 m for the wind-only model to 34.04 m, indicating additional predictive information from the remaining observations and auxiliary features. Removing AERI increases RMSE by 1.7 m under clear skies, 3.1 m under cloudy conditions and 1.8 m in fog/haze, indicating a modest improvement in reference agreement from the infrared inputs in each sky category. Microwave and ceilometer observations make smaller and more consistent contributions. Figure S4 summarises the RMSE of all configurations under each sky condition.

Response to Comment #5: Major Comment 5

Reviewer Comment:

Major Comment 5. Line 260–264: With "It should be noted that the training labels are diagnosed from INTERPSONDE profiles using the Richardson number method, and the input also includes INTERPSONDE wind-field information" you bring up an important point that I have been concerned about while reading the paper up until this point, but I am still left with two confusions:

a. If the INTERPSONDE data are used for the training labels AND as feature data within the model, then why are any of the other input sources other than the INTERPSONDE data needed? If you already have the INTERPSONDE data, why would you need the ML model at all, since your model is trained on the fact that the INTERPSONDE-derived ABLH are truth?

b. When you continue with "Therefore, …" following the above sentence it sounds like you are going to explain how the statement of the previous sentence is mitigated, but instead you just explain what a test set vs validation set is, and I don't see how that has anything to do with your previous point. A test set will still refer to the truth labels determined from the INTERPSONDE product, as does the validation set. This leaves me lost as to what you are trying to do with this few sentence section.

Response:

Thank you for your suggestions. We address these concerns through complementary evaluations: input-configuration comparisons quantify the contribution of the additional observations, the Doppler-lidar experiment tests estimation without sounding-derived inputs at inference, and the held-out launch records assess performance outside model development. The revised manuscript distinguishes these roles and provides the supporting results, as detailed below.

Regarding (a), wind is the only sounding-derived predictive input; sounding temperature, humidity and pressure are used for the label diagnosis. On the same 834 launch-time records, the full-input model reduces RMSE from 46.09 m for the wind-only model to 34.04 m, showing that the other observations and auxiliary features collectively provide additional predictive information about the Ri-defined reference (Table R4 in our response to Major Comment 4). The practical application is instead examined using the Doppler-lidar configuration, in which tower and lidar winds replace the sounding input. On 661 paired launch times, RMSE changes from 36.6 to 53.5 m and R^2 from 0.95 to 0.89 (Table 1). This configuration uses no sounding-derived inputs at inference, although its training and evaluation labels remain sounding-derived. The 10–72 m gap is filled by interpolation between tower and lidar observations.

Regarding (b), the revised evaluation design separates model selection from launch-time performance assessment. To limit temporal dependence, 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. The remaining records are randomly divided into training and validation subsets. The 7,459 validation records guide checkpoint selection and early stopping; the 834 launch-time test records are reserved for evaluation and are not used for training or model selection. This makes the launch-time results an assessment on records withheld from model development, with the same Ri-based definition used consistently across configurations. The shared-source concern is examined separately by replacing INTERPSONDE wind with tower and Doppler-lidar observations: the resulting RMSE of 53.5 m and R^2 of 0.89 show that the estimator retains substantial skill without access to sounding-derived inputs at inference. Together with the full-input versus wind-only comparison, these experiments provide evidence both for additional predictive information from the other inputs and for an application without concurrent sounding input. The reference remains a sounding-derived Ri diagnosis, so the reported skill is interpreted as agreement with this diagnostic target.

The revised manuscript presents this evidence in the discussion of the two wind configurations, the input-contribution analysis and the launch-time evaluation. The validation/test paragraph explains the separation of model selection and performance assessment, while the Doppler substitution and wind-only comparison directly address the input dependence and the value of the other observations. The corresponding manuscript passages are provided below.

The specific revisions are provided below.

Item 1 (lines 387–393).

This dependence reflects both the information carried by the wind profiles and their shared use in constructing the Ri-based reference labels. On the same 834 launch-time records, the full-input model reduces RMSE from 46.09 m for the wind-only model to 34.04 m, indicating additional predictive information from the remaining observations and auxiliary features. Removing AERI increases RMSE by 1.7 m under clear skies, 3.1 m under cloudy conditions and 1.8 m in fog/haze, indicating a modest improvement in reference agreement from the infrared inputs in each sky category. Microwave and ceilometer observations make smaller and more consistent contributions. Figure S4 summarises the RMSE of all configurations under each sky condition.

Item 2 (lines 319–328).

This section evaluates the performance of the condition-adaptive multi-source network (CMNN) against the Ri-based reference ABLH. To limit temporal dependence, 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. The remaining development data were divided into training and validation sets, with the validation set comprising 7,459 randomly selected records used for checkpoint selection and early stopping. Performance is reported separately for this internal validation set and for an evaluation set of 834 launch times withheld from model development. The launch-time evaluation also enables comparison with alternative sounding-based ABLH diagnoses. In a supplementary experiment without the temporal buffer, RMSE decreases from 34.04 to 23.35 m on the same 834 launch-time records (Sect. S4.1). This difference demonstrates the sensitivity of the evaluation to temporally adjacent development samples and supports retaining the buffer for the reported test performance.

Item 3 (lines 401–405).

The Doppler-lidar winds extend only to 502 m and may therefore miss wind shear above this height, while reduced availability of upper-level lidar winds under cloudy conditions may further limit the input information. These effects cannot be separated by the present comparison. Nevertheless, the retained performance after replacing the shared wind input indicates that the

agreement of the reference configuration is not solely attributable to the direct use of winds involved in label construction.

Response to Comment #6: Major Comment 6

Reviewer Comment:

Major Comment 6. Sections 4.2 and 4.3: The paper up until this point focuses on the CMNN model and its ability to infer ABLH. This section now pivots to an analysis of ABLH structure and forcings throughout the year, and it is unclear if this analysis was done using the CMNN-derived ABLH or the "true" Ri-based ABLH. If the latter, then the sudden shift to an analysis of ABLH seems out of alignment with the purpose of this paper (which I thought was to share the CMNN methodology and efficacy). Further, this section presents findings that are similar to findings shared in a lot if previous literature. If you are going to shift the focus of the paper to ABLH analysis, this needs to be prefaced, including a discussion on why this analysis is needed (i.e., what is missing from the previous literature that you are adding here).

Response:

Thank you for this comment. The ABLH variability and association analyses in Sects. 4.2 and 4.3 use the CMNN estimates: Figure 9 shows their seasonal distributions, and Figures 12 and 13 examine their relationships with meteorological variables and LLJs. Figure 10 explicitly compares model estimates with the Ri-based reference. The reference-based stability and cloud categories in Figure 11 provide the observational grouping context for these analyses; they are identified separately in the text and caption.

We have added a brief transition at the start of Sect. 4.2 to connect these analyses to the model evaluation. Beyond pointwise errors, they examine the seasonal and regime-dependent behaviour of the model output and its consistency with established Arctic ABL characteristics. Comparisons with the reference are identified explicitly. Relevant MOSAiC studies are cited to place similar patterns in context. The analyses relate the estimator's behaviour to established Arctic ABL processes.

The specific revisions are provided below.

Item 1 (lines 428–439).

Sections 4.2 and 4.3 examine seasonal and regime-dependent patterns in CMNN-estimated ABLH and their consistency with previous MOSAiC studies. The CMNN-estimated ABLH shows marked seasonal variability during MOSAiC, reaching a monthly mean of 356.7 m in May and its lowest monthly means in January (129.9 m) and July (140.6 m) (Figure 9). The monthly distributions are computed from the 10-min estimates and contain between 1,040 (June) and 3,537 (December) records per month; the sample counts quoted in Sect. 4.1 refer to the launch-time evaluation set. The low January mean coincides with strong lower-atmospheric stratification, with a mean potential-temperature gradient of 24.4 K km^{-1} over 0–400 m. By May, this gradient decreases to approximately 2.7 K km^{-1}, consistent with reduced suppression of vertical mixing and a deeper ABL. The summer shallowing is consistent with limited surface warming over melting sea ice (Vihma et al., 2014b; Intrieri et al., 2002). Warm, moist air advected over colder sea ice can also favor a shallow stable surface layer, potentially decoupled from an overlying cloud or fog layer, providing a further possible explanation for the low summer ABLH (Tjernström et al., 2019; Vihma et al., 2014b). These seasonal patterns in the CMNN estimates are consistent with previous MOSAiC findings (Jozef et al., 2023a, 2024; Peng et al., 2023).

These seasonal patterns in the CMNN estimates are consistent with previous MOSAiC findings (Jozef et al., 2023a, 2024; Peng et al., 2023).

Item 2 (Figure 9 caption).

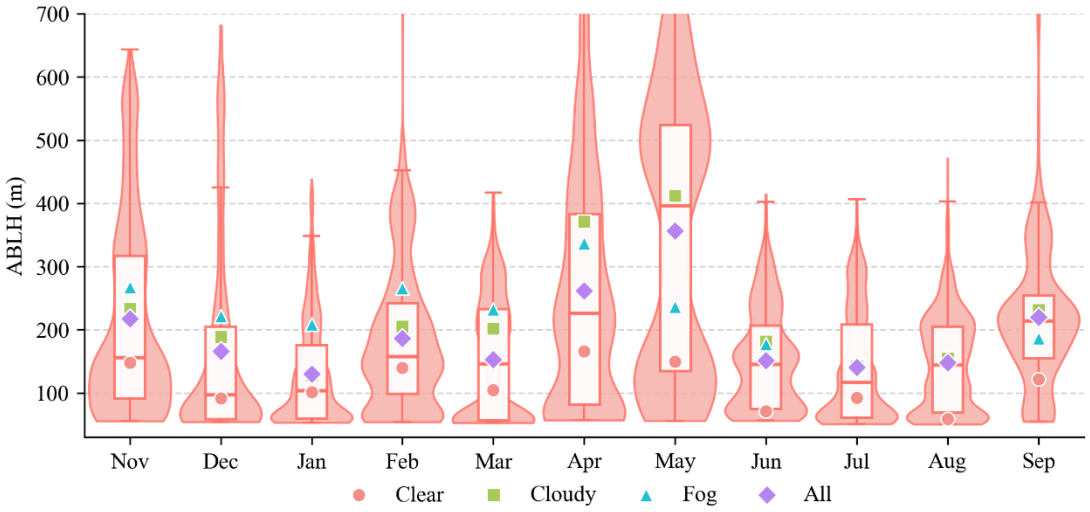


Figure 9. Monthly distribution of the CMNN-estimated ABLH. Pink violins show the overall distribution density; embedded white boxplots indicate the 25th, 50th, and 75th percentiles. Colored symbols denote monthly mean ABLH under different weather conditions.

Item 3 (Figure 10 caption).

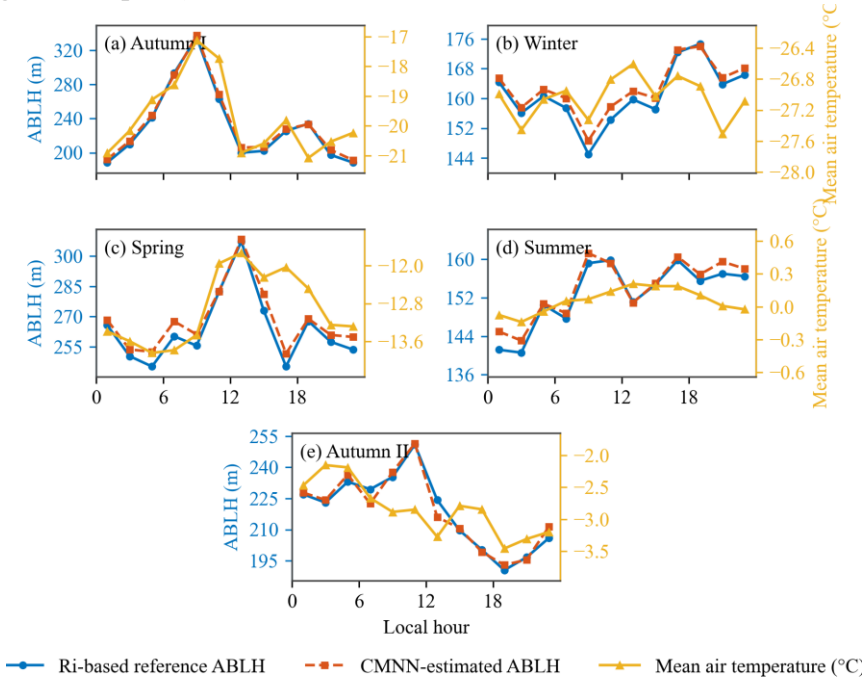


Figure 10. Diurnal variation of the 2-hour mean ABL height in five seasonal periods following Shupe et al. (2022): (a) autumn I (2019), (b) winter, (c) spring, (d) summer and (e) autumn II (2020), as a function of local hour. The blue line is the Ri-based reference ABLH, the orange dashed line the CMNN-estimated ABLH, and the yellow line the mean near-surface air temperature (°C). The periods differ in duration and sampling, so comparisons between their diurnal cycles are qualitative.

Response to Comment #7: Major Comment 7

Reviewer Comment:

Major Comment 7. The paper needs discussion on the impact of sample size under varying conditions on the performance of the model and the significance of the results.

Response:

Thank you for your suggestions. The paper reported error metrics without saying how many records supported each of them, which made it impossible to judge which differences between conditions were meaningful. We have added that information, together with an uncertainty analysis, and we have rephrased the places where a month-to-month difference was described as if it rested on a large sample. The revised text reports the sample sizes and bootstrap confidence intervals directly; the earlier qualitative caveat has been removed.

Estimation errors vary with both sky condition and reference ABLH. The launch-time evaluation includes 244 clear-sky, 393 cloudy and 197 fog/haze records. RMSE is higher under cloudy conditions than under clear skies or fog/haze, by 29.4 m (95 % bootstrap CI 15.2–44.0 m) and 19.6 m (95 % CI 5.0–33.4 m), respectively. Errors also increase with reference ABLH, from 12.1 m for 50–100 m layers to 15.1 m for 100–200 m, 26.8 m for 200–400 m and 87.9 m above 400 m. Normalized by the mean reference ABLH in each bin, RMSE is about 10 % at 100–400 m, compared with 20.0 % at 50–100 m and 16.1 % above 400 m; shallow layers therefore have small absolute errors but comparatively large relative errors. Sampling is uneven across months and height ranges (Figure S2), with 114 launch-time records in December but only 33 in June and 46 in May. The deepest bin contains 91 records, and its RMSE has a wide 95 % confidence interval of 57.4–115.8 m, indicating that the error magnitude for deep layers is less precisely estimated.

The specific revisions are provided below.

Item 1 (lines 336–341).

Estimation errors vary with both sky condition and reference ABLH. The launch-time evaluation includes 244 clear-sky, 393 cloudy and 197 fog/haze records. RMSE is higher under cloudy conditions than under clear skies or fog/haze, by 29.4 m and 19.6 m respectively. Errors also increase with reference ABLH, from 12.1 m for 50–100 m layers to 15.1 m for 100–200 m, 26.8 m for 200–400 m and 87.9 m above 400 m. Normalized by the mean reference ABLH in each bin, RMSE is about 10 % at 100–400 m, compared with 20.0 % at 50–100 m and 16.1 % above 400 m; shallow layers therefore have small absolute errors but comparatively large relative errors.

Item 2 (lines 342–344).

Sampling is uneven across months and reference-ABLH ranges (Figure S2), with 114 launch-time records in December but only 33 in June and 46 in May. For reference ABLH above 400 m, only 91 records are available, and the RMSE has a wide 95 % confidence interval of 57.4–115.8 m, indicating greater uncertainty in the estimated error for deep boundary layers.

Item 3.

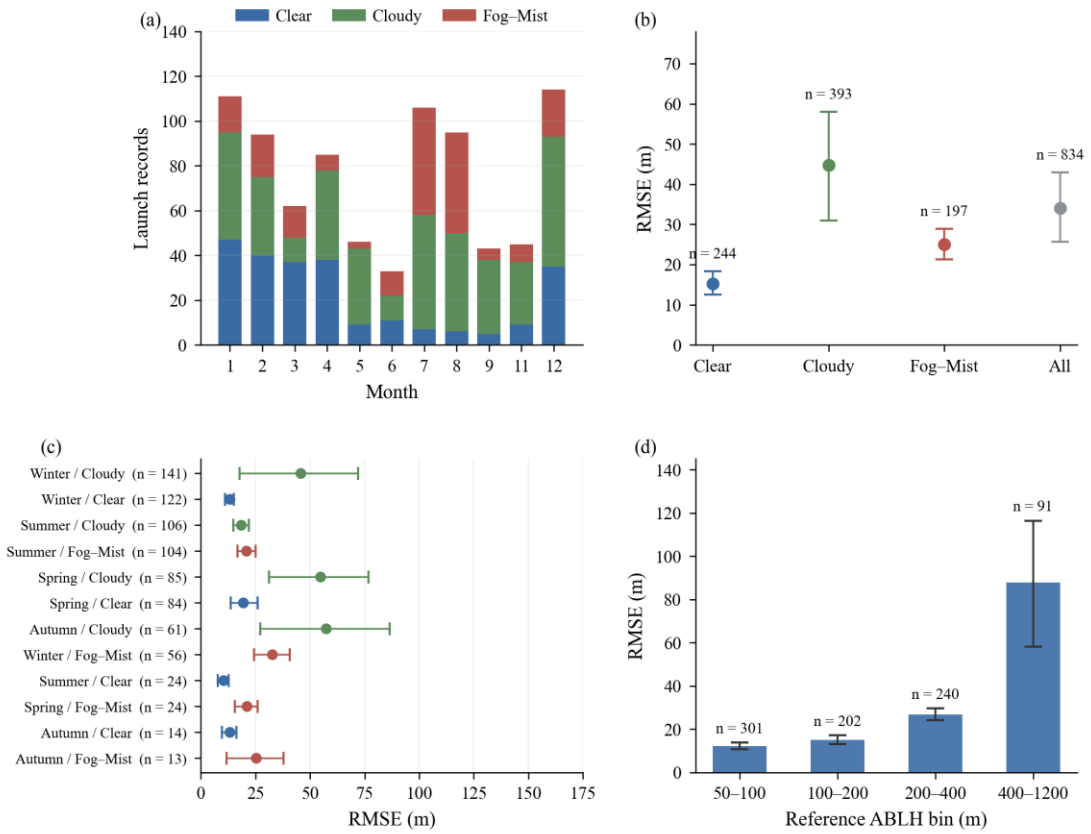


Figure S2. Sample size and the uncertainty of the launch-time evaluation. (a) Number of launch records per month, separated by sky condition. (b) RMSE with 95 % bootstrap confidence intervals for each sky condition and for all records. (c) RMSE with 95 % bootstrap confidence intervals for each season and sky condition, sorted by sample size. (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.

Response to Comment #8: Major Comment 8

Reviewer Comment:

Major Comment 8. Line 342–354: The discussion in this section is largely well-known characteristics and drivers of Arctic ABL seasonal variability which have been discovered/supported by many previous studies. Please add more commentary to the discussion about how your study specifically enhances/supports what has already been discovered. You should include citations throughout this paragraph for the previous studies that already present these phenomena.

Response:

Thank you for your suggestions. We have strengthened the connection between the CMNN-derived patterns and previous Arctic ABL studies. The contribution of this analysis is to show how the multi-source estimator represents previously documented seasonal and sky-condition-dependent variability, complementing the pointwise performance assessment in Sect. 4.1. In the model output, the May mean ABLH reaches 356.7 m, with lower monthly means in January (129.9 m) and July (140.6 m); the May clear-sky and cloudy means are 149.5 and 412.3 m, respectively. These results quantify the seasonal and conditional patterns expressed by the estimator and provide a basis for interpreting its behaviour in the context of existing MOSAiC

observations. The model–reference comparison adds a further assessment of temporal structure: the Doppler-lidar configuration reproduces the reference diurnal cycles with correlations of 0.92 in winter, 0.92 in spring and 0.97 in summer. Together, these analyses show which documented patterns are represented by the model and where its estimates depart from the reference, across the seasonal cycles.

The revised discussion attributes the physical explanations to the relevant literature. The seasonal paragraph cites Vihma et al. (2014), Intrieri et al. (2002) and Tjernström et al. (2019), and explicitly relates the patterns in the CMNN estimates to Jozef et al. (2023a, 2024) and Peng et al. (2023). The sky-condition comparison cites Jozef et al. (2024) and Peng et al. (2023) for the cloud-radiative interpretation, while recognising that wind and air-mass differences may also contribute. The diurnal analysis follows the seasonal definition of Shupe et al. (2022). This framing connects the quantitative model results to established physical understanding: the seasonal and conditional analyses assess the consistency and limitations of the model output, while the reference comparison assesses its representation of the reference time structure. These analyses complement the pointwise error metrics by showing how the CMNN estimates represent seasonal and sky-condition-dependent ABL variability and how their temporal structure compares with the reference.

The specific revisions are provided below.

Item 1 (lines 428–439).

Sections 4.2 and 4.3 examine seasonal and regime-dependent patterns in CMNN-estimated ABLH and their consistency with previous MOSAiC studies. The CMNN-estimated ABLH shows marked seasonal variability during MOSAiC, reaching a monthly mean of 356.7 m in May and its lowest monthly means in January (129.9 m) and July (140.6 m) (Figure 9). The monthly distributions are computed from the 10-min estimates and contain between 1,040 (June) and 3,537 (December) records per month; the sample counts quoted in Sect. 4.1 refer to the launch-time evaluation set. The low January mean coincides with strong lower-atmospheric stratification, with a mean potential-temperature gradient of 24.4 K km^{-1} over 0–400 m. By May, this gradient decreases to approximately 2.7 K km^{-1}, consistent with reduced suppression of vertical mixing and a deeper ABL. The summer shallowing is consistent with limited surface warming over melting sea ice (Vihma et al., 2014b; Intrieri et al., 2002). Warm, moist air advected over colder sea ice can also favor a shallow stable surface layer, potentially decoupled from an overlying cloud or fog layer, providing a further possible explanation for the low summer ABLH (Tjernström et al., 2019; Vihma et al., 2014b). These seasonal patterns in the CMNN estimates are consistent with previous MOSAiC findings (Jozef et al., 2023a, 2024; Peng et al., 2023).

These seasonal patterns in the CMNN estimates are consistent with previous MOSAiC findings (Jozef et al., 2023a, 2024; Peng et al., 2023).

Item 2 (lines 445–450, 454–455).

Figure 9 also shows differences in monthly mean CMNN-estimated ABLH among sky conditions. Cloudy and fog/haze cases generally have higher mean ABLH than clear-sky cases. The clear–cloudy difference is particularly pronounced in May, with means of 149.5 and 412.3 m, respectively, a difference of approximately 263 m in the monthly mean. Enhanced downward longwave radiation under cloudy conditions can weaken near-surface inversions, providing a possible explanation for this contrast (Jozef et al., 2024; Peng et al., 2023). Wind and air-mass differences may also contribute to this contrast. The fog/haze grouping includes conditions beyond shallow summer advection fog. These contrasts are consistent with the higher estimated ABLH under weaker lower-level stratification and stronger low-level winds.

Item 3 (lines 461–462, 469–470).

The relationship between ABLH and near-surface air temperature varies among seasons (Figure 10). The five periods follow the energy-budget-based definition of Shupe et al. (2022).
The CMNN model also reproduces the diurnal cycle of the reference, with correlations of 0.92 in winter, 0.92 in spring and 0.97 in summer (Figure 10).

Response to Comment #9: Major Comment 9

Reviewer Comment:

Major Comment 9. Figure 10: This figure appears to directly match the formatting of Figure 3 in Jozef et al. 2024 (https://acp.copernicus.org/articles/24/1429/2024/). Since this is not a common figure format, but rather was designed specifically for the Jozef et al. 2024 paper, proper credit needs to be given to Jozef et al. 2024 for the design of this figure. If there are any other figures in this paper which were directly based on unique figures of other papers, please add the proper credit for those as well.

Response:

Thank you for your suggestions. We have explicitly credited the panel layout to Jozef et al. (2024, their Figure 3) in both the main text and the caption of Figure 11 (Figure 10 in the original submission). We adopted this layout because it presents the annual frequencies of the different categories together with their seasonal decomposition in a compact, consistent arrangement, making comparisons across categories and seasons straightforward. This is particularly useful for summarising the joint distribution of stability classes and cloud conditions.
The adaptation concerns the graphical layout. The plotted frequencies and seasonal proportions were calculated independently from the observations and classification criteria used in this study; no plotted values or graphical elements were reproduced from the published figure. Our figure presents the SBL, NBL and CBL classes and the sky and cloud–ABL categories defined in our methods. Thus, the layout provides a useful framework for communicating our own classification statistics, with its design source explicitly acknowledged. We have also reviewed the remaining figures for borrowed designs and identified no additional cases requiring this attribution.
The specific revisions are provided below.

Item 1 (Figure 11 caption).

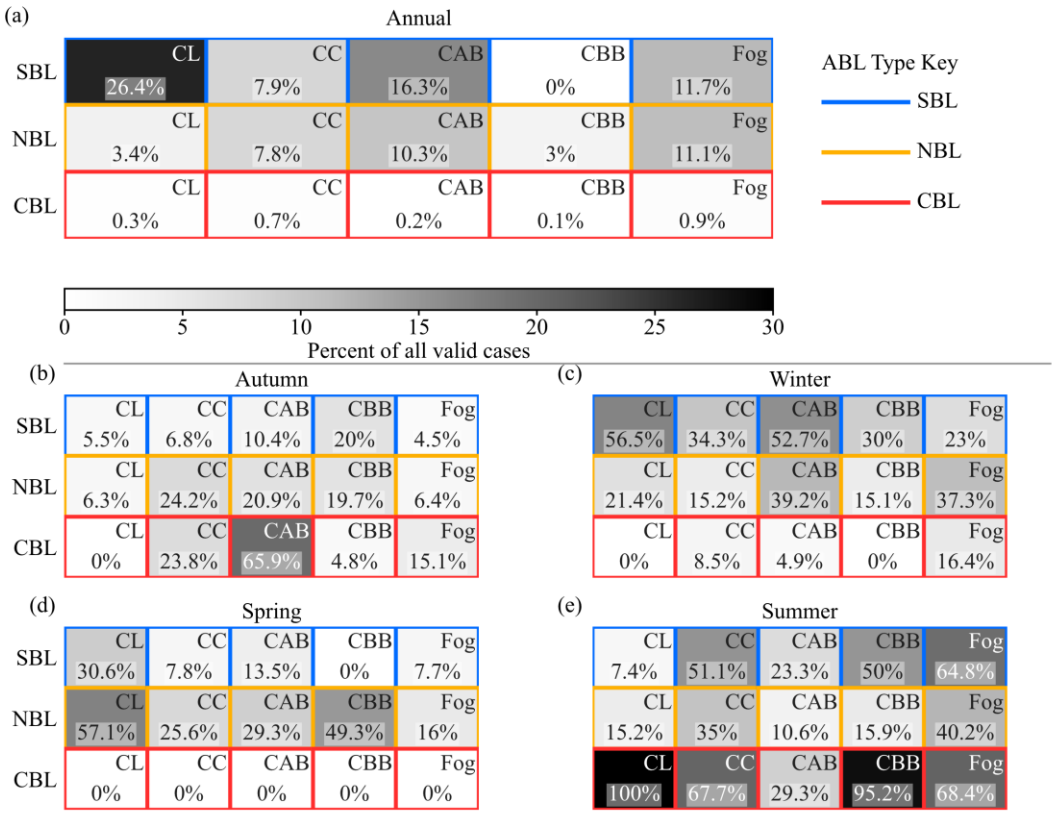


Figure 11. (a) denotes the proportion of each ABL type relative to the total samples. SBL: stable ABL; NBL: near-neutral ABL; CBL: convective ABL. CL: clear-sky conditions; CC: coupled cloud–ABL; CAB: cloud base height more than 100 m above the ABL height; CBB: cloud base height below the ABL height; fog: foggy conditions. The numbers for (b,c,d,e) represent the proportion of each ABL type and weather type occurring in that season relative to the annual total of that type. The classification and the proportions are based on the reference ABLH. The panel layout follows Figure 3 of Jozef et al. (2024), adapted here to the ABL types and sky conditions defined in this study.

Item 2 (lines 472–478).

Records are first grouped into clear-sky, cloudy and fog/haze categories. Cloudy records are then classified by the relative positions of cloud-base height (CBH) and reference ABLH into three groups: coupled (cloud base within 100 m above the ABL top), CBH-above-ABLH (cloud base more than 100 m above the ABL top) and ABLH-above-CBH (cloud base below the ABL top). The 100 m threshold is adapted from the 200 m criterion of Zhang et al. (2024) to account for the shallower Arctic ABL. Following Liu and Liang (2010), the ABL is classified as SBL, NBL or CBL. Figure 11 presents the distribution of these stability classes across the cloud categories, with the panel layout adapted from Jozef et al. (2024, their Figure 3).

Response to Comment #10: Major Comment 10

Reviewer Comment:

Major Comment 10. Line 439–443: This is the first time you truly address the interconnectedness of the INTERPSONDE data used for both the truth labelling and within the model features. This discussion needs to be moved to the methods section and heavily expanded upon. Just because the wind speed profiles provide key controls on ABLH does

not fully justify using the same data for both truth labelling and model features. You need to provide a more compelling reasoning for the decision to use these data within the model features, and/or a discussion on how using these data within the model limits the viability and value that the CMNN model provides. Currently I am not convinced that their use leads to the production of a valuable model (see my major comment 4).

Response:

Thank you for this comment. The methods section now explains the rationale for the INTERPSONDE reference configuration, its relationship to the Ri-based labels, and the role of the Doppler-lidar configuration. Sect. 3.1 explains that the reference configuration is used to assess multi-source estimation performance and input contributions, and serves as the benchmark for evaluating the replacement of sounding winds by tower and Doppler-lidar observations. Sect. 3.2 explains how the shared wind input affects interpretation of the reference results. The Discussion summarises the application scope of the two configurations.

The experiments support these complementary roles. On the same 834 launch-time records, the full-input model reduces RMSE from 46.09 m for wind alone to 34.04 m, quantifying the additional predictive information from the remaining observations and auxiliary features. The Doppler-lidar configuration then tests operation without sounding-derived inputs at inference: on 661 paired launches it achieves an RMSE of 53.5 m and an R^2 of 0.89, compared with 36.6 m and 0.95 for the reference configuration. These results establish both the multi-source gain relative to the wind-only baseline and the performance attainable when tower and lidar observations replace sounding wind. The Ri-based labels are held fixed across the wind-source comparison, providing a consistent diagnostic target. Differences in vertical coverage and measurement characteristics also contribute to the performance change. The revised methods therefore distinguish the benchmark role of sounding wind from the practical use of the estimator with tower and remote-sensing inputs.

The specific revisions are provided below.

Item 1 (lines 182–193, 200–205).

The input features fall into two groups: instrumental observations and auxiliary parameters. The instrumental observations comprise passive brightness temperatures from 22 microwave channels (HATPRO and MiRAC-P) and 50 AERI channels selected by mRMR; an active ceilometer backscatter profile with a vertical resolution of 10 m and 200 levels; and the INTERPSONDE wind profile on 100 levels with 20 m vertical spacing, spanning 10–1990 m. The thermodynamic information enters through the radiances instead. The auxiliary parameters include cloud state, cloud-base height and related single-level descriptors. The model trained with these inputs is hereafter referred to as the reference configuration. INTERPSONDE wind profiles are used both as inputs to the reference configuration and in the diagnosis of the Ri-based reference ABLH, creating shared wind information between the model inputs and reference labels. This dependence may facilitate reproduction of the wind-related relationships embedded in the reference diagnosis, so part of the reported agreement may reflect access to information used to construct the labels. To assess and reduce this dependence, we use input ablation, alternative wind observations and independent-site evaluation to examine the incremental contribution of the other observations, the dependence on shared wind inputs and the transferability of the input–reference ABLH relationship. Second, we train a Doppler-lidar configuration, hereafter the Doppler configuration, in which INTERPSONDE winds are replaced by tower winds at 2, 6 and 10 m and Doppler-lidar winds at 16 range gates between 72 and 502 m. All other inputs, reference labels and the training protocol remain identical. The 10–72 m gap between the tower and lidar observations is filled by interpolation between the two observation sources. The two configurations are compared at the

same launch times with paired predictions to examine whether the model can still reproduce variations in the reference ABLH without directly receiving the INTERPSONDE wind profiles used in its diagnosis (Sect. 3.1).

Item 2 (lines 188–199).

INTERPSONDE wind profiles are used both as inputs to the reference configuration and in the diagnosis of the Ri-based reference ABLH, creating shared wind information between the model inputs and reference labels. This dependence may facilitate reproduction of the wind-related relationships embedded in the reference diagnosis, so part of the reported agreement may reflect access to information used to construct the labels. To assess and reduce this dependence, we use input ablation, alternative wind observations and independent-site evaluation to examine the incremental contribution of the other observations, the dependence on shared wind inputs and the transferability of the input–reference ABLH relationship. First, microwave, AERI, ceilometer backscatter and wind-profile inputs are removed separately while retaining the same reference-label definition and evaluation procedure, and the configurations are compared at identical test times. A wind-only configuration is also compared with the full-input configuration to assess whether the other observations and auxiliary features provide additional predictive constraints beyond wind information. The no-wind experiment further examines how much agreement with the reference ABLH is retained by the remaining inputs without directly supplying the wind profiles used in the Ri diagnosis.

Item 3 (lines 596–610).

Several limitations follow from the design of the framework. The model is trained against a Richardson-number reference and inherits the threshold and vertical resolution of that diagnosis, including the minimum resolvable height. The reference configuration uses sounding-derived wind as an input, and its labels are diagnosed from the same sounding system. The Doppler-lidar configuration reported in Sect. 4.1 uses no sounding-derived inputs at inference time, but both configurations are trained and evaluated against sounding-derived reference labels. The Doppler-lidar configuration is less accurate, and its limited height coverage may reduce information on upper-level wind shear. The Doppler-lidar wind has to be interpolated across the 10–72 m gap between the tower and the first lidar gate. The evaluation is further constrained by sampling: the launch-time records cover four times per day and under-represent periods without lidar observations. The NSA experiment supports transfer of the no-microwave configuration between the two tested sites; evaluating further sites and testing site-specific calibration would establish its wider applicability. A low-level wind observation that resolves the 10–72 m layer (from a short-range lidar, a tethered system or an uncrewed aircraft) would improve wind-profile coverage across this gap. A model trained against labels that are themselves independent of the sounding network, and a temporally blocked evaluation, remain the natural next steps. The choices Ri_c = 0.35, zs = 30 m and a persistence criterion of two consecutive levels on the 20 m grid are empirical. A depth-stratified comparison of the reference labels and the three model configurations with the alternative PANGAEA diagnostic product, derived from the same radiosonde observations, is given in Table S2.

Response to Comment #11: Major Comment 11

Reviewer Comment:

Major Comment 11. Line 451-452: This is a strong statement to make, that I'm not sure is supported. The model retains some cross-site generalization capability, but since the RMSE is greater and R2 value lower (especially for cloudy conditions) for the NSA site versus from the MOSAiC data, one takeaway I get is that a new model should be trained for the NSA site

in order to achieve optimal performance, not that the MOSAiC CMNN model can be used ubiquitously.

Response:

Thank you for your suggestions. We have updated the cross-site evaluation so that the same Ri-based label definition is used at both sites. The no-microwave model, applied to NSA without retraining, achieves an RMSE of 44.00 m and an R^2 of 0.94 on 449 launch-time records, compared with 35.67 m and 0.95 on 834 launch-time records at MOSAiC. Table R11 reports the overall and sky-condition-specific results for this configuration and an additional model trained without either microwave or wind inputs. Within each site, both configurations are evaluated on identical timestamps, reference labels and sky categories. The comparison uses 834 launch-time records at MOSAiC and 449 launch-time records at NSA. At NSA, the model uses the same input types as during MOSAiC, including the radiosonde-interpolated (INTERPSONDE) wind profile, and the reference ABLH is diagnosed from the same interpolated profiles with the Richardson number method. The 449 records arise because the cross-site comparison is restricted to radiosonde launch times, matched to the 10-min estimates within 15 min, rather than using the complete interpolated 10-min series; this keeps the NSA evaluation consistent with the launch-time evaluation at MOSAiC. The condition-specific launch-time samples at NSA are 171 clear-sky, 240 cloudy and 38 fog/haze records.

The sky-condition comparison shows that reference agreement is similar in overall terms but degrades at NSA: cloudy RMSE is 46.48 m at MOSAiC and 48.78 m at NSA, while clear-sky RMSE increases from 15.04 to 34.69 m and fog/haze RMSE from 28.24 to 49.34 m. Removing wind raises the overall RMSE to 78.25 m at MOSAiC and 87.90 m at NSA. Thus, the strong contribution of wind is also evident in the cross-site experiment. The retained overall skill and the comparable response to wind removal support transfer between the two tested sites, while the condition-specific differences identify where further site-specific evaluation is useful. The comparison quantifies transfer performance across the sampling conditions and reference-ABLH distributions of the two sites. Site-specific calibration or retraining can be tested to assess potential further gains.

Table R11. Cross-site comparison of the no-microwave configurations with and without wind. All rows use identical records within each site: MOSAiC uses 834 launch-time records, and NSA uses 449 launch-time records. Cloudy combines ceilometer status 1–3; clear sky is status 0 and fog/haze is status 4. All RMSE values are in metres. NSA evaluation uses the MOSAiC-trained models and normalisation parameters determined during MOSAiC training, without retraining. The additional wind-removal comparison is reported in this response.

Configuration	Sky condition	MOSAiC n	MOSAiC RMSE (m)	MOSAiC R^2	NSA n	NSA RMSE (m)	NSA R^2
No microwave; wind retained	All	834	35.67	0.95	449	44.00	0.94
No microwave; wind retained	Clear sky	244	15.04	0.98	171	34.69	0.91
No microwave; wind retained	Cloudy	393	46.48	0.94	240	48.78	0.89

No microwave; wind retained	Fog/haze	197	28.24	0.95	38	49.34	0.94
No microwave or wind	All	834	78.25	0.76	449	87.90	0.77
No microwave or wind	Clear sky	244	39.55	0.83	171	63.21	0.69
No microwave or wind	Cloudy	393	94.98	0.73	240	95.32	0.77
No microwave or wind	Fog/haze	197	77.40	0.64	38	126.22	0.30

Note: n is the number of evaluation records at each site and R^2 is the coefficient of determination.

The specific revisions are provided below.

Item 1 (lines 541–548).

On the MOSAiC launch-time evaluation set, the no-microwave configuration achieves an RMSE of 35.67 m and an R^2 of 0.95 across 834 launch-time records. RMSE is lowest under clear skies (15.04 m), followed by fog/haze (28.24 m) and cloudy conditions (46.48 m), with corresponding R^2 values of 0.98, 0.95 and 0.94. When applied to NSA without retraining, the same configuration yields an overall RMSE of 44.00 m and an R^2 of 0.94 across 449 launch-time records (Figure 14). At NSA, RMSE is 34.69 m under clear skies, 49.34 m in fog/haze and 48.78 m under cloudy conditions, with corresponding R^2 values of 0.91, 0.94 and 0.89. Both sites use the same Ri-based label definition. Figure S4 compares these NSA results with the MOSAiC configurations by sky condition.

Item 2 (lines 549–553).

The NSA results support cross-site transfer of the no-microwave configuration: the overall R^2 is close to that at MOSAiC (0.94 versus 0.95), although the RMSE is 8.33 m higher (44.00 versus 35.67 m). Cloudy conditions have the highest absolute RMSE at MOSAiC, whereas at NSA the RMSE is highest in fog/haze; relative to MOSAiC, RMSE increases at NSA under all three sky conditions, from 15.04 to 34.69 m under clear skies, from 46.48 to 48.78 m under cloud and from 28.24 to 49.34 m in fog/haze.

Item 3 (lines 553–556).

This condition-dependent pattern identifies where the MOSAiC-trained model retains reference agreement and where its performance changes. The two evaluations differ in temporal sampling and reference-ABLH distributions, so the comparison supports transfer between the tested sites while motivating further evaluation across sites and conditions. Site-specific calibration or retraining can be tested for additional gains.

Item 4 (lines 556–557).

These statistics describe agreement with the diagnostic reference; they are not accuracy statistics against an independently observed ABLH.

Response to Comment #12: Minor Comment 1 (Line 7)

Reviewer Comment:

Line 7: The acronym ABL appears here to stand for "Arctic boundary layer" but should be "atmospheric boundary layer" (I assume this way a typo). Sentence should read "continuous observations if the central Arctic atmospheric boundary layer (ABL)…".

Response:

Thank you for your suggestions. The acronym is now defined explicitly as "atmospheric boundary layer", with the geographical reference as a separate modifier. The revised sentence reads: "The

atmospheric boundary layer (ABL) over the Arctic is governed by stability, clouds, and wind shear, but is poorly constrained because of the scarcity of in situ observations."

We also checked every other use of the acronym. In the Introduction the term is defined correctly ("The atmospheric boundary layer (ABL) is the turbulent layer ..."), and the remaining occurrences of "Arctic boundary layer" in the text are prose descriptions without the abbreviation, so they need no change. The abstract now defines ABL once, correctly, and the abbreviation is used consistently thereafter.

Response to Comment #13: Minor Comment 2 (Line 8)

Reviewer Comment:

Line 8: Write out what MOSAiC stands for.

Response:

Thank you for your suggestions. MOSAiC is spelled out in the abstract, but its first use in the main text relied on the acronym alone, which is what a reader moving from the abstract into the Introduction would notice. The first main-text occurrence (in the opening paragraph of the Introduction, after the revision made for major comment 1) now carries the full name, and the later mentions in the related-work paragraph and in Sect. 2.1 use the acronym.

While checking the surrounding text we also corrected a missing space in "MOSAiC(Shupe et al., 2022)" in Sect. 2.1. Once the first use is expanded, the later references (the section overview, the expedition description and the data section) can use the acronym, so no further change is needed.

Although observations from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) and other field campaigns have substantially advanced knowledge of Arctic ABL structure, …

This study uses atmospheric profiles from Vaisala RS41 radiosondes released from the helicopter deck of R/V Polarstern during MOSAiC (Shupe et al., 2022), between September 2019 and October 2020.

Response to Comment #14: Minor Comment 3 (Lines 9 and 163)

Reviewer Comment:

Line 9: "Retrieve" inherently means that the determined quantity is the truth, however a neural network is only going to predict of infer a value. So I would change the word to "infer" or "predict".

Line 163: Again, I would caution the use of the word "retrieve" as ML models can simply predict or infer values, rather than definitively state the truth. The same comment goes throughout the paper.

Response:

Thank you for your suggestions. We have applied this distinction throughout the paper. A network infers or estimates a quantity; it does not retrieve it in the sense of recovering a measurand, and using "retrieve" for the model output overstates what the method delivers. The verb is now "infer" wherever the network is the subject, the noun is "estimation" where the product of the network is meant, and the adjective is "estimated": for example "CMNN-estimated height" in the caption of Figure 5. The title has been changed accordingly, from "Synergistic retrieval ..." to "Synergistic estimation ...".

We have deliberately kept "retrieval" where it refers to an instrument or a physical product, not to our model, for example the AERI profile retrievals and the reference titles that use the word. Those are standard uses of the term in remote sensing, and replacing them would misdescribe what the instruments do, not clarify what the network does.

The specific revisions are provided below.

Item 1 (manuscript title).

Synergistic estimation of atmospheric boundary layer height from multi-source observations over the Arctic Ocean during MOSAiC

Response to Comment #15: Minor Comment 4 (Line 10, three points)

Reviewer Comment:

Line 10. The end of the sentence "ceilometer backscatter, wind profiles" is missing an "and" before "wind profiles".

Line 10: Spell out RMSE instead of using the acronym in the abstract.

Line 10: How is the truth ABLH value determined, to compare to the neural network prediction? Briefly state this before reporting the RMSE.

Response:

Thank you for your suggestions. We have checked all three points against the current manuscript. The abstract now reads: "... infers ABL height (ABLH) from microwave brightness temperatures, AERI radiances, ceilometer backscatter, wind profiles, and cloud information." The list closes with "and cloud information", so there is no missing connective before the final item.

The abstract also spells out RMSE at first use: "... the network achieves a root-mean-square error (RMSE) of 34.0 m ..." and states how the target is derived: "The reference ABLH used for training and evaluation is diagnosed from Interpolated Sonde (INTERPSONDE) profiles with the Richardson number method."

Response to Comment #16: Minor Comment 5 (Line 19)

Reviewer Comment:

Line 19: The portion of this sentence "and downward longwave radiation also contribute" is grammatically inconsistent with the rest of the sentence. Please revise the sentence to be grammatically correct. Also, it should be "contributes".

Response:

Thank you for your suggestions. The original clause was grammatically inconsistent. The sentence has been removed, and the momentum discussion now states each relationship in a separate sentence with a clear subject. The revised text reads: "The 10-m wind speed, wind stress and friction velocity are positively correlated with the CMNN-estimated ABLH, with Pearson coefficients of 0.732, 0.727 and 0.713, respectively. These variables describe related aspects of near-surface flow and momentum exchange, and their correlations are consistent with a role for mechanically generated turbulence in maintaining deeper ABLs." This removes the choice between "contribute" and "contributes".

Response to Comment #17: Minor Comment 6 (Line 20)

Reviewer Comment:

Line 20: Does the model work well on the other Arctic sites, confirming the generalizability? If so, state this here. Otherwise, the statement is not substantiated.

Response:

Thank you for your suggestions. The abstract now reports the cross-site evidence directly: without retraining, the no-microwave model achieves an R^2 of 0.94 at NSA, compared with 0.95 at MOSAiC (RMSE 44.00 versus 35.67 m). Sect. 4.4 reports the corresponding values, the common Ri-based label definition and the condition-specific results. This supports transfer between the two tested Arctic sites, with RMSE higher at NSA under all three sky conditions.

Without retraining, the no-microwave configuration reaches an R^2 of 0.94 at the North Slope of Alaska, compared with 0.95 at MOSAiC (RMSE 44.00 versus 35.67 m), supporting cross-site transfer.

Response to Comment #18: Minor Comment 7 (Line 27)

Reviewer Comment:

Line 27: You should use the acronym ABL instead of "boundary layer" now that you have defined it (and similar comment throughout the paper). Some places in the paper even have "boundary-layer" (with the dash), so you need to improve the consistency on this.

Response:

Thank you for your suggestions. We have made the terminology consistent. ABL is now used after its definition, and the hyphenated adjectival form has been removed. We replaced "boundary-layer height" with "ABL height (ABLH)" at its first use in the main text and with "ABLH" thereafter. We also replaced "boundary-layer processes" with "ABL processes" and changed "shallow boundary layers" and "deep boundary layers" to "shallow ABLs" and "deep ABLs".

The main text generally uses ABL and ABLH after their definitions. Full terms are retained in definitions, titles, references and some descriptive wording in figure captions.

Response to Comment #19: Minor Comment 8 (Line 40)

Reviewer Comment:

Line 40: I don't know that I would go so far as to say that the Arctic ABL is insufficiently characterized, as there have been several previous works that deeply analyze ABL structure and dynamics throughout the MOSAiC year, as well as from other field campaigns, which have made great strides in this area. Sure, there is still more to learn, but a lot of progress has already been made before this study.

Response:

Thank you for your suggestions. This concern is already addressed by our response to Major Comment 1. The sentence that the reviewer refers to, "the thermodynamic and dynamic structures of the Arctic boundary layer remain insufficiently characterized", has been replaced by a formulation that credits the progress made by MOSAiC and earlier campaigns and narrows the open problem to model representation: "Although observations from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) and other field campaigns have substantially advanced knowledge of Arctic ABL structure, challenges remain in representing its variability in numerical weather prediction at high latitudes, as reflected in documented forecast and reanalysis biases (Jozef et al., 2023a, 2024; Xi et al., 2024; Wesslén et al., 2014; Tjernström et al., 2012)."

No further change is needed here, because the replacement is part of the same edit.

Response to Comment #20: Minor Comment 9 (Lines 45-47)

Reviewer Comment:

Lines 45-47: This sentence is counterintuitive. The longwave radiative cooling is what contributes to the persistent temperature-inversion, and both support a primarily stable ABL, especially in the absence of wind shear, however you are using these factors to support a near-neutral state. This does not make sense to me without more context.

Response:

Thank you for your suggestions. The sentence conflated two different mechanisms. Longwave radiative cooling and the inversion it maintains produce the strongly stable cases, not the near-neutral ones. Near-neutral conditions arise when that stratification is weakened by wind shear, or by cloud-induced warming of the surface relative to the air above it.

The sentence now separates the two. The inversion and radiative cooling are described as maintaining the strongly stable regime, and the weakly stable or near-neutral cases are attributed to the processes that reduce the near-surface gradient, with the citations unchanged.

Previous studies show that stable and near-neutral ABLs dominate in the Arctic. Persistent surface-based inversions, maintained by longwave radiative cooling and the absence of solar

heating, favor strongly stable conditions, whereas weakly stable or near-neutral conditions can arise when wind shear or cloud-induced surface warming reduces the near-surface temperature gradient (Tjernström et al., 2019; Jozef et al., 2023a).

Response to Comment #21: Minor Comment 10 (Line 48)

Reviewer Comment:

Line 48: Is this referring to an elevated inversion with its base around 200-300 m? Or a surface-based inversion capped at 200-300 m? Please clarify.

Response:

Thank you for your suggestions. The reviewer's question exposed a mistake in the original sentence, and we are grateful for it. We checked the source: the 200–300 m figure comes from Jozef et al. (2024), and it describes the base of the elevated capping inversion that is commonly present in the near-neutral regime, not a general statistic of Arctic inversion heights. The original sentence both mis-cited the source and mis-described the quantity.

The sentence now states what the source says, with the correct citation: in near-neutral conditions an elevated capping inversion is commonly present, with its base at approximately 200–300 m and its top extending to 1–2 km (Jozef et al., 2024). This also fits the revised logic of the preceding sentence, where near-neutral cases are tied to processes that weaken the near-surface gradient, not to radiative cooling.

In the near-neutral regime, an elevated capping inversion is commonly present, with its base at approximately 200–300 m and its top extending to 1–2 km (Jozef et al., 2024).

Response to Comment #22: Minor Comment 11 (Line 51)

Reviewer Comment:

Line 51: Should it be "air masses transported northward"? Right now the sentence is not grammatically correct.

Response:

Thank you for your suggestions. Yes, the intended meaning is that air masses move northward, so a past participle is required. The sentence has been revised to replace "transport" with "advected", which is both grammatical and the standard meteorological term.

In summer, warm and moist air masses advected northward may pass over a colder sea-ice surface, favoring the development of temperature and moisture inversions and shallow stable ABLs (Rantanen et al., 2022).

Response to Comment #23: Minor Comment 12 (Line 74)

Reviewer Comment:

Line 74: A few tens of meters and less than 100 m in my mind are the same thing, so I wouldn't use these as bounds for a range. Instead, I'd say something like "…often extremely shallow, frequently less than 100 m and sometimes as low as 20 m…".

Response:

Thank you for your suggestions. The two bounds describe the same part of the range, so the sentence did not actually define a range. It now gives an upper bound and a lower bound that are distinct and both observed: frequently less than 100 m, and sometimes as low as 20 m.

Because the Arctic stable ABL is often extremely shallow (frequently less than 100 m and sometimes as low as 20 m), these single-instrument approaches may be insufficient to fully characterize near-surface ABLH (Mahrt, 2014; Tjernström et al., 2012; Illingworth et al., 2019).

Response to Comment #24: Minor Comment 13 (Line 81)

Reviewer Comment:

Line 81: What do you mean by "information content"?

Response:

Thank you for your suggestions. "Information content" was used without explanation. It refers to the number of independent pieces of information a retrieval provides about the vertical profile: in effect, how many distinct layers the measurement can constrain, as opposed to how finely it is sampled. The sentence now says this explicitly, and distinguishes it from vertical resolution, which is the spacing of the reported levels.

Although microwave radiometers can partly penetrate non-precipitating clouds and provide temperature and humidity profiles, the information content of their estimates (the number of independent pieces of information they provide about the vertical profile) and their effective vertical resolution are limited. Consequently, resolving fine vertical structure within shallow ABLs remains difficult (Martinet et al., 2015; Illingworth et al., 2019).

Response to Comment #25: Minor Comment 14 (Line 83)

Reviewer Comment:

Line 83: Wouldn't all the instruments discussed (not just radiosondes) also be constrained by sparse observing stations and lack of long-term observation over drifting sea ice? I don't think these need to be called out particularly for radiosondes when they are true for all. Only the limitation of launch frequency is unique to the radiosondes.

Response:

Thank you for your suggestions. Sparse sites and the absence of long-term records over drifting ice apply to every instrument in that paragraph, so attributing them specifically to radiosondes was misleading. The sentence now attributes to radiosondes only what is unique to them (a few launches per day, and a small number of launch sites), and states the shared limitation once, as a common constraint, not a defect of one instrument.

Radiosondes provide profiles with high vertical resolution, but they are launched only a few times per day and from a small number of sites. More generally, sustained observations over drifting sea ice are difficult for all observing systems, while the limited launch frequency of radiosondes further restricts their temporal sampling. As a result, radiosonde observations alone may not capture rapid spatiotemporal variations in ABLH (Seidel et al., 2010; Vihma et al., 2014b; Brooks et al., 2017).

Response to Comment #26: Minor Comment 15 (Line 94 and Lines 97-100)

Reviewer Comment:

Line 94: "learn complex mappings from high-dimensional data" is a very limited explanation on the value that machine learning can provide. Plus I don't really know what you mean by "mappings": do you mean correlations/relationships? Please explain in more detail how machine learning adds value to this problem.

Lines 97-100: You should include mention here of the Jozef et al. 2024 paper that used self-organizing maps (a machine learning technique) to characterize ABL structure and stability from MOSAiC data. If you have not thoroughly reviewed the Jozef et al. 2024 paper, as well as Jozef et al. 2023 (which uses multimodal observations), you should do so, as they are very closely related to the goals and content of your paper.

Response:

Thank you for your suggestions. "Learn complex mappings" says little. The paragraph now states concretely what the network provides: it can combine many partially informative observations and learn condition-dependent relationships with the target ABLH without imposing a fixed threshold on a single indicator. "Mappings" has been removed, and "relationships" is used where a statistical association is meant.

We have also added the work the reviewer points to. Jozef et al. (2024) applied self-organizing maps (an unsupervised machine-learning technique) to classify the vertical structure of the

MOSAiC ABL into recurring regimes, and Jozef et al. (2023a) combined multiple observation types to examine the thermodynamic and kinematic controls on stability. Both are now cited where machine learning is first discussed, and the distinction to our approach is stated: their learned representations are used to classify structures, whereas here a supervised network is trained to reproduce a diagnosed ABLH.

Multi-source active and passive remote sensing can provide complementary information on thermodynamic state, humidity, cloud and aerosol structure, and winds, offering a basis for continuous ABLH estimation (Kotthaus et al., 2023; Turner and Löhnert, 2014). Because these observations differ in resolution, error characteristics and weather applicability, and because their relationships with ABLH are nonlinear, methods based on a single indicator or fixed thresholds may have limited flexibility for combining them. Machine learning (ML) offers a different route: a network can combine many partially informative observations and learn condition-dependent relationships with the target ABLH. ML has already been applied to MOSAiC ABL observations: Jozef et al. (2024) used self-organizing maps to classify recurring vertical structures. Complementary analyses by Jozef et al. (2023a) examined the thermodynamic and kinematic drivers of ABL stability from multiple observations. Previous ML studies have also improved ABLH estimation (de Arruda Moreira et al., 2022; Guo et al., 2024). For example, Guo et al. (2024) used ML to merge radiosonde, European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) and Global Land Data Assimilation System (GLDAS) data into a 3-hourly global land ABLH dataset. Xi et al. (2024) used random forests to reduce the ERA5 ABLH bias over the Arctic, reporting an index of agreement of up to 0.71 in the test data. Building on these studies, we investigate whether condition-adaptive fusion of microwave, AERI, ceilometer and wind-profile observations can provide continuous estimates of a Richardson-number-based reference ABLH. We quantify estimation performance and dependence on each observation source across clear-sky, cloudy and foggy conditions, and evaluate a variant that replaces sounding-derived wind inputs with tower and Doppler-lidar observations. The seasonal, cloud-relative and LLJ analyses provide context for interpreting the reference-label variability and the model estimates, with comparisons to established MOSAiC findings.

Response to Comment #27: Minor Comment 16 (Lines 113-115)

Reviewer Comment:

Lines 113-115: The ARM INTERPSONDE product should be explained here when it is first mentioned, rather than later in the paragraph after Figure 1. You should also add a link to the INTERPSONDE dataset.

Response:

Thank you for your suggestions. The product was introduced by its acronym and explained only in the following sentence, and no link was given. The first mention now spells out “Interpolated Sonde (INTERPSONDE)” and carries the dataset DOI, and the explanation of what the product does follows immediately.

Because the routine sampling cannot provide a continuous record, the Atmospheric Radiation Measurement (ARM) Interpolated Sonde (INTERPSONDE) product (https://doi.org/10.5439/1095316; Jensen et al., 2021) is used in addition.

Response to Comment #28: Minor Comment 17 (Line 18)

Reviewer Comment:

Line 18: What is the vertical resolution?

Response:

Thank you for your suggestions. The sentence only said “high vertical resolution” without a number. It now states that the radiosondes measure at a vertical spacing of a few metres. For the

ABLH diagnosis, the INTERPSONDE product interpolates these soundings to a 20 m grid, and the analysis uses the 10–1990 m range.

"…they measure temperature, humidity, pressure, wind speed and wind direction with a vertical spacing of a few metres." "It interpolates the MOSAiC soundings to a uniform grid with 1 min temporal and 20 m vertical resolution; the analysis uses the 10–1990 m range."

Response to Comment #29: Minor Comment 18 (Figure 1)

Reviewer Comment:

Figure 1: The caption is too vague. The figure should be able to stand alone, separate from the rest of the text, so you should clearly state what is being plotted (i.e., the locations of all the radiosonde launches) as well as that the starting and ending points represent the beginning and end of the MOSAiC expedition.

Response:

Thank you for your suggestions. The caption has been rewritten so that the figure can stand alone. It states that the symbols are the locations of the radiosonde launches, that the colour encodes the launch date, that circular symbols denote ice-drift phases and triangular symbols transit periods, and that the blue and black markers are the start and end of the expedition.

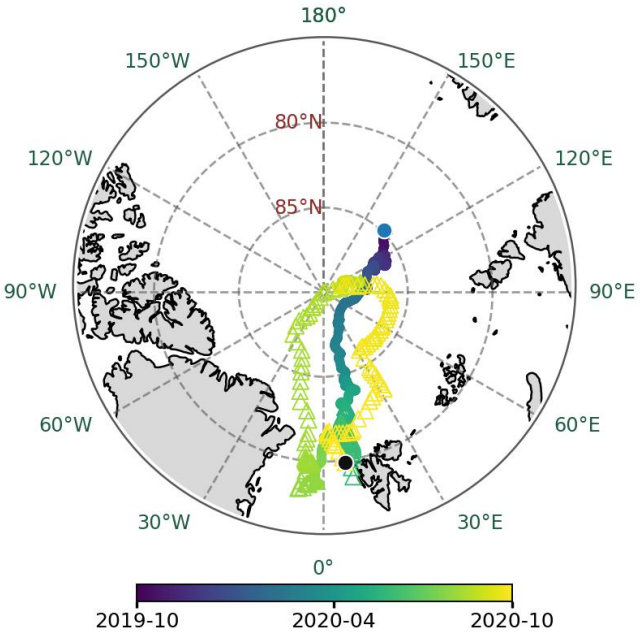


Figure 1. Map of the central Arctic showing the locations of the radiosonde launches during MOSAiC, colour-coded by launch date. Each symbol marks one launch along the drift track of R/V Polarstern; circular symbols indicate ice-drift phases and triangular symbols transit periods. The blue and black markers denote the beginning and the end of the expedition.

Response to Comment #30: Minor Comment 19 (Lines 115-117)

Reviewer Comment:

Lines 115–117: How are you treating the radiosonde profiles that occurred over open ocean during periods of Polarstern transit (e.g., when it had to exit the ice pack between Legs 3 and 4, and at the start and end of the campaign)? These profiles will have different structures than those over the sea ice, and may be missing some of the accompanying data sources that were deployed on the sea ice surface.

Response:

Thank you for your suggestions. These profiles differ, and the surface instruments were absent for some of them. They are retained in the reference series, because excluding them would remove the only profiles from the transit periods and would bias the record toward ice-camp conditions; the text now says so and notes that the corresponding multi-source samples are sparser and few in number.

Profiles launched during transits outside the ice pack are retained in the reference series, but the surface instruments were not deployed during those periods, so the corresponding multi-source samples are sparse.

Response to Comment #31: Minor Comment 20 (Line 127)

Reviewer Comment:

Line 127: What do you mean that it constrains the temp and humidity profiles with the MWRRET data to provide 1 min resolution profiles? Please explain further.

Response:

Between soundings, the interpolation is linear in time at each height, and the relative-humidity profile is scaled by a constant factor so that its integrated water vapour matches the microwave radiometer precipitable water vapor (PWV); temperature is not adjusted by the microwave retrieval.

Response to Comment #32: Minor Comment 21 (Line 129)

Reviewer Comment:

Line 129: How well can the ABLH from the INTERPSONDE be trusted for periods between the radiosonde launches? Do you have any statistics on this?

Response:

Thank you for this comment. We do not have an independent reference ABLH time series between radiosonde launches and therefore cannot directly quantify the accuracy of the INTERPSONDE-derived ABLH during these intervals. The comparison in Sect. 3.2 evaluates differences between diagnostic methods at launch times; it does not measure temporal interpolation error or establish an uncertainty bound for periods between launches. The revised text states this limitation explicitly: “Between launches, no independent reference ABLH is available, so the accuracy of the interpolated ABLH cannot be quantified directly; the launch-time comparison in Sect. 3.2 evaluates differences between diagnostic methods and does not measure temporal interpolation error.”

Response to Comment #33: Minor Comment 22 (Line 130)

Reviewer Comment:

Line 130: As mentioned in the major comments, the dataset at https://doi.pangaea.de/10.1594/PANGAEA.957760 already reports the ABLH for each radiosonde profile. I would strongly suggest you consider using the aforementioned dataset to augment your truth labels to support data consistency across MOSAiC studies.

Response:

Thank you for your suggestions. This suggestion is addressed by our response to Major Comment 2. The Jozef ABLH product is used as an alternative diagnosis rather than as the source of the training labels, which remain the INTERPSONDE/Ri reference. On the 560 launches with all six diagnoses available, our labels are higher than the Jozef diagnosis by 46.3 m on average (study minus Jozef), with RMSE 106.8 m, MAE 54.1 m and r = 0.81; 74.6 % of the launches differ by no more than 50 m, and the median absolute difference is 22.5 m (Sect. 3.2, Figure 3). Figure 3 uses row-minus-column bias, so the Jozef-row/Study-column entry has the opposite sign. Table S2 reports the separate 445-launch subset and is explicitly identified as such.

Response to Comment #34: Minor Comment 23 (Line 130)

Reviewer Comment:

Line 130: What specific Richardson number method did you use? Looks like you explain later on in section 3.2, so you should refer to that section here so the author knows you will be explaining it later.

Response:

Thank you for your suggestions. A cross-reference has been added at the first mention of the Richardson number method, and the method is now named explicitly, so that the reader knows the criterion is defined later.

The temperature, humidity, pressure and wind profiles of INTERPSONDE are used to diagnose the reference ABLH with the bulk Richardson number method introduced in Sect. 3.2.

Response to Comment #35: Minor Comment 24 (Line 131)

Reviewer Comment:

Line 131: What do you mean by stability criteria?

Response:

Thank you for your suggestions. “Stability criteria” was vague and has been removed. The text now specifies the critical value, reference height and persistence requirement: “To determine ABLH from the Ri profile, the critical value must be specified in advance. Following Peng et al. (2023), Ri_c is set to 0.35 and zs to 30 m, and the ABLH is identified from the first pair of consecutive grid levels at which Ri exceeds the critical value (two levels in total on the 20 m grid). The threshold-crossing height is linearly interpolated between the first level of this pair and the valid level immediately below it.”

Response to Comment #36: Minor Comment 25 (Line 140)

Reviewer Comment:

Line 140: I would say “despite that” rather than “because” to start this sentence.

Response:

Thank you for your suggestions. The original sentence used a causal link where a concessive or contrastive one was meant. The causal construction has been removed; the revised sentence no longer says that cloud and fog effects explain why the microwave radiometer can observe. It states directly that the microwave channels carry temperature, water vapour and cloud liquid water information and complement lidar and infrared observations under non-precipitating cloudy conditions.

These channels carry temperature, water vapor and cloud liquid water information and complement lidar and infrared observations under non-precipitating cloudy conditions.

Response to Comment #37: Minor Comment 26 (Line 142)

Reviewer Comment:

Line 142: Remind the reader what information is lost/can’t be measured very well under cloudy conditions.

Response:

Thank you for your suggestions. The text referred to “information loss in optical remote sensing” without saying what is lost. The sentence now names the two losses: lidar and ceilometer backscatter is attenuated, and optically thick clouds make the layer above cloud top inaccessible to AERI. The microwave channels remain sensitive to the integrated thermodynamic state.

These channels carry temperature, water vapor and cloud liquid water information and complement lidar and infrared observations under non-precipitating cloudy conditions. Under these conditions, lidar and ceilometer backscatter is attenuated and optically thick clouds make the layer above cloud top inaccessible to AERI, whereas the microwave channels remain sensitive to the integrated thermodynamic state.

Response to Comment #38: Minor Comment 27 (Lines 162-163)

Reviewer Comment:

Line 162–163: State explicitly what measurements are taken via active and passive remote sensing. When you list specifics such as wind profiles and cloud conditions in the same sentence as the more vague references to active and passive remote sensing, it sounds odd.

Response:

Thank you for your suggestions. The sentence now names the instruments behind each category: the passive observations are the microwave and infrared radiances, the active observation is the lidar–ceilometer backscatter, and the wind profile and the cloud information are listed separately, not folded into either category.

This study develops a condition-adaptive multi-source neural network (CMNN) that estimates ABLH over the central Arctic Ocean from multi-source observations. The passive remote-sensing inputs comprise microwave brightness temperatures and infrared radiances, while the active remote-sensing input consists of ceilometer backscatter profiles. Wind profiles and auxiliary cloud-state information are included as additional inputs. The model adjusts its weighting according to sky condition. Sky conditions are classified into clear-sky, cloudy and fog/haze categories from the ceilometer product, and the resulting categories are used as conditioning inputs to the network.

Response to Comment #39: Minor Comment 28 (Line 167)

Reviewer Comment:

Line 167: What do you mean by auxiliary parameters? Is this single-level information versus vertical profiles? Clarify the difference between what you get from the multi-source observations and from the auxiliary parameters.

Response:

Thank you for your suggestions. The distinction is essentially as the reviewer suspects. The instrumental observations are vertical profiles or multi-channel radiances; the auxiliary parameters are single-level descriptors of the ambient state, specifically cloud state, cloud-base height and related single-level descriptors, and they carry no vertical structure.

The input features fall into two groups: instrumental observations and auxiliary parameters. The instrumental observations comprise passive brightness temperatures from 22 microwave channels (HATPRO and MiRAC-P) and 50 AERI channels selected by mRMR; an active ceilometer backscatter profile with a vertical resolution of 10 m and 200 levels; and the INTERPSONDE wind profile on 100 levels with 20 m vertical spacing. The auxiliary parameters include cloud state, cloud-base height and related single-level descriptors.

Response to Comment #40: Minor Comment 29 (Line 171)

Reviewer Comment:

Line 171: 100 vertical levels between what min and max altitude, and with what resolution? Speaking of resolution, can you provide any information about the resolution of the observations from the other instruments discussed in this paragraph?

Response:

Thank you for your suggestions. The Methods section now gives the vertical coverage and resolution of each observation type. The 100 INTERPSONDE wind levels span 10–1990 m with 20 m spacing, and the ceilometer backscatter profile has 200 levels with 10 m resolution. The AERI spectra are recorded with an unapodized resolution of approximately 0.5 cm^{-1}, and the microwave channels have instrument- and channel-dependent passbands: for MiRAC-P the 183.31 GHz channels listed by Mech et al. (2019, their Table 2) have bandwidths of 0.2–1.0 GHz, while the 243 and 340 GHz window channels have a signal bandwidth of approximately 4 GHz. These

values are now stated in the Methods section together with the channel counts of the two microwave radiometers.

Response to Comment #41: Minor Comment 30 (Lines 177-179)
Reviewer Comment:
How are you sure that this method does not remove some of the fine scale variability that is important to ABL structure? Please discuss.
Response:
Thank you for your suggestions. We retain the percentile-based gradient clipping and cumulative reconstruction because their aggregate effect is small in the lower part of the profile and the preprocessing comparison showed a reduction in model RMSE. To examine the structural effect directly, we compared the log-compressed profiles before and after this step on the 834 launch-time records. Within the lowest 48 range gates, only 0.064 % of adjacent-level differences were clipped, and the mean absolute change in reconstructed log-compressed backscatter was 0.0118, equivalent to 0.39 % of the mean original log-compressed signal. The median absolute change was zero. This indicates that the dominant lower-profile structure is largely retained while extreme gradients are limited.
The clipping acts on adjacent-level differences. Profiles are reconstructed from the lowest gate upward, so clipping at higher gates cannot alter the lower gates. We have revised the Methods to describe the retained procedure accurately and distinguish it from the subsequent central-difference calculation of structural features.
The revised Methods section states: "Negative backscatter values are set to zero, and the profile is compressed as $x' = \ln(1 + x)$. Adjacent-level differences are clipped at their 0.5th and 99.5th percentiles, and the profile is reconstructed by cumulative summation before structural features are calculated. Gradient and curvature features are then computed by central differences."

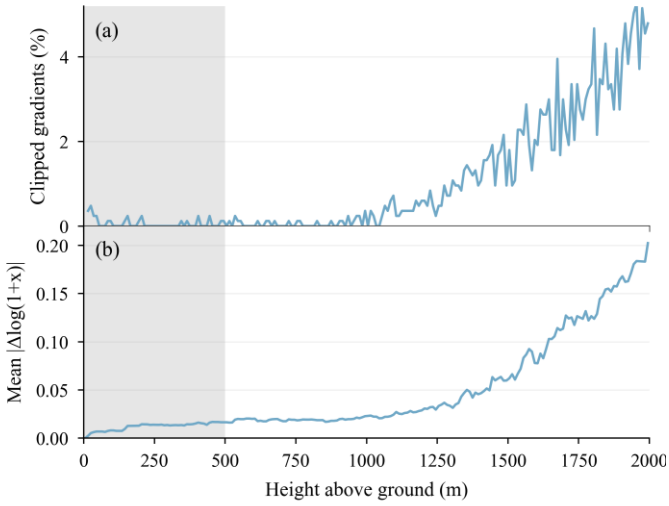


Figure R2. Vertical distribution of the effect of gradient clipping and cumulative reconstruction, obtained from the 834 MOSAiC launch-time profiles. (a) Fraction of adjacent-level differences clipped at each range gate. (b) Mean absolute change in the reconstructed log-compressed backscatter profile relative to the unclipped profile, in log1p units. Grey shading marks 0–500 m, where 0.064 % of adjacent-level differences are clipped and the mean absolute change is 0.012, equivalent to 0.39 % of the mean feature. The effect increases towards the top of the profile.

Response to Comment #42: Minor Comment 31 (Lines 181-183)
Reviewer Comment:

This sentence is hard to follow. Are "gradient, curvature, and characteristic-height information" all applicable to the backscatter profiles? If so, you should split this sentence up more so it is clear that this is separate from the wind profiles, or use semicolons…

Response:

Thank you for your suggestions. Yes: gradient, curvature and characteristic heights all refer to the backscatter profile. The sentence already sets the three groups off with parentheses: wind-profile statistics, backscatter structure, and spectral shape. The three backscatter items are listed inside the "backscatter structure" parenthesis, separate from the wind-profile statistics.

Finally, derived features are added in three groups: wind-profile statistics (speed, direction and vector shear), backscatter structure (vertical gradient, curvature and characteristic heights) and spectral shape of the microwave and infrared radiances.

Response to Comment #43: Minor Comment 32 (Line 187)

Reviewer Comment:

Why do you suddenly use the acronym "BLH" when this has not been used previously in the paper?

Response:

Thank you for your suggestions. That was an inconsistency. The acronym BLH has been removed, and ABLH is used throughout. ABLH is defined at its first use as "ABL height (ABLH)", so no unexplained acronym remains at this point.

Response to Comment #44: Minor Comment 33 (Line 198)

Reviewer Comment:

Using the criteria of four consecutive levels being above the threshold from Jozef et al. 2024 necessitates that your data has similar vertical resolution as the data used in Jozef et al. 2024. Can you confirm this?

Response:

Thank you for this comment. The effective vertical resolution of the INTERPSONDE profiles differs from that of the profiles used by Jozef et al. (2024). In that study, the bulk Richardson number was calculated on a 5 m grid, with the threshold required to remain exceeded over at least 20 m, corresponding to four consecutive 5 m levels, and with a critical value of 0.5. Our primary labels are diagnosed on the 20 m INTERPSONDE grid using $Ri_c = 0.35$, a reference level of $z_s = 30$ m and two consecutive grid levels in total on the 20 m grid. We use the two-level criterion as an operational persistence criterion on the available 20 m grid and describe it as a methodological choice.

We additionally performed a sensitivity comparison using the same Richardson-number profiles interpolated from 20 to 5 m. The interpolation only provides a common diagnostic grid and does not create independent 5 m observations. On this grid, we applied $Ri_c = 0.50$ and the four-level, 20 m persistence criterion used by Jozef et al. The 5 m diagnostic grid contains a level at 30 m; the sensitivity search nevertheless starts at 35 m to match the lowest height at which Jozef et al. retain profiles with trustworthy data. Among 24,730 records for which both diagnoses are available, the mean ABLH is 178.5 m with the primary criterion and 194.4 m with the alternative criterion. The alternative-minus-primary difference has a mean bias of 15.8 m, an MAE of 25.0 m, an RMSE of 38.0 m, a median difference of 13.6 m and a correlation of $r = 0.981$. Excluding the 21.4 % of alternative labels located at the 35 m lower boundary gives a bias of 26.0 m, an MAE of 26.0 m, an RMSE of 38.6 m and $r = 0.985$.

The alternative criterion therefore produces a systematic upward shift in the diagnosed heights but preserves their broad variability. This comparison assesses the joint sensitivity to the diagnostic

grid, the Richardson-number threshold and the persistence criterion; it does not isolate the effect of any individual choice.

Response to Comment #45: Minor Comment 34 (Line 224)

Reviewer Comment:

Line 224: What other observations?

Response:

Thank you for your suggestions. The original sentence referred to unspecified additional observations. The revised sentence names them: radiosondes and lidar. Radiosondes provide in situ thermodynamic profiles and lidar provides the vertical backscatter structure, so AERI spectral radiances are interpreted jointly with these measurements.

The infrared branch encodes AERI hyperspectral radiances. Infrared spectra can provide ABL structural information when combined with radiosondes and lidar (Sawyer and Li, 2013).

Response to Comment #46: Minor Comment 35 (Line 227)

Reviewer Comment:

Why not use the potential temperature profiles from the INTERPSONDE product? Potential temperature is usually one of the more valuable features for determining ABLH.

Response:

Thank you for your suggestions. The labels themselves are diagnosed from those profiles. Using the INTERPSONDE temperature as an input would make the estimation partly a restatement of the label construction, and the same overlap that we discuss for the wind would be much larger. The thermodynamic information enters through the AERI and microwave radiances instead, which are independent measurements of the same atmosphere.

INTERPSONDE temperature and humidity are deliberately not used as inputs, because they are the profiles from which the reference labels are diagnosed; the thermodynamic information enters through the radiances instead.

Response to Comment #47: Minor Comment 36 (Line 234)

Reviewer Comment:

Line 234: What type of weather features are used?

Response:

Thank you for your suggestions. The weather-related inputs are the sky category (clear-sky, cloudy and fog/haze), derived from the ceilometer detection status, and the cloud-state flags that encode cloud presence and cloud-base validity. The sky category forms the condition vector for FiLM; the cloud-state flags are concatenated with the auxiliary features.

Response to Comment #48: Minor Comment 37 (Section 3.3)

Reviewer Comment:

Section 3.3: Do you apply any normalization or standardization to the input data before training begins, so that the model is not impacted by differences in the ranges and max/min values of the different input features? If so, explain what is done. If not, explain why not.

Response:

Thank you for your suggestions. All inputs are standardized with scikit-learn's RobustScaler. Separate scalers are fitted for each input block (HATPRO, MiRAC-P, backscatter, u and v wind, AERI when used, physics features, and labels) on the training set only, and the stored training scalers are applied unchanged to the validation and test sets. For each feature, RobustScaler subtracts the training median and divides by the training interquartile range (75th minus 25th percentile), so differences in the ranges and extrema of the input features do not affect training. The labels are scaled in the same way for training and inverse-transformed for evaluation. No Gaussian noise is added.

All inputs are standardized with scikit-learn's RobustScaler, fitted separately for each input block on the training set only and then applied unchanged to the validation and test sets. For each feature, the scaler subtracts the training median and divides by the training interquartile range.

Response to Comment #49: Minor Comment 38 (Lines 232-240)

Reviewer Comment:

A lot of ML model terminology is used throughout these paragraphs, but without doing a lot of outside research, a reader may not know what each of these layers/steps really does. Can you try to explain some of these layers with plain language…

Response:

Thank you for your suggestions. Plain-language explanations are now given where each component is introduced. The AERI branch uses channel gating and statistical-feature projection because the selected channels are not contiguous in wavenumber. The fusion layer reweights the branch vectors according to sky category. The output heads combine several small predictors with learned weights and add a learned residual offset to reduce remaining systematic bias.

…In practice this layer reweights the branch vectors according to the sky category, so that the branches carrying useful information at that moment dominate the representation.

…The MoE head contains several small predictors and learns when each of them should be trusted, so that a single model can specialize across clear-sky, cloudy and fog/haze conditions. The residual head then adds a small learned offset that removes part of the systematic bias that remains after the first prediction, particularly for shallow ABLs.

Response to Comment #50: Minor Comment 39 (Line 242)

Reviewer Comment:

Why are high boundary layers underestimated? Is it because there are not as many examples of these in the training set?

Response:

Thank you for your suggestions. We checked the actual model-development training split. Of its 17,404 records, 1,502 (8.63 %) have reference ABLH above 400 m, and 707 (4.06 %) reach or exceed 500 m. This confirms that deep ABLs form a relatively small part of the training distribution, which can favour estimates closer to the more densely represented lower heights. Uncertainty in the Ri diagnosis under weak stratification may also contribute. The composite loss increases the weight of underestimated and deep-layer samples to reduce this bias. Performance above 400 m is evaluated on 91 launch-time records, with a 95 % RMSE confidence interval of 57.4–115.8 m. The revised text now uses the training distribution to discuss sample imbalance and the launch-time results to quantify deep-layer performance.

To reduce the systematic underestimation of deep ABLs and the imbalance of samples across weather conditions, the model is trained with a composite loss. The loss combines an asymmetric Huber term, a mean squared error (MSE) term and a relative-error term, so that absolute error, relative error and residual-direction bias are constrained simultaneously, and it assigns larger weights to underestimated samples and to minority weather types and ABLH range bins. A batch-level term further constrains prediction bias and dispersion under each weather condition. Deep ABLs are relatively uncommon in the training set: 1,502 of 17,404 records (8.63 %) exceed 400 m. This imbalance, together with uncertainty in the Ri diagnosis under weak stratification, may contribute to underestimation. Deep-layer samples receive additional training weight, and the 91 evaluation records above 400 m quantify performance in this range.

Response to Comment #51: Minor Comment 40 (Line 247)

Reviewer Comment:

What do you mean by "ABLH height interval"?

Response:

Thank you for your suggestions. It refers to the ABLH ranges used to assign different loss weights: the very-low (≤120 m), low (120–250 m), mid-range (250–500 m) and high (≥500 m) ranges. The text now uses “ABLH range bins”; the earlier “ABLH height interval” has been removed. The revised sentence reads: “… it assigns larger weights to underestimated samples and to minority weather types and ABLH range bins.”

Response to Comment #52: Minor Comment 41 (Line 258)

Reviewer Comment:

I appreciate the fact that you re-iterate the model inputs here in a simple manner. You could, however, introduce this sentence with something like "As a reminder, the model inputs include…"

Response:

Thank you for your suggestions. The sentence now begins with "As a reminder".

As a reminder, the framework integrates microwave radiometer, infrared hyperspectral radiometer, ceilometer, wind profiles, and auxiliary cloud information, with training labels generated by the Richardson number method of Sect. 3.2 (Ri_c = 0.35, a 30 m reference level and a two-level persistence requirement).

Response to Comment #53: Minor Comment 42 (Lines 267-270)

Reviewer Comment:

True model performance can only be determined by the independent validation set… If you are going to list both the test RMSE and the validation RMSE, please explain what you can glean from assessing these against each other: does this point to model overfitting? Or is it within normal statistical variation?

Response:

Thank you for this comment. The training, validation and test sets have different roles. The training set is used to fit the model weights. The validation set (7,459 random records) is used for checkpoint selection and early stopping, and its RMSE quantifies performance during model selection. The test set (834 launch times) was not used for training or model selection and yields RMSE values of 15.27, 44.71 and 25.06 m under clear-sky, cloudy and fog/haze conditions, respectively.

The validation RMSE values are 12.0, 17.8 and 15.0 m, compared with 15.27, 44.71 and 25.06 m for the launch-time test. The difference is largest under cloud. Random validation records and launch-time records sample different evaluation conditions, and differences in composition, temporal dependence and reference-label uncertainty can contribute to the gap. The validation values describe the performance used during model selection, while the launch-time results quantify performance on records withheld from model development.

These complementary results identify cloudy conditions as the main challenge for generalisation within the MOSAiC evaluation. We therefore use the launch-time results as the principal reported test performance and provide the condition-specific sample sizes and uncertainty intervals.

Response to Comment #54: Minor Comment 43 (Line 271-272)

Reviewer Comment:

Have you done some statistical analysis to show that an RMSE of 57.2 m counts as "highly consistent" with the truth? This is a relatively large RMSE, so I don't know if I would claim "highly consistent" for this one unless you have statistics to back it up.

Response:

Thank you for your suggestions. The claim was too strong. The text now states what the statistics support: the correlation exceeds 0.9 in all three categories, but the cloud-condition RMSE (now

44.7 m) is several times the clear-sky value and a substantial fraction of the ABLH itself, so the agreement is good under clear skies and moderate under cloud.

Published Arctic estimates place this error in context. An improved Richardson-number algorithm has a median absolute error of about 29 m against manually identified ABLH, while classical methods range from 50 to 97 m (Peng et al., 2023). MWR-only mixing-layer retrievals differ from lidar by about 340 m (Cimini et al., 2013). In Xi et al. (2024), ERA5 PBLH differs from MOSAiC observations with a centred RMSE of 268 m; their Random Forest correction reduced the centred RMSE to 44 m for training, 94 m for validation and 135 m for test, with $R^2 = 0.6$ and IA = 0.71. RO-derived PBLH differs from radiosondes by 288–691 m depending on season (Chang et al., 2024). The present test RMSE of 34.0 m is of the same order as the algorithm-based estimates and lower than the MWR, reanalysis and RO differences.

The model therefore reproduces the reference well under clear skies but only moderately under cloudy conditions, where the RMSE is higher than under the other conditions and some cases show large discrepancies.

Response to Comment #55: Minor Comment 44 (Lines 284-290)

Reviewer Comment:

It is no surprise that the wind-field-related features… contribute most substantially… as the wind profiles from the INTERPSONDE are also used for the truth labels… Again, it is no surprise that when you take away the data that are used to derive the truth labels, then the model does a lot worse. Please address that this result may be at least partially due to how the data are used, rather than purely feature importance.

Response:

Thank you for your suggestions. The revised text states that the wind dependence reflects both the information carried by the wind profiles and their shared use in constructing the Ri-based reference labels.

One of the inputs is derived from the same sounding system that defines the labels, which bounds how independently the estimation can be validated; the Doppler-lidar configuration reported in Sect. 4.1 reduces, but does not remove, that dependence, and it is less accurate; its limited height coverage may reduce information on upper-level wind shear.

Response to Comment #56: Minor Comment 45 (Lines 302 and 303)

Reviewer Comment:

Line 302: Rather than "frequently used by the model" it would be more accurate to say "considered strongly by the model"…

Line 303: What is "It" that provides constraints? Instead of "It", say what it is that you are referring to.

Response:

Thank you for your suggestions. The conclusion now states: “In the SHAP decomposition, infrared observations contribute most under clear skies, microwave observations provide consistent constraints under moist and cloud-affected conditions, and the ceilometer contributes a smaller share that varies little between conditions.”

Response to Comment #57: Minor Comment 46 (Lines 303-304, 314-315 and 321-322)

Reviewer Comment:

Line 303-304, line 314-315 and line 321-322: In these sentences you provide a somewhat definitive explanation of how the model uses the data, however with ML models you can never really know for sure what the model is getting out of the data, so your assessment can only be an inference. Please soften the language in these statements.

Response:

Thank you for your suggestions. We have softened these statements. The text now presents the SHAP and ablation results as indicative of how strongly the model draws on each source; we replaced “relies on” with “draws on” and avoid definitive claims about internal use. The revised sentence reads: “Taken together, the SHAP and ablation results indicate that CMNN draws on different observation sources under different sky conditions.”

Response to Comment #58: Minor Comment 47 (Line 321)

Reviewer Comment:

Line 321: What does “stable” mean in this context? Is it a stable ABL, or a consistent contribution?

Response:

Thank you for your suggestions. “Stable” here meant a consistent contribution across weather conditions, not a stable ABL. We have replaced it with “consistent”: “their contribution is consistent but not dominant” and “microwave observations provide consistent constraints”.

Response to Comment #59: Minor Comment 48 (Lines 323-324)

Reviewer Comment:

While it is clear that the model is able to consider the multi-source data differently depending on certain ABL conditions… I'm not convinced that the multi-source observations provide any value over the Ri-based radiosonde method… Please add some discussion about how the model provides value under the varying atmospheric/ABL conditions that the Ri-based method does not or cannot provide.

Response:

Thank you for your suggestions. The conclusion describes the model output as a 10-min ABLH estimate and reports condition-dependent errors in Sect. 4.1. The Doppler-lidar configuration additionally evaluates estimation without sounding-derived inputs at inference time, while retaining sounding-derived labels for training and evaluation. This distinction separates the available prediction inputs from the source of the diagnostic reference.

The larger autumn error may partly reflect the limited vertical coverage of the Doppler-lidar winds, which extend only to 502 m and may miss shear associated with deeper ABLs. Reduced availability of upper-level lidar winds under cloudy conditions may also contribute to the performance loss. These differences reflect changes in wind-data coverage and measurement characteristics as well as the removal of the shared INTERPSONDE input. Nevertheless, the Doppler-lidar configuration estimates the Ri-defined reference ABLH without sounding-derived inputs at inference time, supporting its potential use at sites with suitable tower and remote-sensing observations but no routine sounding programme.

Response to Comment #60: Minor Comment 49 (Line 326)

Reviewer Comment:

What does "P1" stand for? Period 1? Please explain.

Response:

Thank you for your suggestions. Yes, P1 denotes period 1. The three periods differ in wind, humidity, near-surface temperature and ABLH: P1 (1–5 May) has weak winds, high relative humidity, low near-surface temperatures and relatively low ABLH; P2 (5–11 May) has increased wind speed, a weaker near-surface temperature gradient, reduced near-surface stability, and more variable ABLH that is often higher than in P1; P3 (11–15 May) has continuing near-surface warming and ABLH generally higher than in P1.

Response to Comment #61: Minor Comment 50 (Line 326 and Figure 7)

Reviewer Comment:

Line 326 and Figure 7: The statement "low-level equivalent potential temperature was relatively low" does not provide much value without comparison to some other value. Is it low in comparison to the values aloft, or something else? I see in Figure 7 you show equivalent potential temperature anomaly in the colorbar, but what is the anomaly with respect to? With no stated comparison for the equivalent potential temperature, it is difficult to interpret and analyze the top panel this figure.

Response:

Thank you for your suggestions. The text now describes the measured variables during P1 directly: "During period 1 (P1, 1–5 May), weak winds and low ABLH estimates were consistent with minimal turbulent mixing, while near-surface temperatures were low and relative humidity was high."

Response to Comment #62: Minor Comment 51 (Line 328)

Reviewer Comment:

Line 328: You comment on "suppressed turbulent mixing" but given the data, there isn't much of a source that would encourage turbulent mixing, so "minimal turbulent mixing" may be more accurate

Response:

Thank you for your suggestions. "Suppressed" has been replaced by a description of minimal turbulent mixing: "During period 1 (P1, 1–5 May), weak winds and low ABLH estimates were consistent with minimal turbulent mixing, while near-surface temperatures were low and relative humidity was high."

Response to Comment #63: Minor Comment 52 (Lines 329-330)

Reviewer Comment:

Lines 329-330: What do you mean thermodynamic conditions improved?

Response:

Thank you for your suggestions. The original phrase has been removed. The text now identifies the wind and temperature-gradient changes: "Wind speed increased, while the near-surface vertical potential-temperature gradient decreased from positive to near-zero or slightly negative values, indicating a transition from stable toward near-neutral or weakly unstable stratification."

Response to Comment #64: Minor Comment 53 (Line 331)

Reviewer Comment:

Line 331: Are you suggesting these short-term jumps are real, or are an artifact of the model?

Response:

Thank you for your suggestions. The network is deterministic at inference, with dropout and training-time noise disabled. Short-term changes can therefore reflect variation in the atmospheric profiles and sensitivity to changing input conditions, including cloud-category transitions; the present comparison does not separate these contributions or establish that each step is a physical change. The revised P2 discussion states this qualification without using events outside that period: "During P2, short-term variability can reflect both changes in the atmospheric profiles and model sensitivity to changing input conditions, including cloud-category transitions."

Response to Comment #65: Minor Comment 54 (Figure 7)

Reviewer Comment:

Figure 7: Can you provide an explanation for the missing data in this figure? Additionally, it would be helpful if you also included a line for the true Ri-based ABLH from the validation set, so the reader can understand how well the model did. If it's too busy to add another line to the top 3 panels, consider adding a 6th panel that shows both the model-predicted ABLH and the true Ri-based ABLH together.

Response:

Thank you for your suggestions. In the revised Figure 8, black dots show the Ri-based ABLH at radiosonde launch times, so the model estimate and the reference can be compared directly; a sixth panel was not added because the overlay keeps the figure compact. The caption now explains the gaps: "Gaps correspond to periods when one or more of the required input observations, such as AERI radiances, were unavailable."

Response to Comment #66: Minor Comment 55 (Lines 350-351)

Reviewer Comment:

In summer, there is also a frequent condition where warm moist air from over the open ocean is advected over the colder sea ice surface, decoupling the lowest part of the atmosphere from the (often foggy) layer above, and creating a shallow, stable ABL. This should be discussed.

Response:

Thank you for your suggestions. The seasonal discussion now includes warm, moist advection as a possible explanation for shallow summer ABLH, with appropriate qualification.

Warm, moist air advected over colder sea ice can also favor a shallow stable surface layer, potentially decoupled from an overlying cloud or fog layer, providing a further possible explanation for the low summer ABLH (Tjernström et al., 2019; Vihma et al., 2014b).

Response to Comment #67: Minor Comment 56 (Figure 8 and Section 4.2)

Reviewer Comment:

Figure 8 and section 4.2: Are the ABLH values in the figure from the Ri-based truth data, or from the CMNN model? Consequently, is your analysis based on the ABLH values from the Ri-based truth data, or from the CMNN model? If you are using the truth data here, this further raises the question as to why the CMNN model-based ABL determination adds value.

Response:

Thank you for your suggestions. Figure 8 in the original manuscript is Figure 9 in the revised manuscript. This monthly-distribution figure and the seasonal analysis in Sect. 4.2 use the CMNN-estimated ABLH. Figure 10 retains the Ri-based reference ABLH for comparison with the estimated diurnal cycles. The opening of Sect. 4.2 now identifies the analysed series explicitly.

Response to Comment #68: Minor Comment 57 (Line 360)

Reviewer Comment:

The ABLH under foggy conditions is often shallow in Arctic summer, due to the advection of warm moist air over the colder sea ice surface… How do you explain a deeper ABLH under foggy conditions in these results? Also, foggy conditions usually do not occur in the Arctic sea ice pack in Winter, so how are you getting foggy examples in Winter? Is it only a few samples? If so, please add discussion on how the limitation in sample size affects the significance of the results.

Response:

Thank you for your suggestions. The fog/haze grouping includes both fog and haze conditions and therefore spans a broader range of atmospheric states than shallow summer advection fog. We have added the group means and sample counts to the revised paragraph. For this descriptive winter comparison, winter is defined as December–February, separately from the energy-budget periods used for the diurnal cycles in Figure 10. Mean CMNN-estimated ABLH is 231.5 m in fog/haze (n = 1,732 records at 1,724 distinct times) and 112.6 m under clear skies (n = 3,883 records at 3,853 distinct times). The mean equivalent-potential-temperature difference between 100 and 50 m is 0.77 and 2.46 K, respectively, while mean wind speed over 10–190 m is 8.75 and 8.31 m s^{-1}. These contrasts are consistent with the higher estimated ABLH under weaker lower-

level stratification and stronger low-level winds. The records are temporally dependent, so the comparison is descriptive and does not treat them as independent events. In summer, warm, moist advection over cold sea ice can instead maintain a shallow stable surface layer beneath fog or cloud, as discussed with reference to Tjernström et al. (2019) and Vihma et al. (2014b). The revised paragraph is reproduced below.

Figure 9 also shows differences in monthly mean CMNN-estimated ABLH among sky conditions. Cloudy and fog/haze cases generally have higher mean ABLH than clear-sky cases. The clear–cloudy difference is particularly pronounced in May, with means of 149.5 and 412.3 m, respectively, a difference of approximately 263 m in the monthly mean. Enhanced downward longwave radiation under cloudy conditions can weaken near-surface inversions, providing a possible explanation for this contrast (Jozef et al., 2024; Peng et al., 2023). Wind and air-mass differences may also contribute to this contrast. The fog/haze grouping includes conditions beyond shallow summer advection fog. For a descriptive winter comparison using December–February records, mean CMNN-estimated ABLH is 231.5 m in fog/haze (n = 1,732 records at 1,724 distinct times) and 112.6 m under clear skies (n = 3,883 records at 3,853 distinct times). The mean equivalent-potential-temperature difference between 100 and 50 m is 0.77 and 2.46 K, respectively, while mean wind speed over 10–190 m is 8.75 and 8.31 m s^{-1}. These contrasts are consistent with the higher estimated ABLH under weaker lower-level stratification and stronger low-level winds. The comparison is descriptive because the records are temporally dependent.

Response to Comment #69: Minor Comment 58 (Lines 367-368)

Reviewer Comment:

Atmospheric convection is very rare in the Arctic, even in summer, as the sea ice surface is not able to heat to the point of supporting buoyant thermals. What you are likely seeing here is a reduction of the temperature gradient as the near-surface atmosphere warms through either downwelling longwave or shortwave radiation, rather than buoyancy from surface heating.

Response:

Thank you for this clarification. The original explanation placed too much emphasis on buoyant convection driven by surface heating over sea ice. We have removed this interpretation. The revised text instead discusses weakened stable stratification as a possible condition allowing existing turbulence to mix over a greater depth. Radiative changes, advection and mechanical mixing provide the physical context for interpreting these cycles. For summer, we now emphasize the constraint on surface warming imposed by melting sea ice.

Response to Comment #70: Minor Comment 59 (Line 371)

Reviewer Comment:

You mention a weak relationship in winter, but the relationship in summer is not much stronger, as the surface temperature in summer is typically locked to the freezing point of the sea ice… Please add more discussion on what drivers during summer could be more important than near surface temperature.

Response:

Thank you for your suggestions. We have expanded the summer discussion to explain why near-surface air temperature provides only a partial account of ABLH variability over melting sea ice. The revised text reads: “In summer, melting constrains the sea-ice surface temperature, limiting surface warming. Near-surface air temperature shows only a weak correspondence with the diurnal variation in ABLH, which may also reflect mechanical mixing and cloud-radiative effects.” This identifies wind-driven mixing and cloud-related radiative changes as additional processes relevant to interpreting the summer estimates. The discussion is framed as an interpretation of the

diurnal behaviour; the correlations over the full analysed record are reported separately in Table S3.

Response to Comment #71: Minor Comment 60 (Line 380)

Reviewer Comment:

Line 380: What do you mean by "weather samples"?

Response:

Thank you for your suggestions. “Weather samples” referred to the individual 10-min records used in the classification, not to weather stations or separate weather cases. We have replaced the phrase with “records”. The revised sentence reads: “Records are first grouped into clear-sky, cloudy and fog/haze categories. Cloudy records are then classified by the relative positions of cloud-base height (CBH) and reference ABLH...” This makes clear that each record is assigned to one sky-condition category before the ABL-type classification.

Response to Comment #72: Minor Comment 61 (Line 389)

Reviewer Comment:

If you have turbulent mixing in the case of a decoupled cloud layer and ABLH, the turbulent mixing is likely driven by wind shear rather than any influence from the clouds. You can use this to support your discussion.

Response:

Thank you for this suggestion. We have clarified that wind shear can sustain turbulence within the ABL when the cloud layer is dynamically decoupled from it. The revised discussion also distinguishes this physical interpretation from the height-based classification, since a cloud base above the estimated ABL top does not by itself establish turbulent decoupling. The revised text reads: “When CBH exceeds ABLH by more than 100 m, the cloud base lies above the estimated ABL top. If the cloud layer is dynamically decoupled from the ABL, wind shear may sustain turbulence below it without direct mixing driven by cloud-top cooling.”

Response to Comment #73: Minor Comment 62 (Lines 399–400 and 406–410)

Reviewer Comment:

These are known phenomena that have been previously published. Please add citations for previous works that have presented these findings.

Response:

Thank you for your suggestions. The cloud-category and stability discussion cites Zhang et al. (2024), Jozef et al. (2024) and Peng et al. (2023). The revised wind-shear statement is conditional: “If the cloud layer is dynamically decoupled from the ABL, wind shear may sustain turbulence below it without direct mixing driven by cloud-top cooling.”

Response to Comment #74: Minor Comment 63 (Line 421)

Reviewer Comment:

What do you mean by "regulating near-surface stability"?

Response:

Thank you for your suggestions. The phrase “regulating near-surface stability” has been removed. The revised discussion states the role of each variable directly: downward longwave radiation and the surface–air temperature difference, Ts − T2m, are positively correlated with the CMNN-estimated ABLH, consistent with reduced surface radiative cooling and weaker near-surface stability; latent heat flux shows a moderate positive correlation and may partly reflect covariation with wind speed and the surface–air humidity difference. The passage reads: “Downward longwave radiation and the surface–air temperature difference, Ts − T2m, show positive correlations with the CMNN-estimated ABLH (r = 0.409 and 0.372, respectively). These

relationships are consistent with reduced surface radiative cooling and weaker near-surface stability..."

Response to Comment #75: Minor Comment 64 (Line 422)

Reviewer Comment:

Please write out what LLJ stands for when it is first mentioned on this line, rather than on the following line.

Response:

Thank you for your suggestions. The acronym is now expanded at its first occurrence in the introduction: "…the dynamical influence of low-level jets (LLJs)." At the definition in Sect. 4.3, the text reads: "An LLJ is identified as a local wind-speed maximum with a decrease of more than 2 m s^{-1} above the core, following Stull (1988)."

Response to Comment #76: Minor Comment 65 (Lines 423-431)

Reviewer Comment:

The relationship between LLJs and ABL structure is another phenomenon that has been previously studied, so please include citations throughout this section where your results agree/disagree with previous studies. Importantly, a thorough analysis of the correlation of LLJ characteristics to ABL stability and structure during MOSAiC was already presented in Jozef et al. 2024. You should compare and contrast the Jozef et al. 2024 results to yours…

Response:

Thank you for your suggestions. We have expanded the comparison with Jozef et al. (2024) to include the relationship between jet-core winds, ABL stability and boundary-layer depth. Their radiosonde analysis found stronger LLJ-core winds and deeper ABLs under weakly stable and near-neutral conditions, bringing the core closer to the ABL top. In the continuous CMNN estimates, mean ABLH is 219.1 m in LLJ cases and 151.1 m in non-LLJ cases; core speed correlates with ABLH at r = 0.684, with a partial correlation of 0.51 after controlling for the below-core minimum wind speed (Table S4). The revised discussion relates these continuous-series statistics to the structural relationship reported by Jozef et al. The occurrence comparison is also retained: 43.6 % of the analysed 10-min records here and approximately 76 % of the 1,377 radiosonde profiles in their study.

Response to Comment #77: Minor Comment 66 (Line 450)

Reviewer Comment:

Line 450: Why do you think you see the greatest degradation under cloudy conditions?

Response:

Thank you for your suggestions. In the evaluation using the same Ri-based label definition at both sites, RMSE increases at NSA under all three sky conditions, from 15.04 to 34.69 m under clear skies, from 46.48 to 48.78 m under cloud and from 28.24 to 49.34 m in fog/haze. The largest absolute increase therefore occurs in fog/haze and clear-sky conditions, and the smallest under cloud. Sect. 4.4 now reports these condition-specific differences and discusses them alongside differences in the observing environment, temporal sampling and ABL structure.

Response to Comment #78: Minor Comment 67 (Lines 452-455)

Reviewer Comment:

I'm not sure I agree with this takeaway. The issue is not that the model may be learning MOSAiC-specific label relationships, as an independent validation set is used for performance evaluation with the MOSAiC data too. The takeaway is rather that for both the NSA and the MOSAiC validation data, the physical relationships used to derive the true ABLH are getting embedded within the CMNN model weights… It's ok that the model

prediction and the truth data are coupled, as long as the model provides some value over the truth data, and that's what I am struggling to see.

Response:

Thank you for your suggestions. Evaluation against Ri-derived labels measures how well the CMNN reproduces that diagnostic definition. The practical contribution is supported by the alternative-input and transfer experiments. In the Doppler-lidar configuration, sounding-derived wind inputs are replaced by tower and Doppler-lidar observations, yielding an RMSE of 53.5 m and an R^2 of 0.89 on 661 paired launches without sounding-derived inputs at inference. The NSA experiment provides a complementary test: the MOSAiC-trained no-microwave configuration is applied without retraining and achieves an RMSE of 44.00 m and an R^2 of 0.94 across 449 launch-time records, compared with 35.67 m and 0.95 at MOSAiC. Both sites use the same Ri-based label definition. The additional no-microwave/no-wind comparison shows a comparable degradation at both sites (Table R11). Together, these results demonstrate estimation from alternative wind observations and retained cross-site reference agreement. The revised discussion reports the sky-condition differences and identifies further site evaluations and calibration experiments as ways to assess wider applicability.

Response to Comment #79: Minor Comment 68 (Line 464)

Reviewer Comment:

What is the "improved" Richardson number method? Refresh the reader on what this means.

Response:

Thank you for your suggestions. We have removed “improved” and specified the settings in the conclusion, with a reference to Sect. 3.2: a critical Richardson number of 0.35, a reference height of 30 m, and a two-level persistence requirement. As described in Sect. 3.2, ABLH is identified from the first pair of consecutive grid levels exceeding Ri = 0.35 (two levels in total on the 20 m grid). The threshold-crossing height is linearly interpolated between the first level of this pair and the valid level immediately below it.

Response to Comment #80: Minor Comment 69 (Line 476)

Reviewer Comment:

It is not just the cloud impacting the ABL, as the ABL may also impact the cloud layer. So maybe instead of saying "cloud impacts" here, you could say "cloud-ABL interactions".

Response:

Thank you for your suggestions. Clouds and the ABL can influence each other. The conclusion now describes ABLH differences among cloud-base categories and relates these contrasts to the diverse cloud–ABL structures documented during MOSAiC.

Response to Comment #81: Minor Comment 70 (Lines 473-484)

Reviewer Comment:

In this section of the conclusion, please state clearly where your results add to or contrast from previous studies, versus where your results agree with previous studies.

Response:

Thank you for your suggestions. The conclusion now identifies the cloud–ABL structures as consistent with Jozef et al. (2024) and Peng et al. (2023), and the associations of ABLH with wind, radiative and thermodynamic variables, together with the higher ABLH in LLJ cases, as agreeing with Jozef et al. (2023a, 2024). The contributions of the present study are the multi-source estimation framework, the evaluation of its performance and input dependence under different sky conditions, the alternative configuration using tower and Doppler-lidar winds, and the transfer experiment at NSA. The conclusion also makes clear that the reported skill measures agreement

with a Ri-based diagnostic reference and discusses the remaining dependence on the sounding system. These contributions extend the estimation and evaluation of ABLH; the physical associations are presented as consistent with established MOSAiC findings.

Response to Comment #82: Minor Comment 71 (Line 483)

Reviewer Comment:

Line 483: This statement is too strong, as there may be other factors at play – instead of "their occurrence increases" I would suggest saying "their occurrence is correlated to an increase of".

Response:

Thank you for your suggestions. The conclusion now describes the difference between LLJ and non-LLJ cases: "Mean CMNN-estimated ABLH is also higher in LLJ cases than in non-LLJ cases (219.1 versus 151.1 m)."

Response to Comment #83: Minor Comment 72 (Line 489)

Reviewer Comment:

Line 489: Did your results actually show this, or is this just something you are predicting?

Response:

Thank you for your suggestions. The conclusion now reports the demonstrated cross-site performance of the no-microwave configuration. The model trained on MOSAiC reproduces the Ri-based reference at NSA, supporting its application across the two sites. Further site-specific evaluation and calibration are identified as next steps.

Response to Comment #84: Minor Comment 73 (Section 5)

Reviewer Comment:

Section 5: What were the limitations of this analysis? What open questions are left after this study? What future steps could be taken to expand upon this work? You should try to end the conclusion with a look toward potential future work, based on the limitations and open questions from this study.

Response:

Thank you for your suggestions. The conclusion now ends with a discussion of the limitations and future work. It identifies the dependence on the Ri-based reference definition, the shared sounding source of the wind inputs and reference labels, the incomplete vertical coverage of the Doppler-lidar configuration, and the sampling limitations of the launch-time evaluation. The remaining questions concern performance against an independent ABLH reference and generalization across time and sites. The proposed next steps include wind observations resolving the 10–72 m layer, training and evaluation using reference labels independent of the sounding system, a temporally blocked evaluation, and further cross-site testing and calibration.

Response to Comment #85: Minor Comment 74 (Original Figure 9; revised Figure 10)

Reviewer Comment:

Figure 9: You need units for Mean Surface Temp in the axis labels. Also, why suddenly use "PBLH" in the legend, which it is "ABLH" throughout the rest of the paper.

Response:

Thank you for your suggestions. We have revised both the figure and its caption. The legend now reads "Ri-based reference ABLH" and "CMNN-estimated ABLH", the unexplained abbreviation PBLH is no longer used, and the temperature series is labelled mean near-surface air temperature with the unit °C on the axis and in the legend. The caption uses the same series names and units, and the revised figure is reproduced in this response letter.