Author Responses

We thank reviewer #3 for the helpful suggestions and comments. Our responses and revisions are enumerated below.

Reviewer #3:

Lines 263–266: The LightGBM configuration is not described in sufficient detail for reproducibility. Please report the main hyperparameters, including the number of boosting iterations or trees, learning rate, maximum tree depth or number of leaves, and any regularization or subsampling settings. Please also state briefly whether these parameters were based on defaults, manual tuning, or an optimization procedure. An extensive hyperparameter analysis is not required.

Good point. We added some clarifying text:

"After testing the output of multiple classification packages while also tuning certain hyperparameters for the pixel segmentation task, we ultimately decided to utilize the gradient boosting classification algorithm in the LightGBM software package … The LightGBM segmentation model is trained using weights that consider the number of training samples for each class, such that the effective class weight ratios are $1:\sqrt{1/3}:\sqrt{1/3}:\sqrt{1/3}:1$ *for the clear, thin, intermediate, thick, and mask classes, respectively. … Beyond this customized weights hyperparameter, the segmentation model is configured with the following non-default hyperparameters, which produced the best classification performance: 150 trees with 31 leaves and a feature fraction of 0.75 in each tree (i.e., 1 of the 4 features is excluded from each tree); a bagging frequency of 1 (i.e., bagging is applied in each iteration) with bagging fraction of 0.70; a learning rate of 0.01; and a deactivated early stopping option during training."*

Lines 218–230: The training dataset is described as containing approximately 81 million pixel samples. Since neighbouring pixels within an image are highly correlated, this value should not be interpreted as the number of independent observations. Please clarify explicitly that the training–test separation was performed at the image level rather than the pixel level, and note that the number of annotated images better represents the number of independent samples.

We agree that the effective training dataset sample size is somewhat smaller than 81 million due to neighboring pixel correlations but likely not by much. This is discussed extensively throughout section 2.2, where we justify this argument (different lighting and pixel properties as a function of pixel zenith and azimuth angles, as well as the angle between a given pixel and the solar disk position). Note that, as discussed in that section, the classification is applied at the pixel level and NOT at the image level, as done for example in CNN algorithm applications, and therefore the number of annotated images strictly does not better represent the number of independent samples.

The text was updated:

"The training data for the segmentation model consists of roughly 81 million samples (pixels), though the effective sample size is somewhat smaller due to correlations among neighboring pixels. Those training samples were gathered from 513 images collected during …"

Lines 269–293: The uncertainty metric is a useful component of the product. However, the probability threshold of 0.5 used to identify uncertain pixels appears to have been selected primarily through expert judgement. Please provide a clearer justification for this threshold and

discuss its influence on the resulting uncertainty estimates. If readily available, a brief sensitivity assessment using nearby threshold values would be helpful, but an extensive additional analysis is not required.

We agree with the reviewer that this threshold has been selected primarily through expert judgment to adjust given algorithm's "level of decision confidence", which also depends on its hyperparameter configuration. The uncertainty metric and its values are already discussed extensively in the text:

"For a given cloudy class pixel … we set the pixel as "uncertain" if the RSS of all cloudy class decision probabilities is below an arbitrarily selected threshold of 0.5, which we evaluated using expert judgment … for a given clear class pixel, we set the pixel as "uncertain" if the clear class decision probability is below the same threshold of 0.5 … we think that the total uncertainty can be used as a general measure of confidence in the segmentation of a given scene (e.g., low uncertainty — higher confidence; high uncertainty — lower confidence, with a higher likelihood of a challenging or artifact-contaminated scene)… the resolved cloud cover, serving as our best estimate, should be considered together with the total uncertainty, as a measure of the lower and upper bound of our estimate. … We think that incorporation of the uncertainty in analyses is especially critical in cases where it is substantial (e.g., exceeding arbitrary values such as 30%, 40%, or 50%), in which case the resolved cloud cover should be taken with a grain of salt, as the convolution of the scene's cloud configuration with the spatial configuration of the masked sectors has many more degrees of freedom."

Lines 493–519: The comparison with ceilometer and KAZR observations provides a useful independent consistency check. However, ASI-derived cloud fraction and zenith-pointing active remote sensing measurements do not represent fully equivalent quantities, even when the near-zenith product and 5-minute averages are used. The former samples a spatially extended sky region, whereas the latter instruments sample a narrow atmospheric column. Please emphasize this distinction when interpreting the agreement and clarify that the comparison is primarily a consistency assessment rather than validation against an equivalent reference measurement.

We added the following text:

"We acknowledge that the KAZR-CEIL combination is treated here as insensitive to SZA for all practical purposes and therefore assumed as unbiased on a first order basis, noting that most solar effects are typically detected when compared against nighttime data, which are excluded from this analysis. We also recognize that because of the field-of-view differences from the ASISKYCOVER, the KAZR-CEIL combination does not represent fully equivalent quantities even when the ASISKYCOVER near-zenith (NZ) product and 5-min running means are used. Therefore, the following comparison should primarily be treated as a consistency assessment rather than validation against a benchmark measurement."

Lines 520–535: The year-long evaluation focuses mainly on the SGP central facility. Please qualify statements concerning transferability to other ARM deployments and other all-sky-imager platforms. Differences in camera configuration, aerosol conditions, surface reflectance, contamination, and cloud climatology may affect performance. It would be useful to clarify which components transfer directly and which require site-specific calibration, masking, or retraining.

We think that related qualifying statements are already provided throughout the text, for example:

"In general, cirrus clouds pose the greatest classification challenge, especially at larger SZA values... Thus, deployments to polluted sites or cirrus-dominated sites could benefit from recalibration of the model weights and potentially revisiting the training dataset."

"While one could suggest that this gradual enhancement is a mere artifact, the KAZR-CEIL curve (yellow) shows a similar behavior… the cloud cover anti-correlation with SZA was not detected in an equivalent 1-year bulk analysis we performed of ASISKYCOVER output for the ARM ENA site (not shown)."

"… these optical effects … are most pronounced close to the solar disk… Moreover, the equivalent ENA analysis (not shown) did not indicate any such artificial patterns, suggesting that this artifact tendency could be driven by local conditions (pollution, etc.) and influenced by the specific ASI instrument deployed at SGP."

"This agreement with the KAZR-CEIL is largely seen throughout all SZA ranges when accounting for the uncertainty ranges, in both the full FOV and NZ products … We note that this full-FOV correspondence with point-measurement datasets might be somewhat different, depending on the prevailing cloud types (see Riley et al., 2020; Wagner and Kleiss, 2016)."

"Future development of the ASISKYCOVER algorithm will include … dedicated configurations for highly polluted sites and events."

Lines 335 and 354: "ASISKICOVER" should read "ASISKYCOVER".

Fixed. Thank you.

Line 377: "The Grey-shaded regions" should read "The grey-shaded regions".

Fixed.

Line 480: Replace the comma after "otherwise)" with a period.

Done.

Line 482: "bottm" should read "bottom".

Fixed.

Line 483: "bi-model" should read "bimodal".

Corrected. Thank you.