2026/06/01

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

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 hyper-parameters 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.

Apart from this, I have a few additional 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 ensured by design that always one of the patterns is present.

- 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.

- The annotation practice needs more details.

    - 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.

    - 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.

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

- 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.

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?

- l. 28: EUREC^{4}A

- l. 29: Barbados **C**loud **O**bservatory

- 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?

- 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.

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

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

- 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.

- 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.

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

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

- l. 75: I assume that the months were chosen to reasonably well 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?

- 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)?

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

- 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.

- 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.

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

- 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?

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

- 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.

- 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 be used? Were the 70x70 pixel subset be restricted to cases where the annotations were covering the entire subset?

- 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.

- 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.

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

- 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 checkout those individual scenes. L. 344 points into a similar direction where the feature with self-similarity on a smaller scale are easier detected than those at a larger scale.

- 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?

- 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?

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

- l. 447: speculative.