

Overall Impression:

This paper discusses the development and performance of an ML model that uses multiple complementary observational datasets for atmospheric boundary layer height (ABLH) prediction. The paper also discusses ABL characteristics and key drivers, and the seasonal variability of such. While the study introduces a new methodology for multimodal ML-based ABLH inference, I fail to understand how this ML-based methodology adds value over traditional algorithmic ABLH detection methods. I also feel that discussion of previous MOSAiC-specific literature on ABL structure and dynamics is largely lacking. Below I provide several major comments regarding my concerns with the methods and the lack of credit for previous works on this subject, which should be addressed in order for this paper to be suitable for publication.

Major Comments:

1. While the introduction gives a good general background on Arctic ABL structure and dynamics based off 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.
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.

3. 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.
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.
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 INTERPSODNE product, as does the validation set. This leaves me lost as to what you are trying to do with this few sentence section.
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 of 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).
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.
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.
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 off unique figures of other papers, please add the proper credit for those as well.
10. Line 439 – 443: This is the first time you truly address the interconnectedness of the INTERPSODNE 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).

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 R^2 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.

Minor Comments:

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 of the central Arctic atmospheric boundary layer (ABL)...".

Line 8: Write out what MOSAiC stands for.

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

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.

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".

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.

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.

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.

Line 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.

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.

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

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...”.

Line 81: What do you mean by “information content”?

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.

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.

Line 99: Here and in the next several sentences/paragraphs, you start to mention MOSAiC but have given no explanation of what MOSAiC is or what the acronym stands for yet. This should be included before at the first mention of MOSAiC, or any reader unfamiliar with the campaign will miss why it is relevant/important until they get to the next section.

Line 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.

Line 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.

Line 18: What is the vertical resolution?

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.

Line 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.

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.

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?

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.

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.

Line 131: What do you mean by stability criteria?

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

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

Section 2: Add more obvious citations to where each dataset can be found, as all the MOSAiC datasets should be public.

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.

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.

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.

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?

Line 177 – 179: How are you sure that this method does not remove some of the fine scale variability that is important to ABL structure? Please discuss.

Line 181 – 183: 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 to separate the broader categories when commas are used to separate characteristics within categories.

Line 187: Why do you suddenly use the acronym “BLH” when this has not been used previously in the paper?

Line 198: 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?

Line 224: What other observations?

Line 227: Why not use the potential temperature profiles from the INTERPSONDE product? Potential temperature is usually one of the more valuable features for determining ABLH.

Line 234: What type of weather features are used?

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.

Line 232 to 240: 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 to communicate what is really happening at each step? Otherwise, it is hard to trust the methodology as much.

Line 242: Why are high boundary layers underestimated? Is it because there are not as many examples of these in the training set?

Line 247: What do you mean by “ABLH height interval”?

Line 258: 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...”

Line 267 - -270: True model performance can only be determined by the independent validation set that the model has not seen before. 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?

Line 271-272: 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.

284 – 287: It is no surprise that the wind-field-related features (from the INTERPSONDE product, right?) contribute most substantially to the result, as the wind profiles from the INTERPSONDE are also used for the truth labels. This again points to my concern that the same data used for the truth labels are used in the ML model (so then why do you need the ML model?)

Line 287 – 290: Again, it is no surprise that when you take away the data that are used to derive the truth labels (the INTERPSONDE wind fields) from the model input, 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.

Line 302: Rather than “frequently used by the model” it would be more accurate to say “considered strongly by the model” as the model always “uses” all the data, it just weighs the influence of the data differently

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

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.

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

Line 323-324: While it is clear that the model is able to consider the multi-source data differently depending on certain ABL conditions/atmospheric state, I'm not convinced that the multi-source observations provide any value over the Ri-based radiosonde method used for the training and validation sets. 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.

Line 326: What does "P1" stand for? Period 1? Please explain.

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.

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

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

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

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.

Line 350-351: 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.

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.

Line 360: 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, discussed a few comments prior. 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.

Line 367-368: 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.

Line 371: 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, and thus does not change much diurnally. Please add more discussion on what drivers during summer could be more important than near surface temperature.

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.

Figure 9: You include both the actual and the predicted ABLH in the figure, but provide no discussion on the differences. If you are going to provide both lines, you should add some discussion on what you can glean from the different lines. For example, there is a larger gap between the ABLH from the truth and the predicted throughout the full summer diurnal cycle – why?

Line 380: What do you mean by “weather samples”?

Line 389: 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.

Line 399-400 and line 406-410: These are known phenomena that have been previously published. Please add citations for previous works that have presented these findings.

Line 421: What do you mean by “regulating near-surface stability”?

Line 422: Please write out what LLJ stands for when it is first mentioned on this line, rather than on the following line.

Line 423 – 431: 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, and add discussion on how your study adds to what was already presented in Jozef et al. 2024 on LLJs.

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

Line 452-455: 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 (such as "physically meaningful shear and mechanical-mixing information"), so the model prediction and the truth data can never be fully decoupled. 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.

Line 464: What is the "improved" Richardson number method? Refresh the reader on what this means.

Line 476: 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"

Line 473-484: 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.

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".

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

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.