Response to Reviewer 2

"Automated Analysis of Ripple-Scale Gravity Wave Structures in the Mesosphere Using Convolutional Neural Networks"

Jiahui Hu et al.

We thank reviewer for the careful and constructive evaluation. The comments led us to strengthen the algorithm description, quantitative validation, event definitions, sensitivity analysis, and figure presentation. The reviewer's original comments are reproduced below in regular type. Our responses are shown in *slanted* type.

Major comments

Reviewer 2:

Please insert line numbers the next time.

Response:

We thank the reviewer for this suggestion. Continuous line numbers have now been added throughout the revised manuscript to facilitate review and cross-referencing.

Reviewer 2:

The overall impression is that the manuscript was put together rather quickly as it shows several deficiencies such as missing references, figures that are not discussed, and misleading statements (see details below).

Response:

We thank the reviewer for the candid and constructive assessment. We agree that the original manuscript contained several omissions and statements that required greater clarification and qualification. We therefore undertook a comprehensive revision of the manuscript rather than addressing the comments only through isolated editorial changes. The revised manuscript now provides a more explicit description of the complete detection and validation pipeline, including image differencing and MAD normalization, patch construction and background sampling, the SE-CNN architecture and training procedure, full-image sliding-window inference, component filtering, temporal linking, and frame- and object-level matching. We also added a blinded single-reviewer audit of 720 strict-zero-overlap disagreement patches to characterize the morphology of cases that differed between the reference catalog and the automated detections. In addition, we revised the interpretation of the validation results to distinguish patch-level classification performance, frame-level recovery, object-level matching, and disagreement-patch morphology. We removed or qualified claims that could imply that the manual catalog is exhaustive, that

the classifier is independent of manual-label subjectivity, or that the detected morphology establishes a particular physical generation mechanism. The Discussion and Summary now explicitly state the limitations of the selected validation cohort and the inability of the present analysis to establish archive-wide occurrence rates or the physical origin of individual structures.

Reviewer 2:

There is a large difference between the number of selected ripple events from the network (952) compared to the manual catalog (720). This difference is explained by a better selection capability of the network. I have several questions regarding this:

(a) The authors write that a visual inspection afterwards shows that a substantial fraction of the sampled model-candidate/reference-unmarked crops were judged to contain clear ripple-like morphology. Can you show examples such that readers can judge for themselves?

(b) Is the change in selection capability caused by the network only, or is it mainly caused by the normalization using MAD? Are the ripples also better visible to the human eye after applying the normalization?

(c) Why is the difference present in one season (autumn) only? Do you have an explanation for that?

(d) Why does this overestimation not occur in the test data set? Are there any comparable patches included in the test data set? If not, what changes would you expect if you include such patches, which are manually marked as non-ripple patches, in the training set and the test set? In other words, how robust is the outcome of the network against changes in the preselected ripple and non-ripple patches?

Response:

We thank the reviewer for raising these important questions. We agree that the difference between automated and manually catalogued detections cannot by itself be interpreted as evidence that the network has better selection capability. We therefore revised the analysis to separate the effects of the complete preprocessing-and-classification pipeline, the composition of the validation cohort, and the incompleteness of the manual reference catalog.

(a) In addition to the representative full-image examples shown in Fig. 3, we conducted a blinded single-reviewer audit of 720 strict-zero-overlap disagreement patches. The audit comprised 177 reference-marked/component-unmatched patches and 543 model-candidate/reference-unmarked patches. Among the 543 model-candidate/reference-unmarked patches, 411 were classified as clear ripple-like, and 132 as non-ripple or artifact. The clear-ripple yield was 411/543 = 75.7%, with a night-cluster bootstrap 95% interval of 70.4%–80.2%. Among the 177 reference-marked/component-unmatched patches, 82 were classified as clear ripple-like and 95 as non-ripple or artifact, corresponding to a clear-ripple fraction of 46.3%. We emphasize that these values are descriptive results for deliberately selected disagreement cases. In particular, the 75.4% conditional clear-ripple yield is not an archive-wide precision

estimate because the model-candidate/reference-unmarked patches were selected from manually positive validation frames.

(b) We agree that the original experiment does not isolate the contributions of preprocessing and network architecture. The detector used in this study consists of signed two-minute image differencing, spatial MAD normalization, and SE-CNN classification, so the increased number of automated candidates cannot be attributed to the SE mechanism alone. We therefore do not claim that the SE mechanism itself is responsible for the larger number of automated detections. The revised manuscript now describes the effect of the preprocessing explicitly. Time differencing suppresses stationary and slowly varying structures while enhancing structures that change intensity or position between successive frames. MAD normalization then rescales the resulting image using the spatial median and MAD of each complete difference image. Thus, the input representation itself changes the visibility and selection characteristics of the detector. We therefore avoid attributing the automated–manual difference uniquely to the neural-network architecture. Instead, we interpret it as a property of the complete processing pipeline.

(c) The automated–reference difference is largest in autumn but is not restricted to autumn. Within the selected manually positive validation cohort, autumn contained 380 automated tracks and 147 reference tracks, giving a difference of 233 tracks. This represented 74.0% of the summed positive automated-minus-reference differences across seasons. However, these are conditional track counts from a selected manually positive cohort rather than exposure-normalized seasonal occurrence rates. The seasonal results depend on the composition of the validation cohort, the model-score threshold, the image-plane region of interest, and the provisional temporal-linking criteria. We have therefore removed the previous causal interpretation of the autumn excess. The revised Discussion identifies possible methodological contributors, including seasonal differences in image background and quality, training-data composition, reference-annotation behavior, and fragmentation or duplication during temporal linking. Because the study does not include coincident wind and temperature measurements, we do not assign the autumn difference to tides, winds, gravity-wave filtering, instability, or secondary-wave generation.

(d) The patch-level test set and full-image deployment evaluate different aspects of the problem. The reconstructed patch dataset contains 725 reference-centered positive patches and 725 deterministically sampled background patches. The background patches were not independently adjudicated and may contain unannotated ripple-like morphology or artifacts. Consequently, the held-out patch metrics characterize classification performance under the specific reference-centered-positive and sampled-background procedure. They do not estimate archive-wide precision, specificity, quiet-frame false-alarm rate, or performance against an independently adjudicated hard-negative set. The revised Results explicitly make this distinction and note that the disagreement audit was not used to retrain the classifier. The 132 model-candidate/reference-unmarked patches classified as non-ripple or artifact therefore represent potentially useful difficult negatives for future work, but they are not treated in the present study as independently confirmed negatives. We have consequently removed the earlier interpretation that the difference between automated and manual counts directly demonstrates superior network performance and instead present the disagreement as a combination of model detections, reference-

label uncertainty, preprocessing effects, and the limitations of the manual reference catalog.

Minor comments: Introduction

Reviewer 2:

At some points the authors claim that the network is objective or reduces a subjective bias. I find these statements inappropriate as the network is trained with a manual and therefore subjective selection. It should also recall a former subjective bias. Please rephrase these statements.

Response:

We thank the reviewer and agree with this criticism. The original wording was too strong. Because the SE-CNN was trained using manually labeled ripple and non-ripple examples, it inherits the definitions and uncertainties of those labels and cannot be regarded as independent of the subjectivity of the manual reference catalog. We have therefore removed statements describing the automated method as "objective" or unbiased. We instead describe its methodological advantage as the consistent and reproducible application of a learned morphological criterion across the analyzed images. We also explicitly acknowledge that the classifier remains dependent on the manually labeled training data.

Minor comments: Methodology

Reviewer 2:

There are better references for the height than Hill and Taylor (1991), such as Baker and Stair (1988) and others (von Savigny et al. (2015), Garcia-Comas et al. (2017), Wüst et al. (2020)).

Response:

We agree that Taylor and Hill (1991) was not an appropriate reference for the nominal OH-emission-layer altitude. We replaced that citation with references directly addressing OH-layer altitude and variability, including Baker and Stair Jr. (1988), von Savigny et al. (2015), García-Comas et al. (2017), and Wüst et al. (2020). We also revised the wording so that approximately 87 km is described as the nominal OH emission peak altitude rather than as a fixed altitude, and we explicitly note that the layer altitude varies temporally and geographically.

Reviewer 2:

Can you provide more information regarding the selection of the ripple and non-ripple patches? Are only perfect ripple patches taken, or are they randomly chosen from the whole catalog?

Response:

We thank the reviewer for requesting clarification. The positive patches were not selected only because they were visually "perfect" ripple examples, nor were the two classes randomly sampled from the entire image archive. The reconstructed patch dataset contained 725 positive examples, with one 41 × 41-pixel patch centered on each available corrected reference-annotation coordinate. For each positive patch, one background patch was selected from the same source image using a seeded deterministic sampling procedure. Background centers were restricted to the central field of view, required to be at least 60 pixels from every reference annotation in that image, and were separated from other selected background centers. Importantly, the sampled background patches were not independently adjudicated as true non-ripple examples and may contain unannotated ripple-like morphology or artifacts. We therefore clarify that the patch-level metrics characterize discrimination under this particular reference-centered positive and sampled-background procedure, rather than against an independently confirmed hard-negative set.

Reviewer 2:

The vanillaCNN is first mentioned below Fig. 2. What is the difference between vanillaCNN and SE-CNN? Is the vanillaCNN used in the following? Otherwise, you might delete these parts from the manuscript.

Response:

We thank the reviewer for pointing this out. We replaced the term "vanilla CNN" with "baseline CNN" throughout the revised manuscript. The baseline CNN uses the same convolutional architecture and training framework as the SE-CNN but omits the squeeze-and-excitation blocks. It is retained solely as an architectural comparator for evaluating the effect of the SE mechanism. The SE-CNN, rather than the baseline CNN, is used for the full-image inference and automated ripple catalog.

Reviewer 2:

Can you explain the large drops in the F1 score at some particular epochs?

Response:

We thank the reviewer for this observation. We examined the validation precision and recall curves separately. The largest temporary decreases in validation F1 were accompanied primarily by decreases in validation recall, whereas validation precision varied less strongly. Because the validation set contained only 194 patches, a small number of samples crossing the classification threshold can produce relatively large changes in recall and consequently in F1. Thus, the temporary F1 reductions do not indicate a persistent degradation of the model; rather, they reflect fluctuations in the classification of a relatively small validation set during training.

Minor comments: Results

Reviewer 2:

Can you give the criteria for spatial coincidence that have to be fulfilled for the two selections (manual and automated) to be classified as overlapping?

Response:

We thank the reviewer for requesting clarification of the spatial-matching criterion. Each reference annotation is represented by a 41 × 41-pixel support centered on its catalog coordinate, while each retained automated detection is represented by the filled mask of its thresholded component. A reference support and an automated component are considered spatially overlapping when their masks share at least one pixel in the same frame. No minimum overlap fraction, intersection-over-union threshold, or centroid-distance criterion is imposed. At the frame level, an evaluable manually positive frame is classified as spatially recovered when at least one eligible reference support overlaps at least one retained automated component. At the object level, maximum-cardinality one-to-one assignment is performed among all spatially overlapping reference-support/component pairs. Manual-object recall is calculated as the number of assigned reference supports divided by the number of eligible reference supports. The automated-component assignment fraction is reported separately and is not interpreted as precision because the manual reference catalog is not assumed to be exhaustive.

Reviewer 2:

Fig. 4 and Fig. 5 are not referenced in the text.

Response:

We thank the reviewer for pointing this out. The revised Results section now explicitly references and interprets Figures 4 and 5. Figure 4 is described as the daily comparison of provisionally linked reference and automated tracks and their automated-minus-reference differences, while Figure 5 is described as the seasonal comparison of frame-, object-, and track-level counts within the selected manually positive validation cohort. We also explicitly state that these track counts are conditional on the selected validation cohort and are not exposure-normalized archive-wide occurrence rates.

Reviewer 2:

In Figs. 5–8 the only two colors used are green and red. This may be problematic for people with color blindness.

Response:

We thank the reviewer for this suggestion. The figures have been revised to improve color-vision accessibility and visual distinction between reference and automated results. The revised figures use distinct colors together with different marker/line representations, including blue and orange for the principal reference and automated series and black for differences. The figure captions have also been revised to make the meaning of the plotted quantities explicit.

Reviewer 2:

I find Fig. 4 hard to read. Maybe two panels above each other instead of one are better.

Response:

We thank the reviewer for this suggestion. Figure 4 has been redesigned as two vertically stacked panels. The upper panel shows the daily reference and automated track counts, while the lower panel shows the automated-minus-reference difference together with a horizontal zero line. The revised layout and larger graphical elements improve the readability of both the count comparison and the seasonal/day-to-day differences.

Reviewer 2:

Figure 6 is not discussed in the text. Please include the corresponding text or delete the figure.

Response:

We thank the reviewer for pointing this out. Figure 6 has been revised and is now explicitly introduced and discussed in the Results section. The revised figure shows the seasonal distributions of provisionally linked reference and automated track durations, together with one-to-one overlapping reference–automated track pairs whose durations fall within the same displayed bin. We emphasize that these durations are products of the provisional temporal-linking procedure and are not independently verified physical event lifetimes. The corresponding Results and Discussion text also clarifies that the distributions are conditional on the selected manually positive validation cohort and are not exposure-normalized seasonal occurrence rates.

Reviewer 2:

The text describing Fig. 7 is not precise enough as at some points the statements are only true for one season or specific situations and not in general. For example, only in autumn does the automated selection show a higher fraction of very short-lived ripples, and the manual selection does not show more long-lived ripples in winter. Please add more text here and improve your statements.

Response:

We agree with the reviewer that the previous interpretation of the lifetime distributions was too general and, in some cases, misleading. The former Figure 7 has therefore been removed. In particular, we removed the previous claims that the manual and automated lifetime distributions were generally similar and that the manual catalog contained a higher fraction of long-duration winter events. In the revised manuscript, the relevant duration analysis is presented in the new Figure 6. Rather than interpreting the distributions as independently measured physical event lifetimes, we apply the stated provisional temporal-linking procedure to the reference and automated detections and present the resulting track-duration distributions descriptively. We explicitly caution that these tracks are algorithmically constructed and are not independently verified continuous physical events. The revised Results describe the observed duration-bin distributions directly, while the Discussion addresses possible effects of detection continuity, score-threshold crossings, component fragmentation, and temporal linking without interpreting the differences as physical event persistence.

Reviewer 2:

Why are there no manual selections of ripple with lifetimes longer than 5 minutes in Fig. 7?

Response:

We thank the reviewer for highlighting this issue. The apparent absence of manual events longer than 5 min in the former Figure 7 resulted from the previous plotting procedure rather than from an observational constraint on ripple duration. The manual catalog contains frame-level annotations and does not provide independently adjudicated physical event start and end times. We have therefore removed the former Figure 7 and no longer interpret that histogram as a measured manual-event lifetime distribution. In the revised analysis, reference annotations and automated detections are linked using the stated provisional temporal-linking procedure, and the resulting track-duration distributions are presented in Figure 6. Reference tracks therefore now occur in the 0–5, 5–10, and 10–20 min bins. We emphasize that these durations characterize the output of the linking procedure and should not be interpreted as independently verified physical event lifetimes.

Reviewer 2:

At the end of the section a statement about the network being more objective is given again. I find this inappropriate as the network is trained with subjective manual selections.

Response:

We agree. We removed the description of the automated approach as "objective." The revised manuscript instead describes the method as a reproducible and scalable implementation of a learned morphological criterion while explicitly acknowledging that the criterion is derived from manually labeled training data and therefore inherits uncertainties and potential biases in those labels.

Minor comments: Discussion

Reviewer 2:

There are again wrong or at least misleading statements about the observation frequencies as a function of lifetime. In total, the amount of detections of very short-lived ripples should be very similar, as only in autumn does the automated selection show a higher fraction (compare Fig. 7).

Response:

We agree with the reviewer that the previous statements about detection frequency as a function of lifetime were too general and could be misleading. In particular, the manual reference catalog contains frame-level annotations but does not provide independently defined event start and end times; therefore, a direct comparison of reference and automated physical event lifetimes is not justified. We have removed the previous general statements about the relative occurrence of very short- and long-lived ripples. In the revised manuscript, Figure 6 presents only the durations of provisionally linked reference and automated tracks produced by the stated temporal-linking procedure. These

distributions are interpreted descriptively rather than as independently measured physical event lifetimes. We also clarify that the track-duration distributions are conditional on the selected manually positive validation cohort and are not exposure-normalized seasonal occurrence rates. The revised text reports the observed duration-bin distributions without attributing differences to physical persistence or occurrence. Longer linked tracks may reflect persistent structures, merging, or repeated association of nearby detections, while differences in the shortest duration bin may reflect detection continuity, score-threshold crossings, fragmentation, and the temporal-linking criteria. We therefore no longer make a general claim that the automated and reference selections have similar or different observation frequencies as a function of physical lifetime.

Reviewer 2:

The complete discussion section lacks comparisons to other studies. Is it possible to add some comparisons to other studies here?

Response:

We thank the reviewer for this suggestion and agree that the original Discussion did not place the results sufficiently in the context of previous observations. We have expanded the revised Discussion to compare the present results with previous studies of ripple-like and small-scale wave structures. Yue et al. (2010), Suzuki et al. (2011), Rourke et al. (2017), and Wüst et al. (2026) are now discussed in relation to seasonal behavior, local-time or directional variability, event duration, and methodological differences in identifying small-scale structures. We also explicitly caution against direct numerical comparison because the studies differ in geographic location, observing period, emission layer, camera characteristics, preprocessing, structure definitions, score thresholds, occurrence denominators, and evaluation units. In particular, the revised Discussion no longer treats the seasonal behavior observed in the selected YRFS validation cohort as a universal ripple climatology. Instead, the results are interpreted as specific to the present validation cohort and detection methodology, and possible physical explanations are separated from the directly observed catalog differences.

We again thank reviewer for comments that substantially improved the transparency and reproducibility of the study.