This manuscript presents a method for detecting small-scale airglow wave structures using a modified YOLOv7. The paper is well written, scientifically sound, and a welcome contribution. A few concerns regarding the methodology need to be addressed, however. Below are my itemized comments.

We thank the reviewer for the very valuable comments. We answered all of them (in orange, see below) and changed the manuscript accordingly.

- Line 173. 'BYOL….'

  Pretraining using BYOL may not be particularly beneficial for this application, and the manuscript provides no evidence that BYOL improves performance. Nevertheless, the authors should at least include further details on the implementation of BYOL and YOLOv7. In particular:

  1. **Which variant of YOLOv7 was used? YOLOv7 has multiple variants with different backbones and capacities.**

     Yolov7-tiny (info added in the manuscript)

  2. **Was the BYOL pretraining performed starting from a randomly initialized network, or from an already pretrained YOLOv7 backbone (for example COCO pretrained)?**

     It was trained from a COCO pretrained network, which is provided by the official paper (info added in the manuscript).

  3. **What data augmentations were used for BYOL during pretraining?**

     Colour Jitter, Random Flips, Random Resized Crop, Normalization, Gaussian Blur (info added in the manuscript)

  4. **Did the authors compare the performance of the BYOL-pretrained model with a non-pretrained (trained from scratch) or standard pretrained YOLOv7 model?**

     Yes, compared to the regular COCO pretrained model. While the training accuracy was not significantly better, the convergence speed was increased.

- YOLOv7 is adapted to output wavelength and orientation. This is in fact a substantial change to the net and these tasks deviate from the original design goal of the YOLO structure. YOLO and its variants are highly optimized for object detection but not much for extracting information from the objects. Adding three additional regression features may severely interfere with its main task (object detection). Generally speaking, regression tasks in neural networks require careful design of the architecture, loss functions, and training strategy. Simply adding three additional regression outputs to the YOLOv7 head is unlikely to work well without further justification or validation.

  Has the author tried using YOLOv7 in its original form and compared the numbers?

  Yes, and the accuracy does not change when no additional parameters are used. However, in this case, it should also be mentioned that the value ranges for the newly generated predictions have been carefully selected.

The wavelength, for example, is calculated in relation to the bounding box size. Since the wavelength must be shorter than the box size, the value range remains exactly the same. Both predicted values (bounding box and wavelength) are initially in the range $(-\infty,\infty)(-\infty,\infty)$ and are then mapped to a range of $(0,4)(0,4)$ using the same formula:
$(2 \cdot \text{sigmoid}(x))^2$
This was done to ensure that the gradients do not differ too much in magnitude and that both bounding boxes and wave parameters are predicted with high precision.

- **Line 284. A validation set is absolutely necessary. The testing set is the one that is optional. In recent work, some studies omit a separate testing set and report validation metrics only, provided that the validation set is sufficiently large and representative. Omitting the validation set, however, is not consistent with standard neural network training practice, since it prevents proper monitoring of overfitting and reliable model selection. Without a validation set, it is impossible to detect overfitting during training. Given that the training set is relatively small (only in the thousands), not using a validation set is a fatal mistake, and the performance metrics obtained during training are likely to reflect overfitting rather than true generalization.**

  We believe this might simply be a naming issue. All model architectures and hyperparameters were fixed *a priori*, and the test dataset was used only for evaluation. There was no adjustment of hyperparameters during training, nor were any other training decisions made based on that dataset. Nevertheless, we repeated the training while splitting the test set into separate test and validation sets of equal size. No overfitting was detected in this setup, and the precision and recall values remained unchanged.

- Line 288. '.....78% are correctly identified'.

  This is not an appropriate way to report regression performance. Regression tasks should be evaluated using continuous error metrics such as MSE or RMSE, and wavelength and orientation should be reported separately with their respective error distributions. Using a binary threshold to count predictions as "correct" obscures the actual performance and does not provide enough information to assess model accuracy.

  The decision to use a binary classification between correct and not correct was done, since metrics like MSE or RMSE would hide this important information. For us, the required minimal level of quality was important to get an estimate of how many predictions are "correct" and therefore how trustworthy the calculated propagation directions (in the later chapter) would be.

- Figure 5 and ~ Line 286.

  The reported performance is subpar for a task that should not be particularly difficult for a modern neural network. This suggests there might be issues with the data, the net config, and/or training. I suggest that the authors retrain the network without the additional regression features, expand the training data if possible, and include a validation set. If the size of the training dataset is the main constraint, using the testing set as the validation set and reporting the validation metrics is also acceptable.

The orientation and wavelength can be handled much more effectively by a dedicated CNN or ViT that processes the image content within the bounding box. Or even better, a DETR-based model would be more suitable for predicting both the bounding box and the orientation. However, adapting the method to DETR would require substantial additional work and is not strictly necessary here.

Thank you for this valuable feedback.

Another training with validation and test set was performed as mentioned above. No overfitting was detected, and the precision and recall values remained unchanged. For the additionally calculated metrics, please have a look at the answer below. Furthermore, we used YOLOv7 also in its original form (without using additional parameters), as written above, and the accuracy did not change.

As mentioned in the manuscript, one inherent problem is the ambiguity of wave events. There were many cases where wave events were on the cusp of being classified as such. Therefore, the model might predict a wave that was not labelled as such, and then be trained not to detect it as a wave anymore, and vice versa. The best way to avoid this issue is to label the probabilities of wave events and adjust the training accordingly. In the current solution, a '50 % wave event is trained to be predicted with either 100 or 0 % confidence, which is 'half wrong' either way.

Expanding the data set would of course be beneficial. However, the project in which this work was performed is finished and Jakob Strutz, who did the AI analysis is working in another job. Therefore, enlarging the data set us not possible.

Line 320. 'In summary, the 2D-FFT provides more accurate results as 78% of the FFT predictions have an error of less than or equal to 2.5° for the orientation and 3% (relative to the labelled wavelength) for the wavelength. For the modified YOLOv7 algorithm, 78% of the results are considered correct, if the wavelength error is less than 10% relative to the labelled wavelength and the error of the angle is less than 10°.'

This is not a fair comparison. The 2D-FFT results are evaluated using an error threshold of 2.5° for orientation and 3 percent for wavelength, while the YOLOv7 results are evaluated using a much looser threshold of 10° and 10 percent. Because the criteria differ by a large factor, the "78 percent correct" numbers for the two methods cannot be directly compared.

While I understand the authors are trying to show that 2D-FFT performs better on normal images, the comparison is still pretty weird. It would be better to compare both methods under the same benchmark. MSE or RMSE is the standard metrics for regression tasks like these.

We fully understand your point. In this section, we intended to highlight that achieving the same accuracy (78%) can be accomplished using stricter requirements for the FFT than for the direct calculation using YOLO.
In the meantime, we performed both calculations using a relative wavelength error of 20% (compared to the labeled wavelength) and an angular error of 10°. This appears reasonable, as wavelength errors of 20% and angle errors of 10° would not significantly affect the calculation of the propagation direction or the classification between secondary waves and ripples.
Using this metric, 89% of the FFT predictions are correct, with an MSE of 7.60 km^2 for the wavelength and 220.71(°)² for the angle. For the YOLO predictions, 84% are correct using the same metric, with an MSE of 10.02 km^2 for the wavelength and 390.10(°)² for the angle.