Response to Referees' Comments

Response to Reviewer #1:

This study uses high quality observations to develop a machine-learning-based scheme for predicting land-coupled boundary layer cloud fraction at a single point location in Oklahoma in the United States. The scheme consists of three machine learning models, which are used in tandem to arrive at cloud fraction predictions at each hour of the day between 8 AM and 6 PM local time. Inputs to these models consist of morning radiosonde profiles of relative humidity, potential temperature, and horizontal winds, surface meteorological conditions from the hour preceding and hour coinciding with the prediction time, and predictions from intermediate steps. The models achieve moderate success on this prediction problem, accurately predicting the cloud base, and approximately predicting the cloud top height and cloud fraction at 10 levels between the predicted cloud base and cloud height. The cloud fraction is generally underestimated, and the cloud top height is generally overestimated below roughly 2 km and underestimated above.

The authors then move on to applying their ML models using data from two reanalysis datasets as input, instead of observations. The idea behind this is to illustrate the shortcomings of these reanalysis datasets in simulating boundary layer clouds at this particular site, and use the ML models as a way to estimate whether errors in simulating the clouds can be attributed to errors in predicting underlying meteorological variables versus the errors introduced by the parameterization scheme used. They conclude that errors in ERA5 can be attributed mainly to the cloud parameterization, while errors in MERRA-2 can be attributed to both errors in the meteorological fields and cloud parameterization.

To me the most interesting aspect of this study was the fact that it trained ML models purely on observations. This was facilitated by the uniquely extensive observations taken at the ARM SGP site. This is a strength in one sense in that the ground truth has strong credibility, but it is also a weakness in another in that it limits the applicability of the trained models (and general approach) to a single point location. The parameterization strategy is also less applicable to general circulation models (GCMs), since GCMs simulate vertical profiles of fields at all grid points at every timestep, so (unlike in the case limited by observations) there is no need to temporally separate vertical profiles from surface meteorological quantities. Nevertheless, it is useful to see that a machine learning model can be trained to predict observed boundary layer clouds better than existing physical parameterizations in models used to produce reanalysis data, at least in an isolated setting. For greater impact, a more generalizable model will be key, but that can be saved for discussion of future work. This study could be worth publishing after addressing some comments and questions.

Response: We appreciate the reviewer's detailed and comprehensive feedback on our study. These comments have significantly contributed to improving the clarity of the manuscript. We have carefully considered these comments and concerns raised and have integrated necessary revisions to address the issues related to the model descriptions, the model structure, the application to reanalysis data, and the limitations of the study. All the comments and concerns raised by the referee have been carefully considered and incorporated into this revision. Our detailed responses to the reviewer's questions and comments are listed below.

General comments:

1. *A cleaner and more complete description of the machine learning approach could*

be helpful. For instance I think the feature importance scores in Table 1, which are illustrated in a more interpretable way in Figure 3, could be replaced by some more metadata about the predictor and target fields (e.g. the short names could be accompanied by longer descriptions, including the data source; see e.g. Table 1 in Payami et al., 2024). The network structures I think could be described in the text. In addition some more details about the networks could be provided. What activation functions were used between the layers? What was the optimizer used to train the networks? What were the loss functions? What was the batch size used during training? What was the learning rate?

Response: We are grateful for the reviewer's suggestion for a cleaner and more complete description of our machine learning approach. We have revised Table 1 to include longer descriptions of each predictor and target field, along with their data sources, similar to the format suggested (e.g., Table 1 in Payami et al., 2024).

In addition, we have included a detailed description of the network structures in the text, as follows:

"The DNN architecture is designed, beginning with an input layer reflective of the selected feature set, which includes morning sounding profiles, surface meteorology and heat fluxes data, and the derived variables such as LCL, BLH_{parcel} and BLH_{SH}. For predicting the current hour BLC, the inputs of surface conditions include data both at the current hour and the previous hour. The input variables for training and validating the deep learning model are detailed in Table 1, including variable names, descriptions, and data sources, together with the ARMBE cloud fraction profiles as the learning target for model outputs.

The architecture of the DNN models was structured and tailored for each module: occurrence, cloud-base, and fraction (or fraction-thickness) estimation. Each module's structure is defined by the number of neurons in its hidden layers. For the occurrence module, the structure consists of four hidden layers with 108, 64, 36, and 24 neurons, respectively. The CBH prediction module is similarly structured with four hidden layers, but consisting of 96, 56, 32, and 24 neurons, respectively. The module for predicting cloud fraction and thickness has a slightly simpler structure, with three hidden layers containing 56, 32, and 24 neurons, respectively.

As the specific configuration, we utilized the ReLU (Rectified Linear Unit) activation function to introduce non-linearity into the DNN. L2 regularization with a strength of 0.01 is applied to mitigate overfitting by penalizing large weights and encouraging simpler models. Batch normalization is implemented at each layer to normalize the inputs, ensuring consistent data distribution and stabilizing the learning process. A dropout rate of 0.2 is used to randomly omit neuron connections during training, preventing overfitting and encouraging the network to learn more robust features. The training process was refined with early stopping, ceasing further epochs when the validation loss ceased to improve, and learning rate reduction, systematically decreasing the learning rate upon encountering plateaus in performance improvement. These callbacks were instrumental in honing the model's performance, ensuring convergence to the accurate estimation of the BLC. Neuron biases are included in the network's architecture and systematically inserted in the hidden layers (Battaglia et al. 2018). The model is compiled using the Adam optimizer with an initial learning rate of 0.01. The loss functions used are mean squared error for regression tasks and Binary Cross-Entropy for binary classification tasks. The batch size during training is set to 32. Early stopping with a patience of 37 epochs is implemented to

prevent overfitting and to restore the best weights when the validation loss ceases to improve."

Table 1: Detailed descriptions of input and output variables used in the deep learning models for predicting boundary layer clouds (BLCs). The table includes the variable names, descriptions, and data sources. For the input parameters, surface meteorology and fluxes are taken from the current and previous hours, while morning profiles comprises 46 values spanning from 0-8 km at 06 LT. Note that the output data is derived from ARSCL (Active Remote Sensing of Clouds). The three outputs correspond to the trigger module, cloud-base module, and fraction-thickness module, respectively.

Variable	**Description**	**Data Source**
Input		
Month	Range from 1-12	Time Record
LT	Local Time	Time Record
PS	Pressure at surface level (2m)	Surface Meteorology Station
RH	Relative Humidity at 2m	Surface Meteorology Station
U	Zonal wind at 2m	Surface Meteorology Station
V	Meridional wind at 2m	Surface Meteorology Station
T	Temperature at 2m	Surface Meteorology Station
LCL	Lifted Condensation Level	Derived from T, RH, PS
SH	Sensible Heat	Energy Balance Bowen Ratio
LH	Latent Heat	Energy Balance Bowen Ratio
RH Profile	Morning RH profiles	Radiosonde
U Profile	Morning U wind profiles	Radiosonde
V Profile	Morning V wind profiles	Radiosonde
θ Profile	Morning potential temperature profiles	Radiosonde
$\mathrm{BLH}_{\mathrm{SH}}$	PBLH derived from sensible heat	Derived from θ Profile and SH
$\mathrm{BLH}_{\mathrm{Parcel}}$	PBLH derived from parcel method	Derived from θ Profile and T
Output		
Trigger	Cloud occurrence	ARSCL
Position	Cloud-base height	ARSCL
Fraction Profiles	Cloud fraction and thickness	ARSCL

2. *The structure of the overall model is complicated. In particular the way of separating the predictions of the top and bottom of the cloud layer from the cloud fraction within the cloud layer is unusual (as opposed to simply predicting a cloud fraction at a static set of vertical levels). How was this arrived upon? In addition, how were the inputs chosen? There are many, and to some extent some could be considered redundant. For instance I gather that the BLH_P, BLH_SH, and LCL inputs are derived from fields that overlap in part with other inputs; does omitting those and retraining lead to significant degradation in skill? Also instead of month and local time, could something more physical like insolation be used, which would capture both effects, and be better suited for generalizability?*

Response: We appreciate the reviewer's comments on the structure of the overall model and the choice of inputs. Below, we provide a detailed explanation addressing the three points one by one. The following discussions have been

incorporated into the revised manuscript to clarify the rationale behind the model design and the choice of inputs.

(1) Model Structures and Separation of Predictions

The decision to use three separate models for predicting BLCs, including triggering, cloud position, and cloud fraction, was driven by the need to capture the different aspects of clouds. To characterize clouds comprehensively, it is essential to consider various aspects rather than relying on a single metric. We believe this approach provides a full overview of cloud information.

Firstly, predicting cloud occurrence is a classification problem, distinguishing it from the subsequent tasks that deal with continuous variables. Therefore, the first model focuses exclusively on cloud occurrence, providing a binary outcome that indicates the presence or absence of clouds. This separation ensures that the classification task is handled independently, optimizing the model specifically for this type of prediction. Once cloud occurrence is determined, the next step involves predicting the cloud position. This second model operates on a regression basis, as it deals with the variables representing the vertical position of the cloud. Finally, the third model focuses on the cloud fraction within the established cloud base and top. This model provides a detailed depiction of the cloud's vertical structure by predicting cloud fraction at multiple levels within the clouds.

The key to this strategy lies in the relative independence between cloud position and cloud fraction. BLCs can occur at various positions with different cloud fractions, and the height of the cloud does not necessarily indicate its fraction. By isolating these tasks, the model can accurately determine different aspects of clouds without interference from other predictive tasks. Using three separate models allows for the optimization of each one for its specific task. This modular approach ensures that different aspects of cloud characterization are captured, enhancing the overall reliability of the predictions. Thus, we believe this approach aligns with physical principles and achieves reasonably good performance. To echo these points, we added the following description to the revised Section 3.1:

“The occurrence module, as the first step, evaluates the likelihood of cloud formation by producing a number between 0 and 1, which we call “trigger” in the following, whose value above 0.5 indicates the presence of clouds. The target data for this module is binary (0 or 1), and the model output is a continuous value between 0 and 1. This occurrence information then feeds into the other two modules in parallel: one for locating cloud boundaries and the other for delineating the vertical shape of the cloud fraction in cloudy layers. While the cloud-base (or boundary) module and the fraction-thickness (or fraction) module are independent of each other, they collaborate to depict the vertical cloud fraction profile.

To represent the vertical structure of BLC in the fraction-thickness module, we segmented the cloud layer from the base to the top into ten levels, with each level's thickness varying according to the overall cloud thickness. These values are then interpolated to create a continuous vertical profile of cloud fraction within the BLC boundaries, offering a detailed depiction of the cloud's vertical extent. The vertical position of the layer changes based on the predicted cloud base and top to accurately represent the vertical structure of BLCs. This dynamic approach allows the fraction module to adjust and focus on the relevant portions of cloud

fraction within cloudy layers. Compared to a static level approach, which requires the prediction of cloud fraction across a fixed vertical extent (e.g., multiple levels between 0-6 km), our method focuses on the shape of the fraction profile. This ensures the model is not constrained by fixed vertical levels, allowing for more efficient and robust estimations."

(2) Derived Inputs:

The inputs were chosen based on their relevance to the physical processes governing boundary layer cloud formation and evolution. Although BLH_{Parcel}, BLH_{SH}, and LCL are derived from other inputs rather than direct measurements, they can offer some information for the formation of BLCs, which is why we include these parameters. The results confirm that these parameters are beneficial. As shown in Figure 3, LCL and BLH_{SH} are not very important and only play a minor role. Meanwhile, BLH_{Parcel} demonstrates a notable impact, which is understandable since the PBLH is a critical factor for the formation of BLCs, and BLH_{Parcel} provides a good representation of PBLH.

It is also important to note that BLH_{Parcel} is derived from surface temperature and morning potential temperature profiles, which themselves are significant inputs. The DNN model can adjust the weight of each input by itself, automatically filtering out less important parameters. After this adjustment, BLH_{Parcel} remains an outstanding factor, demonstrating its significance. Thus, we believe it is generally beneficial to include these parameters. Although LCL and BLH_{SH} may not be crucial for the DNN model, their inclusion can still provide some physical constraints to the process. In general, these inputs contribute helpful information that enhances the model's performance.

(3) Use of Time vs. Insolation:

We acknowledge the potential benefits of using more physically meaningful parameters such as insolation rather than proxies like month and local time. Insolation directly reflects the solar radiation received at the surface, which could enhance the model's accuracy and generalizability across different geographical locations. However, using month and local time also has its advantages. These parameters are readily available and are naturally linked to diurnal and seasonal cycles, which affect the characteristics of BLCs. Moreover, they are easy to obtain from any location. While we recognize the value of incorporating insolation in future work, especially for applications over larger regions, the current use of month and local time provides practical and meaningful inputs for our model.

3. *It is acknowledged briefly as future work, but what challenges might be present in trying to apply this approach globally? One aspect that stands out is that we do not have such high-quality detailed observations of clouds and radiosonde profiles everywhere. How would one address that? Data-driven models typically struggle with generalization, so it is unlikely that the model trained for this specific location would be drop-in applicable in other synoptic regions without being exposed to more diverse training data.*

Response: We recognize the limitation of having high-quality, detailed observations only at specific locations like the ARM SGP site. Meanwhile, it should note that the strategy of using ARM sites has several advantages. First, the long-term datasets cover a wide range of scenarios, making it possible to apply the method to other locations with similar meteorological conditions (e.g., mid-

latitude plains). Additionally, ARM sites are part of a global network with extensive coverage, although many sites have limited measurement periods (several months to several years). We recognize the limitation of having high-quality, detailed observations only at specific locations like the ARM SGP site. In the revised manuscript, we discuss potential strategies for addressing this challenge, such as leveraging satellite data, using transfer learning to adapt models trained on one region to others, and integrating data from multiple observational networks to create a more diverse training dataset. We extensively discuss the limitations and potential future strategies as follows:

"Moving forward, future work is warranted to test and extend this diagnostic tool to different synoptic patterns over a large region, which can be integrated into multiple-scale models or reanalysis data. However, several challenges need to be addressed to achieve this. One significant limitation is the lack of high-quality, detailed observations of clouds and radiosonde profiles globally. This scarcity of data can hinder the model's ability to generalize effectively across different regions. To overcome this, there are several potential strategies. First, using transfer learning techniques can help adapt the model trained in one region to other regions with limited data. Integrating data from global observational networks (i.e., ARM) can also create a more diverse and representative training dataset, capturing a wider range of atmospheric conditions and cloud characteristics. Meanwhile, leveraging satellite data can provide broader coverage and enhance the robustness of the model. We plan to explore these approaches in future work to enhance the model's performance and applicability on a global scale."

Specific comments:

Lines 26-28: this sentence is not clear. Should it be something like "Morning meteorological profiles are the initial conditions and then triggers for the formation of BLCs are identified from surface fields."?

Response: We agree with the reviewer’s suggestion. The sentence has been revised for clarity: "The model takes ARM measurements as inputs including early-morning soundings and the diurnal-varying surface meteorological conditions and heat fluxes and predicts hourly estimates as outputs including the determination of cloud occurrence, the positions of cloud boundaries, and the vertical profile of cloud fraction."

Lines 47-48: "These clouds [...] are the critical part for weather prediction and climate modeling [...]." I might switch from "the critical part" to "a critical part," since clouds are not the only important feature to get right for weather or climate modeling.

Response: Per this comment, we have deleted the term “the critical part”.

Line 78: O'Gorman and Dwyer (2018) did not use observational data; they aimed to use ML to merely emulate (rather than improve upon) a convection scheme in an idealized model. Similarly neither did Gentine et al. (2018); they derived an ML parameterization of convection using data from a more expensive super-parameterized simulation. I think Zhang et al. (2021) is the only study cited here that can be said to have used observational data.

Response: We acknowledge the correction and have revised the statement to reflect this:

"Similarly, ML tools have been applied to leverage observational data for the refinement of convection parameterizations, offering more insights into convective triggering (Zhang et al., 2021). In addition, ML has been used to emulate convection schemes and develop parameterizations using data from advanced simulations (O'Gorman and Dwyer, 2018; Gentine et al., 2018). "

Lines 96-98: "By serving as the cloud parameterization in the reanalysis data, this model advanced the capability of low cloud simulations within reanalysis frameworks." I think I get what is being said here, but it is important to emphasize that this is an offline approach, meaning the clouds are predicted based on output data and not embedded in the simulations that produce the reanalysis data itself (thus they cannot affect things like the radiative heating rates and fluxes in the reanalysis data).

Response: We appreciate the clarification. We have revised the sentence to emphasize the offline nature of the approach:

" By serving as an offline diagnostic tool, this model aims to enhance low cloud simulations within reanalysis frameworks without being embedded in the simulations that produce the reanalysis data itself."

Lines 104-109: it might be helpful to emphasize—if I understand correctly—that while ARM SGP takes measurements of some fields at an array of locations across the general SGP region, they only launch radiosondes regularly at this one particular point location, and therefore this study pertains only to that spot. This is quite different than many ML studies which use either data from reanalysis or climate model simulations for training, which is not directly observed (i.e. so can have its own internal biases) but at least is global in nature, without any missing data in time or space. Citing a paper like Sisterson et al. (2016) might be helpful for those who want more historical background on the SGP site.

Response: We cited Sisterson et al. (2016) to offer useful information for the historical background on the SGP site. We also have included additional context to emphasize the specific location of radiosonde launches in Section 2.1: " Note that all the observations are collected at the central facility of SGP, a fixed location, which is different from other ML studies that use global data from reanalysis or climate model simulations (e.g., O'Gorman and Dwyer, 2018; Shamekh et al. 2023).”

Reference:

Sisterson, D. L., Peppler, R. A., Cress, T. S., Lamb, P. J., & Turner, D. D. (2016). The ARM Southern Great Plains (SGP) Site. Meteorological Monographs, 57(1), 6.1-6.14. https://doi.org/10.1175/AMSMONOGRAPHS-D-16-0004.1

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. Proceedings of the National Academy of Sciences, 120(20), e2216158120.

Line 188: "Launched routinely at multiple times daily [...]" Can this be quantified in some way? E.g. approximately how many times per day is it done? Is the important aspect for this study that a radiosonde was launched roughly every morning? Is that at a particular time of day?

Response: We have quantified the radiosonde launches in the revised Section 2.1:

"We take radiosondes (SONDE) measurements around 6 a.m. local time to offer thermodynamic and wind profiles in the PBL and the free atmosphere

(Holdridge et al. (2011) as initial conditions. SONDE launches typically took place four times per day at the SGP site, usually at 00, 06, 12, and 18 local times."

Lines 169-172: it could be helpful to note the purpose of this reanalysis data up front, contrasting it to the purpose of the observational data described earlier. As I understand it, the reanalysis data is mainly used as a way to illustrate how boundary layer clouds are misrepresented in common data sources and as a way to try to disentangle why that might be the case. Unlike the observational data, it is not used in any way to train the ML models.

Response: Following this helpful suggestion, we have clarified the purpose of the reanalysis data upfront in the revised Section 2.3: " Note that unlike observational data aforementioned, reanalysis data are not used for training the deep learning model, instead they are going to be used to help illustrate how the deep learning model may disentangle the potential causes leading to the biased cloud simulations."

Lines 196-197: "models are purpose-built to simulate the initiation, positioning, and vertical extent of BLCs." It might also be worth adding "at the SGP site," since it is likely that these models would likely not be sufficient at other locations given the limitations of the training dataset.

Response: Per the comment, we have added specificity to the sentence: "This study develops an integrated deep learning model to simulate BLC over the SGP site, whose design is illustrated in Figure 1."

Lines 212-216: "To represent the vertical structure of BLC, we equally segmented the cloud layer from the base to the top into ten levels. For each of these levels, our deep learning models calculate individual cloud fraction values." So the vertical position of the layers your models calculate cloud fraction for change depending on the cloud base and cloud top? How would the cloud fraction network know what portion of the morning profiles were most relevant to the cloud fraction? Why was this more complicated model architecture chosen instead of simply skipping straight to predicting a cloud fraction at a static set of vertical levels?

Response: We appreciate the reviewer's insightful questions regarding our model architecture. We have added detailed clarification to explain the reasoning behind our model architecture in the revised Section 3.1:

"To represent the vertical structure of BLC in the fraction-thickness module, we segmented the cloud layer from the base to the top into ten levels, with each level's thickness varying according to the overall cloud thickness. These values are then interpolated to create a continuous vertical profile of cloud fraction within the BLC boundaries, offering a detailed depiction of the cloud's vertical extent. The vertical position of the layer changes based on the predicted cloud base and top to accurately represent the vertical structure of BLCs. This dynamic approach allows the fraction module to adjust and focus on the relevant portions of cloud fraction within cloudy layers. Compared to a static level approach, which requires the prediction of cloud fraction across a fixed vertical extent (e.g., multiple levels between 0-6 km), our method focuses on the shape of the fraction profile. This ensures the model is not constrained by fixed vertical levels, allowing for more efficient and robust estimations."

Table 1: why is the trigger value an input to the other two models instead of just using the other two models only when the predicted trigger value is greater than 0.5? If I understand correctly, with the current approach there is no guarantee that the classification statistics presented in Figure 4 will be relevant in the full problem.

Response: In our approach, the trigger value, which indicates the likelihood of cloud occurrence, is used as an input to ensure continuity and coherence between the models. Sometimes, the trigger value hovers around 0.5, indicating uncertainty about the presence of clouds. This situation often corresponds to scenarios like broken clouds or residual clouds, typically associated with relatively small cloud fractions. Incorporating the trigger value as an input for cloud fraction estimation helps the model account for these ambiguous situations, thereby enhancing its ability to estimate cloud fraction. While including the trigger value is particularly beneficial for the cloud fraction model, it does not affect the CBH estimation, as this aspect of cloud properties is handled separately.

Figure 4 demonstrates the classification problem and is related to cloud occurrence prediction. The classification significantly affects the statistical estimation of cloud fraction, as cloud fraction is set to 0 if the trigger is less than 0.5. However, this does not affect the regression tasks for cloud base and top height predictions.

These discussions have been incorporated into the revised Section 3 to provide a clearer understanding of the rationale behind this approach.

Line 227: what is the strength of the L2 regularization?

Response: The strength of the L2 regularization is specified: "L2 regularization with a strength of 0.01 is applied to mitigate overfitting by penalizing large weights and encouraging simpler models."

Lines 244-246: "Additionally we incorporate datasets from 2017-2020 as part of our validation process, specifically focusing on data from the untrained period to assess the model's performance." If I understand correctly, this is your "test" dataset in ML parlance. Therefore I might rephrase this as "Additionally we save data from 2017-2020 for testing, specifically focusing on data from this untrained period to assess the model's performance."

Response: Following the comment, we have rephrased the sentence for clarity: "In addition, we save data from 2017-2020 for testing, specifically focusing on this untrained period to assess the model's performance."

Lines 246-248: "The training and validations are both using the more than 20-year BLC observations, as well as the ARMBE products." I'm not sure I totally follow this sentence, since the previous few sentences describe the training data / validation datasets as coming from 1998 - 2016 (which is less than 20 years) and the test dataset coming from 2017 - 2020 (which is also less than 20 years). In general I'm not sure what this sentence adds, since having data from these various sources for the time periods cited (which, yes, are in aggregate over 20 years) is already implied, so I think it could be removed.

Response: We acknowledge the confusion of this statement and have removed the redundant sentence for clarity.

Lines 285-287: for the morning profiles, which as I understand it are multiple input features each, I take it this permutation was done using all the profile values for a *particular variable at once? This seems reasonable, but might be worth describing in the manuscript.*

Following the reviewer's suggestion, we have added a clarification in the manuscript regarding the permutation of the morning profiles, as follows: "When performing the permutation, we shuffle the entire morning profile for each case without altering the internal order of values within the profile. This approach ensures that while profiles are permuted across different cases, the sequential structure of values within each profile remains intact. This method allows us to assess the importance of the profiles as coherent units, rather than disrupting their vertical structures."

Lines 312-320: it is a bit odd to describe these specific input parameters—and how they were derived—only at the moment when describing feature importance (and after discussing some sample model predictions). It would be better to describe this earlier when describing the structure of the different models, e.g. in Section 3.1.

Response: Per this helpful suggestion, we have moved the description of specific input parameters (i.e., LCL, BLH_{Parcel}, BLH_{SH}) earlier in the revised Section 3.1.

Table 1: I'm not sure I see the value of presenting the precise numerical importance scores in addition to the bar chart in Figure 3 (I find the bar chart more interpretable).

Response: Thanks for pointing out. We have decided to retain only the bar chart in Figure 3. The current Table 1 has been revised as the input and output lists.

Line 338: "to identify and simulate from surface meteorology." Should this also include a reference to the morning radiosonde inputs?

Response: Indeed, we have revised this statement as: "……to identify the BLC trigger using morning meteorological profiles and observed surface meteorology and fluxes."

Line 357: "Table 2 complements the Figure 4" It seems Table 2 is completely redundant with Figure 4. I would probably keep Figure 4, since it includes slightly more information.

Response: We agree with the reviewer and decided to keep Figure 4 and remove Table 2 for clarity.

Figure 4: where are the F1 scores shown? From what I can tell, precision, recall, and accuracy are shown, but not F1 scores. Also I do not think it is important to explicitly show the performance within the training data. What matters most is the performance on the held out test data.

Response: We appreciate the reviewer's observations regarding Figure 4. We acknowledge that Figure 4 does not display the F1 score, and we have deleted any descriptions related to the F1 score to avoid confusion. We also agree that the performance within the training data is not important as the performance on the held-out test data. However, we did not test the performance on the training data. For "the training period", we used a 70% training and 30% validation split to ensure model validation. This is the regular procedure. In addition, we performed an independent test on a separate validation period. This approach demonstrates that the DNN model can be applied to future data over this region without the need for retraining, indicating its potential for generalizability and robustness in

practical applications. This clarification has been incorporated into the revised manuscript.

Lines 361-364: "The table highlights the model's robustness, with overall accuracy rates of 92.3% for the trained period and a slightly reduced but still substantial 89.2% for the untrained period." Given that the datasets are imbalanced (i.e. there are fewer occurrences than non-occurrences) the accuracy is perhaps not the best metric to highlight. The precision and recall are both reasonably high, and might be better to highlight. See discussion in this TensorFlow tutorial regarding classification of imbalanced data, in particular the note about the accuracy metric: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data.

Response: Following this constructive comment, we have highlighted precision and recall instead of accuracy in these descriptions: "Moreover, for the trained period, the model achieved a high precision of 88.1% and a recall of 81.2%. For the independent period, the precision and recall remained reasonably high at 76.9% and 75.6%, respectively, demonstrating the model's effective generalization to new data."

Figure 6: again I think showing the results on the held out dataset alone is standard and sufficient.

Response: As noted in the previous response, the term "training period" does not imply using the same data for both training and validation. Instead, it indicates the timeframe during which we allocated 70% of the data for training and the remaining 30% for validation, following common practice. The independent dataset represents a completely different period used for additional testing, ensuring that our model's performance is robust and generalizable. Therefore, we present results for both the training period (standard method) and the independent period. We have clarified this issue in the revised Section 4.2.

Technical corrections:

Line 28: "offer" -> "offers"
Lines 41-42: "stratiforms and shallow cumuli" -> "stratiform and shallow cumulus types"
Line 53: "of land surface" -> "of the land surface"
Line 57: "simulating the boundary layer clouds" -> "simulating boundary layer clouds"
Line 88: "structural structure" -> "structure"
Line 90: "diurnal-varying" -> "diurnally varying"
Figure 6: "Independant" -> "Independent"
Figure 12: "attribute" -> "attributed"
Line 590: "Sesning" -> "Sensing"

Response: Thanks a lot for pointing out. We have corrected all these grammars and typos as suggested.

Response to Reviewer #2:

A deep-learning algorithm is presented that forecasts Boundary-Layer clouds based on morning-soundings and surface fluxes. The network is trained and validated with observational data of ARM's souther great plains (SGP) site. The skill of the new model is analyzed in terms of cloud triggering, their vertical structure and cloud fraction. An attribution of forecast errors is undertaken based on the network model that illustrates major factors influencing the representation of boundary-layer clouds in the context of this model.

The authors present a technically involved analysis using a broad scope of observational and modeling data which are synthesized into a deep-learning neural-network model to represent boundary-layer clouds over the ARM's Souther Great Plains (SGP) site. The analysis is novel and suits the scope of GMD well. In some aspects, the quality of the manuscript does, however, not hold up to the standards of the journal. This holds for (i) the methodological description (which in the present form would not allow to reproduce the findings), (ii) use of English Language and (iii) contextualization of the work in the scope of existing models (mixed-layer models) for the problem considered. I therefore recommend to the editors to reconsider the manuscript for review after the comments below have been addressed.

Response: We appreciate the reviewer's constructive and comprehensive feedback on our study, which are very helpful for improving the clarity and quality of the manuscript. Specifically, we have provided a clearer and more complete description of the machine-learning procedure, improved the overall readability of the manuscript, and discussed the advantages and limitations of our approach compared to mixed-layer models. All the comments and concerns raised by the referee have been carefully considered and incorporated into this revision. Our detailed responses to the reviewer's questions and comments are listed below.

Cross Review: I agree with the comments raised by anonymous Reviewer #1, in particular I want to subscribe to his first general demanding a cleaner and more complete description of the machine-learning procedure.

Response: We appreciate the reviewer's concurrence with the comments raised by anonymous Reviewer #1 and their emphasis on the need for a cleaner and more complete description of the machine-learning procedure. We have addressed this concern by providing a detailed and comprehensive description in Section 3.1, as outlined below.

Major remarks:

1. *The methodological description of the deep-learning model is rather obscure. Even after carefully reading the manuscript, it remains unclear what exactly is input to the model and what is the output? Does the model forecast an entire day or just a single instance? What is given in terms of the surface parameters – the evolution of fluxes up to the moment of forecast? Or the value at this time? How exactly are the trigger, vertical position and horizontal cloud fraction components of the model related? Shouldn't the vertical position and horizontal fraction constrain each other from the perspective of available humidity? Or does one trigger the other – if so in what direction: does non-zero cloud fraction trigger the vertical positioning or vice versa?*

Response: We are grateful for the reviewer's constructive comments on the need for a clearer methodological description. We have revised the manuscript to

provide a more detailed explanation of the inputs and outputs of the deep-learning model, as well as the relationships between its components. In general, we use hourly input and output at the same time period. Trigger serves as the input for the estimations of cloud positions and fraction, as these parameters are only meaningful under the cloudy conditions. In addition, a non-zero cloud fraction is not necessarily associated with a specific vertical position, as the cloud fraction indicates the proportion of cloud within the predicted cloud boundaries rather than at a fixed vertical grid. Specifically, we have included the following description and Table 1 of input and output lists in the revised Section 3.1, as follows:

"The model is purpose-built to consist of three distinct deep learning modules, each responsible for a critical aspect of the cloud simulation: 1) the determination of the BLC occurrence, 2) the height position of the cloud base, and 3) the cloud thickness and the normalized 10-layered shape of cloud fraction within cloud boundaries, which jointly yield the hourly-averaged vertical structures of BLCs. This modular approach ensures that the estimations are specific for each aspect of the BLCs. Combining cloud thickness and cloud fraction in one module is logical because the thickness for 10-layered clouds varies based on cloud thickness, and thickness is potentially related to the fraction, as thicker clouds are sometimes associated with larger cloud fractions. This separation of tasks enhances the overall reliability and clarity of the model in capturing the various characteristics of BLCs. Note that each of the three deep learning modules is built upon a deep neural network (DNN) with multiple hidden layers.

The occurrence module, as the first step, evaluates the likelihood of cloud formation by producing a number between 0 and 1, which we call "trigger" in the following, whose value above 0.5 indicates the presence of clouds. The target data for this module is binary (0 or 1), and the model output is a continuous value between 0 and 1. This occurrence information then feeds into the other two modules in parallel: one for locating cloud boundaries and the other for delineating the vertical shape of the cloud fraction in cloudy layers. While the cloud-base (or boundary) module and the fraction-thickness (or fraction) module are independent of each other, they collaborate to depict the vertical cloud fraction profile.

To represent the vertical structure of BLC in the fraction-thickness module, we segmented the cloud layer from the base to the top into ten levels, with each level's thickness varying according to the overall cloud thickness. These values are then interpolated to create a continuous vertical profile of cloud fraction within the BLC boundaries, offering a detailed depiction of the cloud's vertical extent. The vertical position of the layer changes based on the predicted cloud base and top to accurately represent the vertical structure of BLCs. This dynamic approach allows the fraction module to adjust and focus on the relevant portions of cloud fraction within cloudy layers. Compared to a static level approach, which requires the prediction of cloud fraction across a fixed vertical extent (e.g., multiple levels between 0-6 km), our method focuses on the shape of the fraction profile. This ensures the model is not constrained by fixed vertical levels, allowing for more efficient and robust estimations."

Table 1: Detailed descriptions of input and output variables used in the deep learning models for predicting boundary layer clouds (BLCs). The table includes the variable names, descriptions, and data sources. For the input parameters, surface meteorology

and fluxes are taken from the current and previous hours, while morning profiles comprises 46 values spanning from 0-8 km at 06 LT. Note that the output data is derived from ARSCL (Active Remote Sensing of Clouds). The three outputs correspond to the trigger module, cloud-base module, and fraction-thickness module, respectively.

Variable	Description	Data Source
Input		
Month	Range from 1-12	Time Record
LT	Local Time	Time Record
PS	Pressure at surface level (2m)	Surface Meteorology Station
RH	Relative Humidity at 2m	Surface Meteorology Station
U	Zonal wind at 2m	Surface Meteorology Station
V	Meridional wind at 2m	Surface Meteorology Station
T	Temperature at 2m	Surface Meteorology Station
LCL	Lifted Condensation Level	Derived from T, RH, PS
SH	Sensible Heat	Energy Balance Bowen Ratio
LH	Latent Heat	Energy Balance Bowen Ratio
RH Profile	Morning RH profiles	Radiosonde
U Profile	Morning U wind profiles	Radiosonde
V Profile	Morning V wind profiles	Radiosonde
θ Profile	Morning potential temperature profiles	Radiosonde
BLH_{SH}	PBLH derived from sensible heat	Derived from θ Profile and SH
BLH_{Parcel}	PBLH derived from parcel method	Derived from θ Profile and T
Output		
Trigger	Cloud occurrence	ARSCL
Position	Cloud-base height	ARSCL
Fraction Profiles	Cloud fraction and thickness	ARSCL

2. *Data Representativity should be assessed and discussed. The entire analysis focuses on a single site. This comes with two major restrictions that need careful consideration: (1) the current findings are constrained to the surface configuration of the SGP site; as such the insight to cloud—surface coupling may be substantially limited. (2) When comparing to reanalysis data, it needs to be taken into account that the reanalysis is representative for a large area – in comparison to the local nature of the DNN model and observation data. What is the local heterogeneity of cloud fields in the area covered by an ERA5 or MERRA grid cell? In other words – to what extent is the current method to be understood us a downscaling rather than an improved parameterization. To a lesser, but possibly non-negligible (?), extent, this also applies to time-locality.*

Response: We appreciate the reviewer's comments on the need to address data representativity and the focus on a single site. We have added a discussion in the manuscript acknowledging that our findings are specific to the SGP site and the need for further studies in different surface environments and synoptic patterns to generalize the findings. Meanwhile, the limitations of the study have been discussed, and we have outlined future work as follows:

" Moving forward, future work is warranted to test and extend this diagnostic tool to different synoptic patterns over a large region, which can be integrated into multiple-scale models or reanalysis data. However, several

challenges need to be addressed to achieve this. One significant limitation is the lack of high-quality, detailed observations of clouds and radiosonde profiles globally. This scarcity of data can hinder the model's ability to generalize effectively across different regions. To overcome this, there are several potential strategies. First, using transfer learning techniques can help adapt the model trained in one region to other regions with limited data. Integrating data from global observational networks (i.e., ARM) can also create a more diverse and representative training dataset, capturing a wider range of atmospheric conditions and cloud characteristics. Meanwhile, leveraging satellite data can provide broader coverage and enhance the robustness of the model. We plan to explore these approaches in future work to enhance the model's performance and applicability on a global scale."

Meanwhile, we acknowledge the local heterogeneity of cloud fields in the area covered by an ERA5 or MERRA grid cell in the revised Section 2.3. This inherent discrepancy between the model and observations arises from the difference between point-based and grid-based measurements. However, we believe it is not a significant issue for this study. The point observation at the SGP site, as a plain region, is frequently used as a benchmark to evaluate cloud coverage and fraction for reanalysis data or climate models, which typically have grid sizes ranging from tens of kilometers to over 100 km (e.g., Song et al., 2014; Zhao et al., 2017; Zheng et al., 2023; Zhang et al., 2017). Thus, we believe it is safe to compare our model to reanalysis data, including ERA-5 with a 0.25-degree grid (~30 km) and MERRA-2 with a 2/3° × 1/2° degree grid. For example, Figure R1 demonstrates the local heterogeneity of cloud fields within the grid cells. The four sites around the central site are located at distances of approximately 30-50 km. Their total cloud fractions are very similar, with differences ranging from 0-7% in terms of climatology. In summary, we believe it is suitable to use the point-based observations to represent the grid-based reanalysis data over this region, similar to other numerous studies. It is also important to note that for meteorological sites in plain regions, the diagnostic of clouds from models may largely differ from observations, which is an inherent limitation that applies to most models.

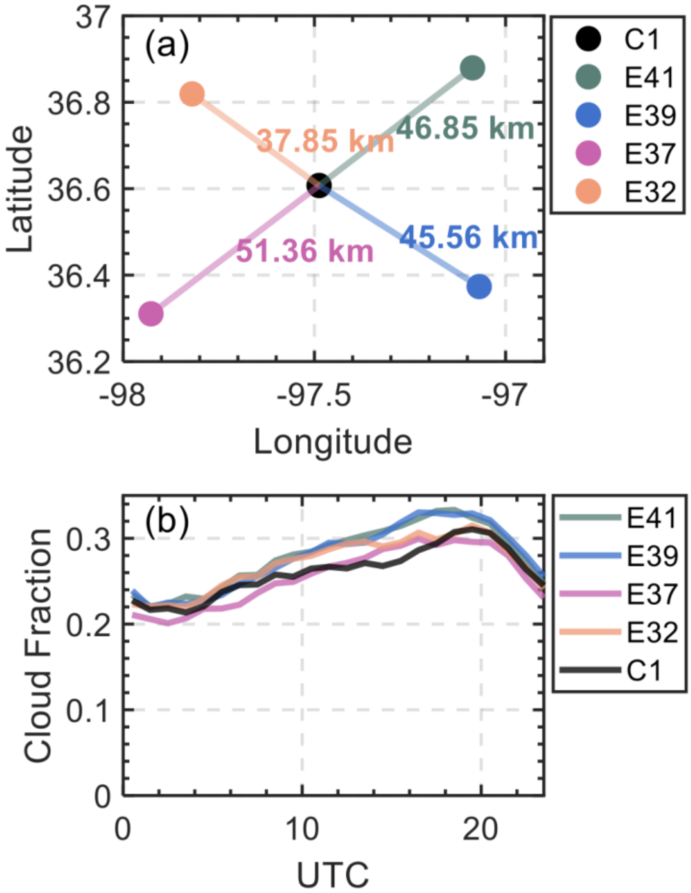

Figure R1. (a) Locations of the five sites, including the central site (C1) and four surrounding sites (E32, E37, E39, E41), with distances between C1 and the other sites indicated. (b) Total cloud fraction for these five sites derived from Doppler lidar, showing variations over time in UTC. The cloud fraction is averaged for the period where all five sites have available data (2016-2023).

Reference:

Song, H., Lin, W., Lin, Y., Wolf, A. B., Donner, L. J., Del Genio, A. D., ... & Liu, Y. (2014). Evaluation of cloud fraction simulated by seven SCMs against the ARM observations at the SGP site. Journal of climate, 27(17), 6698-6719.

Zhao, W., Marchand, R., & Fu, Q. (2017). The diurnal cycle of clouds and precipitation at the ARM SGP site: An atmospheric state-based analysis and error decomposition of a multiscale modeling framework simulation. Journal of Geophysical Research: Atmospheres, 122(24), 13-387.

Zheng, X., Tao, C., Zhang, C., Xie, S., Zhang, Y., Xi, B. and Dong, X., 2023. Assessment of CMIP5 and CMIP6 AMIP simulated clouds and surface shortwave radiation using ARM observations over different climate regions. Journal of Climate, 36(24), pp.8475-8495.

Zhang, L., Dong, X., Kennedy, A., Xi, B. and Li, Z., 2017. Evaluation of NASA GISS post-CMIP5 single column model simulated clouds and precipitation using ARM Southern Great Plains observations. Advances in Atmospheric Sciences, 34, pp.306-320.

3. *Embedment in large-scale models would break the physical consistency of the reanalyzed state. While I agree that there is merit in the (local) DNN cloud representation vs. the large-scale reanalysis, I doubt that the "representation" of*

clouds can be improved by simply using the DNN output in the context of the reanalyses. We should not forget that the reanalysis produces a heavily constrained, physically consistent approximation with the observations. Simply changing the cloud representation would thus, most likely deteriorate many other parameters as it would break the consistency.

Response: We appreciate the reviewer's insightful comment regarding the potential challenges of embedding the DNN cloud representation within reanalysis data. Models or reanalysis data are integral and complex systems. Any modification to a specific parameter may improve that parameter itself but could potentially worsen the performance of others. This is a common issue, not limited to deep learning, but applicable to any notable modifications to models.

In our study, we propose using the DNN as a diagnostic tool for BLCs from both observations and reanalysis data. Regarding the scale issue. As demonstrated in Figure R1, we believe that the local-scale observations in the SGP are representative of reanalysis grids. Thus, the DNN model can improve cloud representation over this site. However, further tests are warranted for its wide application across different sites and regions.

To address the specific concern about breaking the physical consistency of the reanalyzed state, we have revised the manuscript to clarify that our DNN model is intended to complement rather than directly replace existing reanalysis cloud representations. Specifically, we suggest that the DNN outputs can be used in a diagnostic capacity to identify BLCs and help to understand deficiencies in representing BLCs of reanalysis data. This approach allows for targeted improvements in cloud representation without compromising the overall physical consistency of the reanalysis data. Thus, DNN model can provide valuable insights into the local characteristics of BLCs, which can be used to inform the development of more accurate cloud parameterizations in large-scale models for this region. By integrating these insights in a controlled and incremental manner, it is possible to enhance cloud representation while preserving the integrity of these models. These discussions have been incorporated into the revised Section 4.3.

4. *Relation to mixed-layer (single-column) models – what is the added benefit of the rather complex DNN approach in comparison to simple low-order mixed-layer models which also capture the daily evolution of the boundary layer given an initial state and time series of surface fluxes? From a fundamental viewpoint, these models have a number of advantages vs. the DNN as they are (i) a loat cheaper to run, (ii) contain physical reasoning, (iii) are available for many of the large-scale models and thus do not need to be implemented.*

Response: Thanks for pointing out the comparison between the DNN approach and mixed-layer models for representing BLCs. We agree that mixed-layer models have advantages for simulating boundary layer, including lower cost, inherent physical reasoning, and ease of integration with large-scale models (Pelly and Belcher, 2001; Clayson and Chen, 2002; De Roode et al., 2014).

However, our DNN approach offers several distinct benefits that complement these traditional models. Firstly, the DNN can capture complex, nonlinear relationships between various different parameters that may be difficult for several equation to represent accurately. Meanwhile, the DNN model is trained on a large dataset of observational data, allowing it to learn from real-world

examples and potentially identify patterns and relationships that are not explicitly encoded in physical models. This data-driven approach can help improve the accuracy of cloud representation by leveraging the wealth of observational information available. The capability of DNN model is particularly valuable in situations where the interactions between variables are highly complex and not well understood. The advanced capability of the DNN model is demonstrated in this study, showing strong performance in estimating cloud occurrence, position, and fraction.

On the other hand, while mixed-layer models are based on established physical principles and are designed for estimating the boundary layer, they do not address vertical cloud fraction in the same scope as the proposed DNN model. In addition, DNN models generally offer advantages in computational speed compared to physical models. For example, our DNN approach is cost-effective, capable of producing two decades of BLCs with vertical information over the SGP within 1 minute using a single GPU node.

Finally, we propose using the DNN model alongside traditional physical models, not as a replacement. The current DNN model cannot produce detailed cloud microphysics and turbulence information. A hybrid approach can combine the strengths of both methods, leading to a comprehensive and accurate representation of BLCs.

We have revised the manuscript to highlight these points and clarify the complementary role of the DNN approach in the broader context of atmospheric modeling.

Reference:

Clayson, C.A. and Chen, A., 2002. Sensitivity of a coupled single-column model in the tropics to treatment of the interfacial parameterizations. Journal of climate, 15(14), pp.1805-1831.

De Roode, S.R., Siebesma, A.P., Dal Gesso, S., Jonker, H.J., Schalkwijk, J. and Sival, J., 2014. A mixed-layer model study of the stratocumulus response to changes in large-scale conditions. Journal of Advances in Modeling Earth Systems, 6(4), pp.1256-1270.

Pelly, J.L. and Belcher, S.E., 2001. A mixed-layer model of the well-mixed stratocumulus-topped boundary layer. Boundary-layer meteorology, 100, pp.171-187.

5. *Style and use of English language. The manuscript is full of syntactical and grammatical errors, which partly obscures the scientific contents. It needs to be carefully checked by a language editor before it may be re-considered for publication. I will list some recurring errors in my technical comments, but this list is by no means complete.*

Response: We appreciate the reviewer's feedback regarding the English language in the manuscript. We acknowledge that syntactical and grammatical errors can obscure the scientific content and hinder comprehension. We have reviewed the specific errors highlighted in the technical comments and corrected them accordingly. In addition, we have thoroughly revised the manuscript to check grammatical errors and syntactical issues. We believe these revisions have significantly improved the quality of the manuscript, which meets the high standards expected for publication.

Minor remarks:

l. 86: ‘comprehensive data’ – what kind of data? Please use a more telling attribute!

Response: Thank you for pointing this out. We have explicitly described the datasets in the introduction, including radiosondes (SONDE) profiles, surface meteorological measurements, and Active Remote Sensing of Clouds (ARSCL), from the Atmospheric Radiation Measurement (ARM) program at the Southern Great Plains (SGP) site.

l.89 ‘By assimilating morning radiosonde observations’ – the procedure of deep learning is not really an assimilation (please check throughout the manuscript!)

Response: We appreciate your suggestion. We have replaced "assimilating" with "integrating" throughout the manuscript to accurately reflect the process of using morning radiosonde observations in our deep learning model.

l. 91 ‘ [...] uniquely positioned to unravel the complex initiation [...]’ Even after careful assessment of the manuscript, I do not agree on this statement: First, the model is not uniquely positioned as other models exist that can cope with the processes in question (mixed-layer models, LES, etc.); Second, I do not agree that it unravels the initiation and evolution (which would correspond to a causal attribution.)

Response: We have revised this statement to acknowledge that the model is not uniquely positioned, as other models also address these processes. The text now reads: "this deep learning model is capable of diagnosing the initiation and evolution of low clouds".

Paragraph lines 82-92. At the end of this paragraph, it remains unclear what is training, input and output data for the DNN. This should be clarified here, at least qualitatively.

Response: Following this comment, we have added a detailed descriptions for the training, input, and output data for the DNN. In particular, the training data for the deep learning model includes SONDE profiles, surface meteorology, and fluxes, while the outputs are the estimations of cloud occurrence, position, and fraction. More detailed description can be found in the revised Section 2.1 and Section 3.1.

Section 2: “Data and instruments” – the section title does not reflect the contents; it also includes the cloud detection algorithm and a regime classification.

Response: We have revised the section title to "Data, Instruments, and Methodology" to more accurately reflect the contents.

l. 149 Why does CBH need to align with the LIDAR-detected PBL top? A BLC can also be initiated far below the PBL top…

Response: Indeed, we added a clear explanation for the criterion, as follows: "Coupled clouds are identified when the cloud base height (CBH), as derived from the ceilometer, aligns with or is below the lidar-detected PBL top height within 0.2 km, and the calculated surface-based Lifting Condensation Level (LCL, Romps 2017) falls within a maximum allowable range of 0.7 km (Su et al. 2022)."

l. 157/8 “typically lasting more than three hours”. I suppose, this is a threshold for automatic detection, so is it three hours, or not? (“Typically “ is rather confusing here…)

Response: We have revised this statement for clarity. It now reads: "lasting more than three hours," removing the word "typically" to avoid confusion.

l. 153-166 The classification is unclear to me. Is it done per day or per situation? For the stratiform cases, there is a three-hour threshold, but for the cumulus cases, there is no threshold in terms of duration (other than that the clouds emerge after local sunrise). So, how is it possible to characterize days then – these criteria could be evaluated separately for any situation, and certainly there are days in which regime shifts occur.

Response: We appreciate the reviewer's query regarding the classification process. The classification is indeed done per day rather than per individual situation. We recognize that there are situations where transitions between cloud regimes occur, such as the well-known stratiform to cumulus transition. However, in our study, we excluded mixed days and focused solely on days that exhibit purely stratiform or cumulus regimes. This approach ensures that our analysis is clear and specific to each cloud regime without the complexity introduced by transition periods. The revised text now explains the classification criteria more clearly.

Regarding normalization: The target data (cloud trigger, cloud vertical structure and cloud fraction) is already normalized (binary [0,1] , binary vector with elements [0,1] , real vector with real elements from the range [0,1]). Why does normalization need to be applied here? In what sense would it help?

Response: Although normalization is not mandatory, it is a common process used in numerous studies (e.g., Klambauer et al. 2017; Salimans and Kingma, 2016). We apply normalization to both the input and output layers, making them have similar magnitude. This practice offers several benefits: it improves convergence by ensuring features are on a similar scale, stabilizes the training process by preventing issues like exploding or vanishing gradients, and enhances model performance by enabling more consistent weight updates during training. While it is feasible to train without normalization, normalization generally leads to more efficient and effective learning.

Reference:

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. Advances in neural information processing systems, 30.

Salimans, T., & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. Advances in neural information processing systems, 29.

Fig. 1 need to be improved and is incosistent with part of the main text. The RH profiles / morning SONDE is duplicated information and should be within the box entitled "input". Are "profiles" the same as "soundings"? Or is there a difference? LT and MONTH should be combined to the azimuth angle as this is the actual physical parameter that matters and just gets encoded by Month and LT. What is the difference between surface meteorology and surface fluxes? In my understanding, surface fluxes are part of the surface meteorological parameters… The vertical alignment reflecting the geometry of the boundary layer is not appropriate here and confuses. A schematic focusing on input and output would help. If the models for cloud position, cloud fraction and trigger are independent entities, they should be reflected by separate *hidden layers. The tag "cloud position" is potentially confusing here; I suggest to rephrase as "cloud structure" or "vertical position" as position alone might be mistaken for horizontal position. The relation between cloud trigger, cloud position and cloud fraction should be made more clear; If I understand correctly, the trigger is part of the input for the cloud position and cloud fraction models, but this is not appropriately reflected in the schematic. What is meant by time indicators? Are the meteorological parameters / fluxes input as instantaneous values or daily time series? Correspondingly: is the output produced per time instant or rather per day?*

Response: We have revised Figure 1 to address the reviewer's comments and improve consistency with the main text.

- **The input box now includes the morning SONDE, and it is clarified that "morning SONDE" refers to the morning meteorological profiles.**
- **While we agree that using the azimuth angle makes physical sense, local time and month already imply this information, so there is no need to add an extra input parameter.**
- **Surface meteorology typically includes wind, temperature, pressure, and humidity. Although fluxes may be considered part of meteorology, we separate them here for clarity. The detailed input parameters are listed in Table 1.**
- **The idea is to use deep learning as a hidden layer to directly estimate BLCs from surface conditions, rather than resolving the complex PBL processes. Thus, we have marked the PBL in the diagram to indicate that the DNN model bypasses the turbulent and complex PBL processes to directly obtain BLCs from surface conditions.**
- **We have now indicated that there are three separate hidden layers, sharing a similar structure, to reflect the independent entities for cloud position, cloud fraction, and trigger.**
- **The term "cloud position" has been revised to "cloud base" to avoid confusion.**
- **The relationship between the models has been clarified: the trigger value serves as input for the cloud position and cloud fraction models.**
- **The caption now explains that the time indicators refer to local time and month.**
- **It is clarified in the caption that the meteorological parameters and fluxes are instantaneous hourly values.**

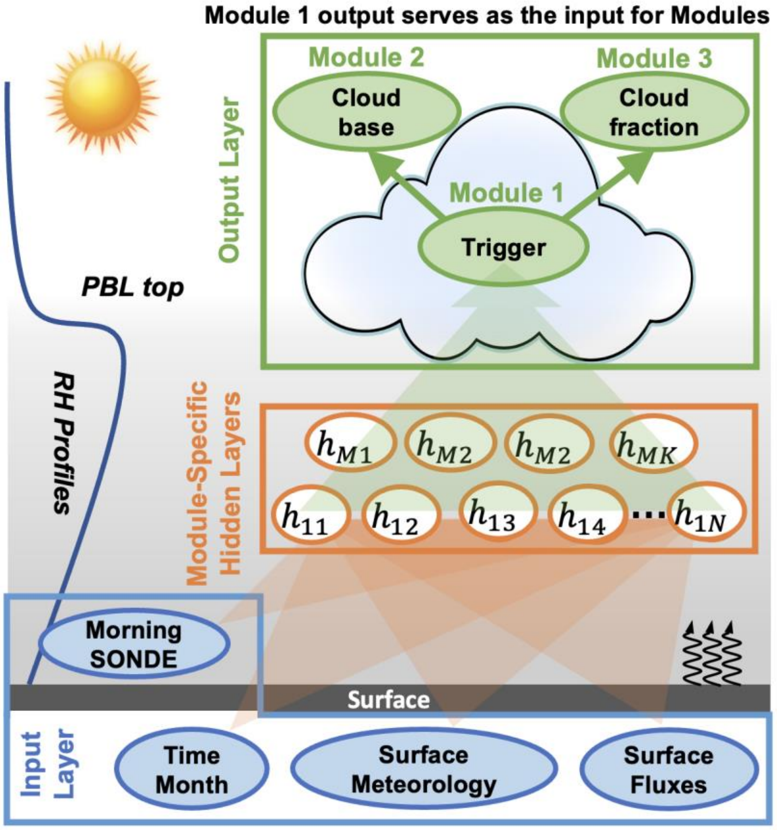

Figure R2 (The revised Figure 1). Conceptual diagram of the deep learning framework for simulating boundary layer cloud (BLC) characteristics over the US Southern Great Plains. Inputs for the deep neural networks (DNN) include morning meteorological profiles from radiosonde (SONDE), time indicators (i.e., local time and month), and surface conditions such as fluxes (curved black arrows) and meteorological data. The relevance of relative humidity (RH) profiles and the planetary boundary layer (PBL) top is emphasized due to their critical role in BLCs development. These variables are processed through multiple layers of hidden neurons (h11 to hMK). Both input and output parameters are hourly, except for the morning SONDE. Separate DNN modules are constructed for each task: Module 1 handles the initiation (trigger) of BLC; Module 2 estimates the cloud base; and Module 3 estimates cloud fraction and thickness. Together, these models synergize to predict the presence, altitude, and stratification of BLC.

Tab. 1 / Text Why is data used if it contributes a negative feature importance?

Thank you for pointing this out. We do filter out the input parameters based on their importance scores, as noted in the original lines 304-305. To clarify, we have extended the description: "After determining the importance scores from the test run, in refining the model, features contributing a negligible or negative effect on performance (i.e., importance scores less than zero) are excluded to ensure only beneficial data is used." In practice, filtering these parameters does not notably affect the results, as the negative values have a magnitude of -0.001.

l. 338/339 "measurements" and "surface meteorology" – what exactly is mean here?

Response: We have revised the text to specify that "measurements" refer to observed surface meteorology and fluxes.

l. 398' "Parameterization" – the DNN model has not parameterization for the cloud top.
Response: We have revised this statement as follow for better accuracy: "This can introduce inherent limitations, as the tops of many clouds may be decoupled from surface influences despite a coupled base, potentially leading to gaps in the DNN's ability to accurately define and estimate the cloud top".

l. 491/ l.559 ' (also compare major point #3) "a more accurate representation" – the DNN is employed here as an offline, a posteriori analysis tool. While you convincingly argue that the DNN has skill to yield a better cloud field, it is misleading to talk about a "better representation" as the DNN is run offline. In fact, the cloud field modified by the DNN is most likely inconsistent with the reanalysis! So, there is no better representation of clouds by just applying the DNN.
Response: We appreciate the reviewer's feedback and have revised the text to clarify that the DNN provides improved cloud field estimations rather than an integrated representation within reanalysis data. These statements have been revised as follow:
"……we achieve a more accurate estimations of cloud fractions for both stratiform and cumulus clouds."
"The implementation of this model within reanalysis datasets like ERA-5 and MERRA-2 has resulted in enhanced cloud field estimations for stratiform and cumulus clouds."

l. 577/8 "advancing our understanding of BLC dynamics" – this is not true. While we get improved cloud fields, the DNN tells little about the dynamics; in fact, the point of ML / deep learning is that we can have forecasts without understanding of the dynamics.
Response: We appreciate the reviewer's insight and have deleted the statement. Instead, we focus on the DNN's capability as a powerful diagnostic tool for improving cloud field estimations.

l. 577/8 "improving the representation of low clouds – see above point for lines 491/559.
Response: We have deleted this statement as noted above.

Technical comments / Typos:
l. 97: 'this model' – which model?
Response: We specified it as the deep learning model.

l. 99: 'we strive to narrow the gaps in boundary layer clouds' – bad style, please rewrite!
Response: We revised it as "we aim to help bridge the existing gaps in representing boundary layer clouds…"

l. 139/140 Syntax incorrect.
Response: We revised this statement as: "We treat BLCs as synonymous with land-coupled clouds, in contrast to clouds that are decoupled from the PBL and land surface."

l. 147/148 please cite the data by their DOI (which is provided on the ARM website)
Response: We added the DOI as suggested.

l. 148 abbreviation CBH for cloud base height is introduced to late; The phrase has been used before.
Response: We defined CBH for the first appearance.

l. 183 'a advanced' – please correct!
Response: Revised as suggested.

l. 581 What are "synoptic regions"?
Response: Revised it as "synoptic patterns".