Response to Referee #1 on the manuscript “Marine cloud base height retrieval from MODIS cloud properties using machine learning”

Julien Lenhardt, Johannes Quaas, and Dino Sejdinovic
16.07.2024

Dear anonymous referee,
We would like to thank you for reviewing the revised manuscript and providing additional comments. Please find our response below, in which the review comments are in bold and followed by our response.

Best regards,

Julien Lenhardt on behalf of the authors

Minor/technical comments

It should be introduced why the reconstruction error, introduced as MSE (l. 211), does not carry a unit. If it is calculated as MSE between output and input field of the AE, it should carry the squared unit of the input variable. Or does it refer to the normalized input? This should be clarified.
The error is indeed applied to the normalised inputs. We added at the end of line 214 in the revised manuscript:
“The MSE considered here between the inputs and outputs of the AE is unitless, as the inputs are standardised before processing to ensure each channel is on similar scales and a more stable model training.”

The line numbers refer to the tracked changes document.

l. 23-25: This implies that the test data set consists of both ceilometer and CALIOP CBH retrievals. I guess, that is not what is meant. Furthermore, the phrase "performs well on both datasets" is not helpful. A more quantitative statement is necessary.
The word “evaluated” was included here to emphasis that the method is trained on the surface ceilometer measurements (line 21) but then evaluated using both surface measurements and CALIOP satellite retrievals. The following sentence at lines 23-25 was modified to:
“The statistical model performs similarly on both datasets, and notably on the test set of ceilometer cloud bases where it exhibits accurate predictions in particular for lower cloud bases and a narrow distribution of the absolute error, namely 379 m and 328 m for the mean absolute error and the standard deviation of the absolute error respectively.”

l. 71-73: "The objective of the developed method is primarily to produce CBH retrievals with reduced uncertainty, and additionally to extrapolate CBH retrievals from local surface observations to a wider spatial and temporal coverage" - Sounds like the ground-based observations are used as input and then together with satellite data extrapolated into space. However, they are only used to train the algorithm.
Indeed, the sentence was reformulated to:
“The objective of the developed method is primarily to produce CBH retrievals with reduced uncertainty, and additionally to provide extended spatial and temporal coverage compared to surface observations.”

l. 129-130: One sentence alone should not constitute its own paragraph. Possibly connect it to the following paragraph.
This sentence is actually the first one of the following paragraph and the corresponding pilcrow was already removed in the track changes manuscript (see last character of line 130).

l. 131: Right before, it is stated that earlier reports are based on human observers. Now it is stated "The CBH is derived using a ceilometer". Whether the authors utilize only CBHs retrieved by ceilometers should be clearly stated.
A clarification was added:
“In the surface observation dataset used in the study, the CBH is derived using a ceilometer, [...]”

l. 136-141: From Fig. 2b the binning of the utilized ground-based CBH data set becomes clear. However, according to this text passage, the authors use the minimum of the bin range for their evaluation. Other than introducing a negative bias, it is unclear how this procedure affects the results.
From choosing to use the bottom value of the bins as label, no real impact would be expected on the CBH predictions as they are considered as classes by the ordinal regression model. Naturally, some of the regression metrics would change (MAE, bias, RMSE). However, no clear comment could be made regarding for example the distribution in Figure 7 where the average might vary depending on the chosen labels.

l. 148-149: Table 1 caption - references to Section 2.1 and 2.2 seem to be mixed up.
Thank you for catching the typo.

l. 187-189: "The level 2 product [...]" - sentence can be removed.
This sentence is indeed a bit redundant with the previous paragraph and was thus removed.

l. 190: "in particular" - these words can be removed.
Removed.

l. 207: "Appendix B" -> "Appendix C" (I suppose is meant here)
Indeed, it was corrected.

l. 265-278: The structure of this new paragraph should be improved. First, they say, data are only taken for the year 2008. Then they say, data are only taken for a single year to avoid correlation between training and test data. Then they say they obtain 500.000 samples that are split into training, validation and testing based on retrieval date. So are all these 500.000 samples from the year 2008? And then how are they split into train, validation and test sets? Adding to the confusion is the statement regarding another test set using data over land for the year 2016. Please, clearly state, which period and location are used for training, validation and test, respectively.
The training, validation and testing datasets are created for the year 2008 and then split following their retrieval date in chunks of 70%, 20%, 10% “The overall built dataset consists of around 500 000 samples which are then splitted for training, validation and testing based on their retrieval date” (l271). To further test the generalisation of the model to unseen data, we build a similar dataset for the year 2016 but for which no train/val/test splitting is done “For further testing, Wwe additionally create a test dataset based solely on data from the year 2016 for further testing which includes tiles not only over ocean but also over land, indicating potential generalisation skill for unseen data including orography influence” (l272). The spatial distribution of the error is then reported for these overall 4 datasets “The spatially averaged reconstruction errors per cloud property channel are displayed in Figure 4 for each of the training, validation and testing datasets previously mentioned” (l275).

l. 354: It should be added that the 9x9-tile-simple method still features the OR. However, instead of using the AE feature vector as input, it uses the 3 cloud properties within the 9x9 tile simply as a flattened vector as input. At least, I assume that is what is done.

A OR model was indeed used with the 9x9 flattened tile data. This was clarified by adding in the sentence at line 354:

“[...] and an OR method relying on the flattened cloud properties of a 9x9 tile centred around the observation [...]”.

l. 357-361: The phrasing appears a bit complicated. The term "the baseline model" is not really defined. There appear 3 "baseline" models if you want (2 trivial, one 9x9 tile) so using this term is a bit confusing. And the error metrics for the two trivial methods should be identical since both always predict the 600m bin. Maybe it would be easier to just state the MA-MAE, and MA-RMSE for the developed AE-OR method and then in increasing order the errors for the other (9x9 method, trivial methods).

The phrasing and label for each mentioned method was clarified in this section.

l. 431-434: First, it is stated that the AE-OR method with immediate-threshold setup (IT) has similar (low) skill compared to the other retrieval. It is also stated, that the AE-OR method with all-threshold setup (AT) performs much better "on par with the other retrievals". If IT and AT methods differ, they cannot both be similar to the other retrieval method. Consider rephrasing this part.

This part was rephrased with the following:

“We first note that the OR method with an immediate-threshold setup fails at predicting adequately the cloud scene base height compared to all the other retrieval products, producing large errors (double-fold in comparison to the all-threshold setup). On the other hand, ORABase performs well with satisfying error measures and uncertainty in the predictions, on par if not better than the two retrievals from Goren et al. (2018) and Noh et al. (2017).”

l. 525 - "pacific" -> Pacific

This was corrected accordingly.