Review: A comparative analysis of deep learning models for classifying shallow mesoscale cloud patterns in satellite images

By: Anna Granberg, Vilma Lundholm, Pouria Khalaj, Manu Anna Thomas, Yifan Ding, Daniel Jönsson, and Abhay Devasthale

Anonymous Reviewer #1

Understanding the dynamics of mesoscale patterns of shallow convection is a very active field of study and availability in observations and compute power allow now thorough investigations at this scale. As both the traditional stratocumulus mesoscale patterns as well as the more recently named shallow cumuli found primarily in the downstream trades, are able to alter the mean radiative effect (e.g. Bony et al. 2020), the scientific community has developed various methods to detect these patterns in various data sources from satellite observations to model output. This manuscript compares the ability of deep neural networks based on three different neural network architectures to classify mesoscale patterns of shallow convection in SEVERI satellite images. The authors study in particular the mesoscale patterns as defined in Stevens et al. (2020) and e.g. Atkinson (1996) with the extension of two additional categories.

While I am not aware of such an in-depth analysis of different models, the added value of this analysis is currently unclear to me due to several reasons:

1. Several models have been previously successfully trained by other authors on the patterns defined in Stevens et al. (2020) and proofed that these patterns have physical differences (e.g. Eastman et al. 2024) and their detections help to further understand their mesoscale dynamics.

2. The three trained models are specific to the given task and all needed their individual hyperparameters including different learning rates (see l. 220). The manuscript titles correctly that this is a comparison between “deep learning models”, but it would be much more informative if more general conclusions could be drawn on which “deep learning architecture” is best and what architectures and training methods used so far are lacking or could improve on.

3. The research outlook reflects this vague applicability of the study’s results as it mentions to scale this study to more annotations and more architectures. Because there are already annotations of a large-subsection of these mesoscale patterns available from previous studies that are more comprehensive (e.g. Rasp et al., 2020 with several domain, multiple annotators), it is unclear what should be achieved.

As the study is well written and the inter-comparison between the models is done in a great detail, it would be a shame if the authors could not investigate more time to draw more general conclusions that go beyond these particular models and tasks.

We thank the reviewer for the thoughtful and detailed assessment of our manuscript. We appreciate your comment that this is an in-depth analysis comparing different models, and we have tried to address your concerns, clarifying the added value of our work and how it is different

compared to the studies mentioned. Please find below the point-by-point response to your comments.

Major comments

The study assumes that a particular region of a satellite image can be attributed to one of the 8 selected mesoscale patterns. The reality seems more complex and an “unknown” category should be included. Please clarify that during the validation only annotated subsets are used, so it is ensured by design that always one of the patterns is present.

Yes, we use only annotated subsets during validation and we get at least one of the 8 mesoscale patterns present in any random image chosen. This is clarified in the revised manuscript. We followed Stevens et al (2020) pre-defined categories and two extended patterns that were found to be very dominant in the images. The model is not evaluated on cases where no pattern is present as this study is aimed at understanding the more commonly seen organizations of the shallow stratocumulus clouds.

This particular region is chosen for two reasons:

1. Previous studies have demonstrated that this region contains persistent shallow stratocumulus decks that organize into distinct mesoscale patterns depending on meteorological conditions (e.g., Stevens et al. 2020; Atkinson 1996). This makes it an ideal testbed for studying pattern classification.

2. SEVIRI scans this region with high temporal (15-minute) and spatial (3–4 km) resolution, enabling us to capture the evolution of mesoscale patterns and is ideal for robust annotation and model training. This temporal resolution is critical for understanding the evolution and is superior to many other satellite instruments for this application.

This is now clarified in the revised manuscript.

- The k-means clustering model is not comparable to the benchmark dataset as the patterns are defined by visual impressions and not by a clustering analysis. It is unclear why this has been included.

Thank you for pointing this out. This is a valid point and we totally agree that results of our k-means analysis are not directly comparable to that of visual inspections by the human annotator. Our purpose by including the k-means clustering was to merely examine the clustering of DINOv2 embeddings as an exploratory approach, and not treating it as a fourth classifier per se. This has been further clarified in the Discussion and the caption of Table 12.

The first paragraph of the Discussion has been revised as: “In this study three supervised deep learning based models — CNN, ResNet50, and DINOv2 — are developed to classify shallow oceanic clouds using satellite image data. Furthermore, an unsupervised clustering of frozen DINOv2 embeddings was performed using the k-means algorithm. This was to examine the extent to which the model can successfully recover the features of the cloud patterns (identified visually) in the latent space. Being merely an exploratory approach, the results of our k-means

analysis are not directly comparable to those of visual inspections by the human annotator. In other words, our k-means analysis is not meant to be interpreted as a fourth model to classify cloud patterns."

The caption for Table 12 is re-written as:
"Table 12. Classification performance per predicted cluster (six clusters). Note that these results merely correspond to cluster memberships according to the k-means algorithm, and therefore cannot be directly compared with the performance metrics of the three supervised models trained on the annotated dataset (identified visually) provided by the human expert."

- The annotation practice needs more details.

The first paragraph in section 2.2 has been replaced by the following paragraph with more details on the annotation.
"The 2021 SEVIRI images are manually annotated by Dr. Manu Anna Thomas according to the cloud patterns described in Fig. 1 (right). The annotator follows the mesoscale pattern definitions from Stevens et al. (2020) and McCoy et al. (2023) as80
mentioned in Section 2.1, along with two additional categories that were frequently observed: a sand pattern, characterized by wavy band-like cloud structures, and a honeycomb pattern, characterized by quasi-circular cloud cells arranged in a polygonal
network. Each SEVIRI image is annotated with a pattern class label and the corresponding bounding box co-ordinates that encompass the full extent of each pattern. The images from the first four days of four months (February, May, August, and
November) during daylight hours (08:00–16:00) are used for the annotation. This temporal sampling captures seasonal variability. Daylight hours are selected because SEVIRI visible channels are used to create these images. The individual bounding
boxes in Fig. 2 denote the different mesoscale patterns, visually identified by the annotator."
An additional paragraph is also included:
"The mesoscale cloud patterns examined in this study do not change significantly over time scales of a few hours as captured in SEVIRI images. As a result, there is a high chance that the annotations which are later retrieved by the models are temporally correlated rather than fully time-independent. Although this is not necessarily a disadvantage, but depending on the intended application, it is important to recognize this temporal correlation when one needs labeled data which are strictly
time-independent."

- As this study does not make use of previous manually labeled datasets (Rasp et al. 2020, Schulz, 2022), further description is needed on what the instructions to the annotators were to identify certain patterns. A definition or reference to previous work for the Sand and Honeycomb patterns seems missing.

All patterns follow Stevens et al. (2020) and Mc Coy et al. (2023) definitions, except for the two extended categories:

Sand pattern: Cloud patterns that give a more wavy band like structure contributes to this label class.
Honeycomb pattern: Quasi-circular cloud cells arranged in a polygonal network.

This is mentioned in Section 2.2 of the revised manuscript.

- It sounds like only one annotator has done the annotations. From past studies it is known that there can be quite some spread in agreement. This needs to be discussed when it comes to interpreting the results of the different neural networks and their challenges distinguishing some of the patterns.

A single expert annotator was chosen to create the labelled data set for the individual cloud patterns. This is because an attempt using annotators without meteorological background resulted in high misclassification rates due to the subtle distinctions in those patterns. In the previous studies by Stevens et al., 2020, the annotators were all scientists with similar backgrounds. A single expert annotator can reduce these misclassifications and at the same time introduce some kind of variability in the labeled data set. Non-expert annotations would require tens of thousands of samples to achieve comparable performance due to higher label noise whereas with a single expert annotator, a couple of thousands of samples would suffice.

- Have the images been served randomly to the annotators or consecutively?

The images are not selected randomly; instead, we chose the first four days from four randomly selected months distributed across the year. This is mentioned in the manuscript.

- The models get served only a subset of the original domain and are lacking context that the human annotator had. This difference and its implications need to be discussed also with respect to the choice of the 70x70 pixel domains.

Thank you for your suggestion. We have added a paragraph on this to the end of Section 4.1 (3.4.1 in the revised manuscript), and also one paragraph to the end of the Discussion (Section 8).

The text added to the end of Section 4.1 (Section 3.4.1 in the revised manuscript):
“Note that the context that the human annotator had regarding the input data differs from that of the models. In particular, the expert meteorologist who annotated cloud patterns (Dr. Manu Anna Thomas) has viewed the complete SEVIRI scenes, albeit, cropped to the area of interest, i.e. the west coast of Africa. However, the models were only fed with the extracted sub-images, which were further transformed into 70x70 pixels. As a result, the task of models was to classify localized cloud structures rather than detection and subsequent classification of such patterns in full scenes. The choice of 70x70 pixels as the resolution of input images is motivated by the fact that it is divisible by the patch size of DINOv2, which is 14x14. Furthermore, as a practical compromise, 70 is an integer value which is sufficiently close to the typical sizes (width and

height) of the extracted images in the benchmark and input datasets. Combined with having a fixed size for all input images, this helps with keeping the effect of required transformations on images to a minimum. A possible shortcoming of using a fixed size for all input images, could be the potential removal of some information, related to the scale and morphology of the cloud patterns. Such information might be crucial for the classifications of some patterns."

The text added to the end of the Discussion section 8 (Section 7 in the revised manuscript):
"It is important to note that the human annotator had access to the full-scene context of SEVIRI files whereas the models were trained exclusively on local
bounding boxes of cloud patterns, identified visually by the human expert. In particular, all models have received images of size 70×70 as input. One implication of this is that the original scale information and morphology of the cloud patterns could be potentially affected."

Further comments

- l. 1: There is no connection made in this manuscript on how this deep learning technique will help climate models to improve. Can this be more widely motivated why there is interest in capturing these mesoscale patterns?

The text below that addresses the reviewer's question is added in the outlook section in the revised manuscript.

"Beyond their value for automated classification, deep learning techniques for identifying mesoscale cloud organization also offer concrete pathways to improve climate models. Once patterns are detected with high accuracy, their associated meteorological conditions can be systematically characterized. This will provide us with more information on the cloud-controlling factors across different patterns. These details can then be used to compute 2D and 3D histograms of cloud physical and microphysical properties and their corresponding meteorology which could then be used to refine the representation of shallow stratocumulus regimes in climate models. These linkages between mesoscale cloud patterns and their governing meteorology provide a physically informed basis for constraining and testing AI-based cloud parameterizations."

- l. 28: EUREC4A

This is corrected to EUREC^{4}A in the revised manuscript.

- l. 29: Barbados Cloud Observatory

This is corrected in the revised manuscript.

- l. 50: These studies have already advanced insights into mesoscale cloud processes that can be used to improve climate models. More studies are certainly needed, but how this study helps to do so needs to be better described. What are e.g. the short-comings of previous (machine learning) approaches that this one (helps to) overcome?

The AI models are advancing rapidly and these models have their strengths and limitations. The current study uses the state-of-the-art AI-models that are suitable for pattern classification and carries out a comparative study. Such detailed assessment has never been done before in the context of organization of low-level mesoscale clouds. The present study also goes beyond the six patterns (flower, sugar, gravel, fish, open cell and closed cell) that have been studied so far and includes sand and honeycomb structures as well.

- l. 51f: It is unclear how this manuscript contributes to the goal of AI4PEX as this study is purely observation based. This needs more explanation. A perfect place for this would be the outlook.

See the response to l. 1 above.

- l.54: Deep neural networks have been used before to classify mesoscale cloud patterns also based on RGB images.

This sentence has been rephrased giving more details such as “However, existing efforts that classify mesoscale cloud patterns from RGB satellite imagery (for e.g. Rasp et al. (2020); Eastman et al. (2024); Chatterjee et al. (2024)) have primarily focused on demonstrating successful classification with individual architectures without systematic comparison of multiple architectures. In contrast, the present study evaluates and compares three distinct neural network architectures on SEVIRI data, and, to our knowledge, is the first to exploit SEVIRI’s high spatial and temporal resolution for mesoscale cloud pattern classification”

- l. 64.: What artifacts and distortions have the images been corrected for?

It is radiometric artifacts and geometric distortions. We have amended the corresponding line to include this information in the revised manuscript.

- l. 66: The guide seems to be the source for selecting the specific channels for RGB and it’s footnote should be better placed accordingly.

This is now corrected in the revised manuscript.

- l. 66f: The guide link provides several guides. Assuming that the RGB Cloud product is meant, this product seems to use the broadband, high-resolution channel for red and green and an

infrared channel for blue. This is different to the spectral bands 0.635µm and 0.81µm indicated here.

The line is rephrased as:
The daytime images in this study are generated by compositing three spectral bands centered at 0.635 µm and 0.81 µm in the visible range, and 10.8 µm in the infrared range which are, in turn, assembled into red, green, and blue (RGB) channels.

- l. 66f: Please indicate here which resolution the RGB images have and later in the text which size translates to for the 70x70 pixel subsets.

This is added in the revised manuscript in Section 2.1: "…SEVIRI scans it with high temporal (15-minute) and spatial (3-4 km) resolution. A line is added to section 3.4.1 in the revised manuscript "this resolution translates into a physical size of 210-280 km."

- Figure 2: Lat/lon or at least scale is missing.

We adapted the caption of the figure to include this information as "Figure 2. Two examples of annotated SEVIRI images. All classes are not necessarily present in every image. The images correspond to the area of interest with latitude and longitude extents of $\phi \in$ [34S, 3N] and $\lambda \in$ [16W, 14E], respectively."

- l. 75: I assume that the months were chosen to reasonably capture the seasons and still have a relatively independent set of months left for validation. Please briefly motivate the choice of months. The domain covers also the deeper tropics, has there been any issue with mis-classification of patterns concerning deep convection by the human annotator?

Yes, the months were indeed chosen to capture the seasonal variability. Previous studies have shown the dominance of these clouds in the subtropical regions, especially over the Atlantic Ocean, where the high temporal resolution data from SEVIRI/MSG is also available. These patterns rarely occur in the deep convective regions when the African monsoon is active.

- l. 79: "…images from the same daytime range as the original data set, including five randomly selected days excluding those of the original data set." Improve clarity. The benchmark dataset consists of 5 randomly selected days that have not been used during training. All 15 min images between 8 - 16 UTC (?) have been annotated.
Is this true? This gives a lot of similar annotations as these mesoscale cloud clusters do not change much, particularly not their overall structure within a few hours. Would it not have been more robust to classify randomly 5*32 images (8h daylight @ 15min -> 32 images)?
This is a valid point. Since the zenith angle of the Sun changes in the images, our intention was to train and test the model given different daylight conditions. We have added a note to the

revised manuscript to inform the reader that the retrieved annotations are not necessarily independent. This information is added towards the end of the first paragraph in Section 2.2 as "It has to be noted that the mesoscale cloud patterns examined in this study do not change significantly over time scales of a few hours as captured in SEVIRI images. As a result, there is a high chance that the annotations which are later retrieved by the models are temporally correlated rather than fully time-independent. Although this is not necessarily a disadvantage, depending on the intended application, it is important to recognize this temporal correlation particularly in contexts where strictly time-independent labelled data are required."

- l. 85: :"only a single mesoscale pattern", a "single cloud structure" sounds like individual cloud cells were the minimal size.

The line is rephrased as: Each bounding box is guaranteed to enclose only one cloud pattern, to the extent that is possible by visual inspection.

- Tab. 1.: It would be helpful to show the "total" below the individual category counts to show that this is the sum of the above.

This is done in the revised manuscript.

- Tab. 1.: "Images" here refers to annotations or bounding boxes? Please rename as it otherwise can be confused with the number of full-domain RGB images.
Yes, we understand the confusion. "images" will be replaced by "annotations".

- l. 114: Please state the GPU model here, as I first assumed this is a typo and should have been 80GB.

The GPUs we used had 8GBs of VRAM, as the task was not very memory intensive. Regarding the exact model of the GPUs, we have used virtual GPUs (i.e. physical GPUs which are further split into smaller virtual GPUs, so that a single GPU can be shared among many users), and therefore do not have the exact model information of the GPUs that we used at the time.

- Figure 3: Do the models use the entire range of probabilities for each class, or is the model always more uncertain about some classes than the others. Have the authors also tried to learn a model for each class individually?

The models always consider all the classes and assign probabilities to all classes. We appreciate the reviewer's suggestion. We have not tried to learn for each class individually but this can be done in a future study.

- l. 143: fine-tuned for which task? Is there a literature reference to the performance of that model?

Thank you for pointing this out. The relevant information regarding the performance and details of the model is given in the URL of the model on Hugging Face (given in footnote 9).

- Section 4.1: Please move before architecture specific sections i.e. 3.1 as this describes important information about the input images that is otherwise missing in section 3.

We appreciate the reviewer for this suggestion. We have now moved the section 4 under Section 3, so it is now Section 3.4 in the revised manuscript. We agree that it now reads better.

- l. 190: “cloud patterns are isolated by bounding boxes in a way that each image contains only one mesoscale cloud pattern”. Describe this in more detail: Has the center-point of the human annotations been used? Were the 70x70 pixel subset restricted to cases where the annotations were covering the entire subset?

We have not used the center-point of the human annotations. We have further clarified this in the revised manuscript.

The corresponding lines in Section 4.1 (Section 3.4.1 in the revised manuscript) are rephrased as: “… to a uniform resolution of 70x70. In particular, for each annotation and the corresponding bounding box defined manually by the human expert, the enclosed region was entirely extracted. Since the bounding boxes differed in size, this resulted in images of regions (cloud patterns) with different sizes, which were then all resized to a fixed size with a resolution of 70x70 pixels (this resolution translates into a physical size of 210-280 km).”

- l. 199: the channel shuffling and focus on structural over chromatic cues is interesting. The infrared channel includes valuable information about the height of the clouds and at least to some extent also the cloud controlling SST. By shuffling the channels this extra information is discarded and the model largely reduced to a monochromatic model. This seems quite harmful, particularly when input images include high clouds. Please comment.

Thank you for this note. As you also stated, we did the shuffling so that models rely more on the morphology, rather than being specifically tied to channels. This does not necessarily lead to RGB channel values being discarded or making the model monochromatic. Moreover, we have used PNG images as input and not the actual SEVIRI channels. As such, shuffling channels in RGB images for detection of objects and patterns is a common practice. In fact, our intention has been to train and test models which solely rely on RGB images to identify cloud patterns without the reliance on any extra information. We agree that if we were to keep the original channels of the SEVIRI, then a more appropriate approach would have been to not shuffle the channels, as individual channels can have information corresponding to different characteristics of the cloud patterns. Moreover, our identified patterns are supposedly all low-level structures. We

acknowledge that this may reduce the scope over which our models can be used in terms of physical parameters.

- l. 229f: How many different setups have been tested to reach an optimal configuration and what has this training optimised for? This is very important as this helps to draw conclusions for the model architecture and not just the model.

Thank you for pointing this out. We have added three more tables to the Appendix, namely tables A3, A4, and A5, to include this information.

- l. 260: please explain in more details how this domain shift is done.

Regarding the domain shift, it arises from the fact that the original dataset and the benchmark dataset were created separately using different tools, and approximately six months apart by the human annotator. In particular, they differ in the size of the bounding boxes as also evident in table 1. We have clarified this in the revised manuscript. Regarding the analysis of the domain shift, Section 7.4 elaborates upon this.
The following lines are added to the beginning of Section 7.4 (Section 6.4 in the revised manuscript): “The domain shift arises from the fact that the original dataset and the benchmark dataset were created separately using different tools, and approximately six months apart by the same human annotator. In particular, they differ in the size of the bounding boxes as also evident in Table 1.”

- l. 324: “confusion in differentiating between Sugar and Flowers…: These are very different mesoscale cloud structures. Is the subset of 70x70 pixels maybe too small, i.e. could it be that a subset contains only the clear-sky region of a Flowers-pattern and therefore detects Sugar? Worth to check out those individual scenes. L. 344 points into a similar direction where the features with self-similarity on a smaller scale are easier detected than those at a larger scale.

Thank you for pointing this out. This was a typographical error on our part and should read as Sugar and Gravel. As you correctly noted, Sugar and Flowers represent distinct mesoscale cloud organization patterns with different morphological characteristics. The manuscript has been corrected accordingly. Also, we had to revise tables 5 and 6 because some of the values of Precision and Recall for CNN and ResNet50 were misaligned with their corresponding cloud patterns.

- l. 410: “annotations that are occasionally incomplete or inconsistent”: please explain. How can annotations be inconsistent with a single annotator? How are these annotations incomplete?

The labeled dataset was created by a single expert annotator with a background in meteorology. A preliminary attempt involving non-expert annotators resulted in misclassification, primarily due to the subtle and often ambiguous distinctions between cloud pattern types. Previous studies

(e.g., Stevens et al., 2020) used annotators with expertise. Using a single expert annotator can ensure higher degree of consistency although we acknowledge, it may introduce subjective bias. The annotator focused on selecting clear, representative examples for each cloud pattern from the SEVIRI images, There are cases where some mesoscale structures within an image may not have been annotated, for example when very small, ambiguous, or partially obscured features were deliberately left unlabelled to avoid over-interpretation.

Due to the use of a single annotator, the variability in the labeling when one uses multiple annotators can not be evaluated and no independent second review was carried out. To compensate for the limited size of the labelled data set and to introduce more diversity, data augmentation techniques such as geometric transformations (e.g., rotations, flips) and intensity variations designed to preserve the physical characteristics of cloud structures were applied. This approach helps the model generalize better and reduces overfitting, thereby mitigating some of the limitations associated with a single-annotator dataset.

We agree that the use of multiple annotators would strengthen the dataset and we consider this an important step for future work.

- l. 413: “benchmark images were overall larger than the original cropped images”: this is mentioned here for the first time. Please clarify at a central place, what the different datasets are: benchmark dataset, original dataset, hold-out test dataset, source data, new benchmark domain. What are their sizes and resolutions? Are they inclusive or exclusive of each other?

Thank you for the question. In fact, we mention this in Section 2.2 and also Table 1. We have added some more clarification to the revised manuscript.
The caption of Table 1 is rewritten as: “The overall statistics of the original and the benchmark data set for each cloud feature. The original and benchmark datasets are separate from each other.”

- l. 440: “possibility to time-aware classification and temporal pattern analysis”:are you thinking of something like Vial et al. 2021?

Yes. Thank you for this reference. We have included this in the revised manuscript.

- l. 447: speculative.

We agree, however, since this is in the research outlook section and we are not making any claims to guarantee the success of this approach, we believe it is worth mentioning.
