Response to Reviewers – "Cell tracking -based framework for assessing nowcasting model skill in reproducing growth and decay of convective rainfall"

Jenna Ritvanen, Seppo Pulkkinen, Dmitri Moisseev, and Daniele Nerini

November 27, 2024

We thank the reviewers for their constructive comments on our submitted manuscript. We have copied the comments of the reviewers in black here and include our response to each individual comment in blue.

Please note that in addition to changes to address the comments from the reviewers, detailed in the replies, the changes listed below were made to the manuscript. We also include a `latexdiff` version of the manuscript to this response that shows the detailed changes from the original submission.

- Updated the L-CNN training subset plotted in Figure 3 and the description of the subset in Section 2.2.4.
- Harmonized the vocabulary in Figure 1 to match the vocabulary used in the text.

Reviewer 1 - Kun Zheng

The authors proposed a cell tracking-based framework for evaluating nowcasting models' ability to predict the growth and decay of convective rainfall. Using the cell tracking-based framework, the authors assess the forecasting ability of different models at the cell scale for local convective systems, rather than focusing on pixel-by-pixel verification. Four nowcasting models (the advection nowcast, S-PROG, LINDA, and L-CNN) are evaluated using Swiss radar data, with L-CNN showing the best performance in forecasting convective cells. However, there are certain issues that need to be addressed:

1. What specific results can be compared between the framework based on cell tracking and existing frameworks to demonstrate its superiority? While the framework assesses models at the cell scale, no comparison is provided with traditional pixel-scale methods. The results mainly focus on highlighting differences among the models but do not demonstrate the advantages of the proposed framework over conventional approaches. Including metrics like MAE, MSE, and SSIM would offer a clearer comparison between the two methods.

   We appreciate the suggestion to clarify the advantages of our cell tracking-based framework over conventional approaches for nowcasting model verification. We agree that the manuscript does not sufficiently highlight these aspects. The main advantages of this framework compared to conventional approaches are

(a) the ability to differentiate the forecast skill based on type of cells (rainfall), e.g. if the cells are growing or decaying,

(b) the ability to separate various aspects of the error according to our interests, e.g. the error in cell (rainfall) area or mean rain rate.

To highlight these points, as suggested, we have added a figure (Fig. 5 in revised version) with CSI and RMSE metrics calculated pixel-by-pixel using a threshold of 4.6 $\mathrm{mm\,h^{-1}}$ (equivalent to the cell identification threshold of 35 dBZ) to make the metrics as comparable as possible. The CSI and RMSE metrics were selected in order to provide metrics for direct comparison with the metrics calculated from the cell tracking -based framework. Pixel-based CSI provides an opportunity for comparison with the different CSI versions used in the framework. RMSE was selected because it penalizes high error values more than MAE (similarly to MSE), but compared to MSE provides the error value in original units. We note that adapting SSIM to the cell-based framework is not straightforward, as it requires calculating the covariance between the possibly different-sized observed and nowcast cells, which is not well-defined.

In addition to the discussion with the new figure in a new Section 4.2, we have added discussion in relevant points to Sections 4.3 and 4.4 to highlight the differences between the information provided by our framework and the conventional frameworks.

2. There is a lack of clear justification or references for many parameter settings in the framework, such as the minimum area threshold for convective cells. This should be addressed with more detailed explanations.

   We have added more justification for the cell identification parameters in Section 3.1. We hope these additions provide a clearer rationale for the parameters used in the framework.

3. It is essential to add a final comprehensive evaluation metric in the framework. Independent evaluation algorithms for cell occurrence and development could hinder broader model applicability. A combined metric reflecting overall predictive capability should be introduced as a final output.

   Thank you for the comment. We agree that adding a final comprehensive evaluation metric would improve the applicability of the framework.

   As the final evaluation metric, we have added a Root Mean Squared Error of the cell volume rain rate where non-existent cells in either the target observations or the nowcasts are considered as zero values (Fig. 6 in revised version). Such error metric penalizes errors in predicting the cell development through incorrectly predicted cell volume rain rate, and errors in predicting the cell occurrence. The cell volume rain rate was selected as the feature of interest because it is impacted by both the size (area) of the cell and the distribution of rainfall inside the cell. Additionally, it is directly associated with e.g. predicted accumulated rainfall. The RMSE value was selected to match the RMSE calculated with the pixel-by-pixel framework (1st comment) to allow comparison between the two.

   The description of the RMSE was added to the methodology in Section 3.4, and the figure (Fig. 6) showing the values was added to the start of section 4.3.

4. Although the significance of cell splitting and merging in convective rainfall evolution is acknowledged, no concrete evaluation methods are provided. The four models tested show limited performance in this aspect, making it difficult to determine if the framework effectively evaluates these processes.

   In the current form of the framework, we do not evaluate how well the models are able to handle splitting or merging. This is because the fraction of cell tracks with splits and merges in the dataset is small (see

Figure 12 / 14 in revised version), explained partly by the emphasis on small cells in the dataset and partly by the configuration of the cell identification and tracking algorithms that aimed to minimise the number of splits and merges. Additionally, removing the splits and merges from the dataset does not alter the results substantially (see the supplementary figures and discussion in Section 4.5). On the other hand, the models evaluated here are also not expected to predict splits or merges in any consistent way, as the blurring occurring in the nowcasts inevitably impacts how the splits or merges are identified. For example, in the identification algorithm applied here, large cells may be split if they contain multiple maxima that are separated by large enough distance or drop in reflectivity. However, the blurring would most likely cause the maxima values and/or the path of least change between the maxima to decrease, which would impact at which time step the cell is considered split compared to observations.

Evaluating the models' ability to reproduce splitting or merging would make more sense for models that are able to produce non-smoothed nowcasts, e.g. generative machine learning models. If one wanted to evaluate how such models reproduce splitting and merging, one could apply a more advanced cell tracking scheme that would handle the splits and mergers of the cells in a more sophisticated way (e.g. [2, 3]). Then one could consider the entire track of the cell, including any child/parent cells resulting from splits or merges, and compare the tracks between the observations and nowcasts.

We have tried to make these points more clear in Section 3.5 and in the discussion in Section 5.

5. Why L-CNN performs better than other models? Analyzing specific factors within the L-CNN model that contribute to this performance could provide insights for improving nowcasting models.

   L-CNN performs most likely performs better than the other models because in it, the local development of rainfall is predicted by a convolutional neural network. The CNN is able to learn non-linear patterns of growth and decay that occur in convective rainfall, and it is also able to produce heterogeneous development inside the cell. This leads to more irregular rainfall values inside the cells that improves forecasting skill. On the other hand, LINDA (that performs second best in multiple aspects) performs Gaussian convolutions to account for the loss of predictability, which leads to homogeneous smoothing inside the cells.

   We have added these points in Section 4.3.3 and in the discussion in Section 5.

6. This framework has difficulty handling complex weather phenomena, especially when multiple weather cells interact in the same area. By only tracking the "most representative" cell and simplifying merge-split operations, the framework's versatility is limited. A more detailed explanation of the framework's applicability is needed from the authors.

   You are correct that the proposed framework has difficulty dealing with complex situations where the cells either merge or split. However, from the analysis we have seen that the number of such cells and their impact on the results is small (please see answer to comment 4).

   How well the framework is able to handle complex weather phenomena depends mainly on the cell identification and tracking algorithm applied. Here, the algorithm is relatively simple, and selected to be such because the models that are being evaluated are not expected to handle complex weather phenomena well either. However, the versatility of the framework can be expanded by applying a more complex cell identification and tracking algorithms. For example, using an algorithm capable of identifying storms hierarchically, e.g. [1], would allow studying also more complex weather phenomena. Although we note that such analysis is sensible only when using models capable of producing non-smoothed nowcasts, since the more complex the identification algorithm, the more blurring in the nowcasts impacts the identification and thus the results.

We have added discussion of this in Section 5.

7. The paper is relatively long, with some redundant sections. Please revising the content in sections "2 Data and nowcasting models" and "3 Cell tracking-based verification framework" . It would be useful to reduce the details on the four models and remove repetitive or non-essential information.

   We thank the reviewer for this helpful suggestion. We have revised the content in sections 2 and 3 to remove non-essential details and reduce redundancy. Please see the Author's tracked changes file for the updates.

Reviewer 2

Thank you for inviting me to review the paper: "Cell tracking-based framework for assessing nowcasting model skill in reproducing growth and decay of convective rainfall" by Ritvanen et al.

The topic of convection is important, as these are often the strongest of storms. Their understanding for the current climate is required, and once verification is achieved of such knowledge, then we can start to project ahead to estimate their statistics for raised levels of atmospheric greenhouse gas concentrations. Furthermore, better assessment for now supports the development of improved weather forecasting systems. Hence this manuscript is useful, starting at the very basic but important level – can we understand short timescale evolution of such storms.

We appreciate the reviewer for recognizing the importance of the topic. We would like to clarify that the primary objective of the manuscript is to evaluate the performance of nowcasting models in predicting convective rainfall, rather than focusing on the detailed understanding of its evolution.

To appeal to the more general non-meteorological reader, it is important that the definitions are set out early. An immediate question is: "Why storm tracking" (which is about predicting into the future) when we are trying to nowcast? Very early on, in the Abstract even, state that "Nowcasting" includes prediction, at least for the short-term ahead. Indeed, also in the Abstract, give a typical timescale for that. Hours?

We have re-written the first sentence of the abstract as *"The rapid temporal evolution of convective rainfall poses a challenge for rainfall nowcasting models that forecast rainfall in time scales ranging from 5 minutes to 6 hours."*

The wording of the Abstract could be tightened. In particular, it would help to make clearer that this is not a paper that develops a tracking methodology, but instead compares four existing tracking algorithms. This should be mentioned earlier in the Abstract, rather than "The framework consists. . . ." (which again gives the impression this paper actually builds a tracking algorithm). Indeed, the title might be better-worded, something like "A comparison of. . . ".

We have re-written the sentence on Line 8 "The framework consists.." to *"In the framework, a cell identification and tracking algorithm is applied first to the input observation rainfall fields, and then separately to the target observation and nowcast rainfall fields where the tracks identified in the input observations are continued."*

There is some ambiguity about the naming of the "four open-source advection-based models" (Line 17). Does "the advection nowcast" have a name, and presumably this is model one? Then models two and three are S-PROG and LINDA. And the 4th models is L-CNN. Ideally there is a consistency of naming and order of model description throughout the entire paper to help the reader.

The advection nowcast refers to a simple Lagrangian persistence nowcast and has no other specific name. We agree that consistent order of the models improves the readability of the paper. In the abstract, introduction and nowcasting model description the models are provided in the order of advection nowcast, S-PROG, LINDA and L-CNN as that provides a logical order for explaining how the models work. Additionally, the advection nowcast, S-PROG and LINDA are available through the pysteps library, while the L-CNN implementation is available online separately. However, later in the results and figures, the L-CNN is shown next to the advection nowcast, as due to the model architecture, L-CNN nowcasts differ the least from the advection nowcast, whereas S-PROG and LINDA are more similar, so placing them side-by-side makes comparison easier.

We have re-written the sentence mentioned from Line 17 "The framework is demonstrated using four open-source advection-based models..." to *"We demonstrate the framework with four open-source nowcasting models: advection nowcast, S-PROG and LINDA models from the pysteps library, and L-CNN model, with data from the Swiss radar network."* to try to clarify it.

The statement of typical timescale would also be useful at the very beginning of the Introduction, i.e. the first sentence. Line 52 talks about lead times of one or two hours.

We have added the timescale of nowcasting (officially from 5 minutes up to 6 hours) to the first sentence of the introduction. We note that the manuscript focuses only on the first hour, since the forecasting skill of the evaluated models is very low after the first hour.

Reading further into the paper, then it becomes clear the authors regard their framework as more of a toolbox of statistical methods to assess storm tracking for short periods ahead. For instance, the caption to Figure 1 implies this. This is all fine, but again, please help the reader to make it clearer earlier in the manuscript what this terminology refers to. The paper is evaluating existing tracking methods.

We would like to clarify that the aim of the framework is to assess grid-based quantitative rainfall nowcasting models that aim to forecast the precipitation intensity in the entire field, not storm tracking -based nowcasting models that aim to forecast the motion of individual storm cells identified from the radar fields. We have tried to clarify this by including for example the mention of quantitative rainfall nowcasting in multiple points in the introduction; please see the provided Author's tracked changes file for the changes.

A key criticism I do have is the presentation in Figure 1, and an opportunity is lost to really illustrate what the paper achieves. First of all, I like the top right part of the diagram that illustrates how observed data is a storm for, say, ... t-2, t-1, t-0. Then storms are observed at t+1, t+2, t+3, t+4. . . Correct? So this is the named "target cell tracks". And then these are compared against the predicted storm features. These are the "Nowcast cell tracks". Hence this part of the diagram makes very clear the approach taken. However, the size needs to be increased, not least so the times "t" and subscripts can be seen.

Once the top right part of the diagram is improved, then the left-hand column makes more sense. The split between "2" and "3" is identical to the split between "2" and "3", top right. Can this be made clearer? Then, as a further enhancement to the diagram, show where the features link to the comparisons. In other words, is there a way to have arrows linking the left to the right? If not, then make it really clear that the "4-5" label on the right-hand side informs boxes "4" and "5" on the left-hand side. Correct?

The caption of Figure 1 could be made much more informative. This is important, because people often extract diagrams from paper .pdfs for use in powerpoint presentations. As far as possible, a diagram and caption should be self-contained.

We thank the reviewer for the constructive suggestions. We have increased the font size in Figure 1, as recommended. Additionally, we have updated the caption and the grouping of the diagrams on the right side to attempt to better convey the correspondence between the flowchart on the left and the diagrams on the right.

Similarly, a little more time spent in Figure 2 would help. What does the outer bound box represent (i.e. is there something special about the domain 41.8N to 49.1N, etc. Is there a good reason why the domain of just one model is shown? (i.e. L-CNN). Please move the colourbar away from the diagram slightly. And please make the font of the tick labels, the colourbar labels and the colourbar title much bigger. Also make the "x" radar locations larger in the diagram.

We thank the reviewer for the helpful suggestions regarding Figure 2. To clarify, the outer bound box represents the area covered by the Swiss rainfall product used as the data in the study. The red box, which represents the domain subset used for training the L-CNN model, is shown only for L-CNN, because it is the only model that has to be trained by using archived data. In L-CNN, the weights in the neural network are trained using the subset; in the other models, the prediction is based solely on a small number of recent observations.

In response to the suggestions, we have re-plotted Figure 2 with the requested adjustments: increased space between the colourbar and diagram, increased fontsize for the text in the figure, and larger "x" markers for the radar locations for better visibility.

Figure 3 could be made really informative, but the presentation is especially poor. I simply cannot see the tick labels on the colourbars, and the individual panel headings I can only just see. As I understand it, the top row is times t-4, t-3, t-2, t-1, t-0, which is radar data used to initialise the tracking algorithms. Then the second row is time t+1, t+2, t+3, t+4, t+5, and this is radar data we hope the tracking algorithms can emulate. Finally, the 3rd-6th rows are the performance of the different tracking algorithms for the same times t+1 tp t+5.

A further issue is that Figure 3 is at a large spatial scale, so to the eye, it is almost impossible to see the changes between times. In other words, almost every panel looks the same. Either zoom in on a particular region of change. Or alternatively, maybe present in colour the edges of change, e.g. blue for where advanced and red for where receded.

The description of the different rows in Figure 3 is correct. To improve the figure, we have increased the font size of the colourbar ticks and the panel headings, as suggested. Additionally, we have zoomed in on a part of the domain and removed one timestep in order to make the individual panels larger; hopefully this will make the figure easier to understand and more visually effective.

I will assume all the statistical analysis is robust and correct, but again, it is important to present the findings well. Figure 4 appears as a key diagram, as histograms of continuing, stopping and starting cells with advective storms. But is this purely data, or are the four models involved somehow? I'm guessing it is purely data, while Figure 5 shows, in a similar way, changes for the four models. I like the format of presentation, but again, the legend for Figure 5 is difficult to see.

The histograms in Figure 4 are purely observed data: cell feature values taken at the last observed timestep of each cell track. We have updated the figure caption to make this more clear.

Additionally, we have increased the legend font size in Figure 5 (Figure 7 in revised version) and all other figures with similar legends.

This paper makes a useful contribution to understanding very short-term forecasting, and I am happy to re-review the manuscript. I can envisage the paper being published, but the authors really need to spend more time on the presentation. It is the underlying science that matters the most of course, but when presentation falls below a threshold, there is a real risk that ambiguities or misunderstandings will occur by readers.

We greatly appreciate the reviewer's positive comments and recognition of the contribution of our work to understanding short-term forecasting models. We also value the feedback regarding the presentation of the manuscript. We agree that clear and effective presentation is crucial for ensuring that the underlying science is communicated accurately, and we have made revisions throughout the manuscript to improve clarity and reduce any potential ambiguities. We hope these changes will address the reviewer's concerns and enhance the overall readability of the paper.

References

[1] Jinyi Hou and Ping Wang. "Storm Tracking via Tree Structure Representation of Radar Data". In: *Journal of Atmospheric and Oceanic Technology* 34.4 (Jan. 25, 2017), pp. 729–747. ISSN: 0739-0572. DOI: 10/f93gwb. pmid: null.

[2] George Limpert, Adam Houston, and Noah Lock. "The Advanced Algorithm for Tracking Objects (AALTO): Advanced Algorithm for Tracking Objects". In: *Meteorological Applications* 22.4 (Oct. 2015), pp. 694–704. ISSN: 13504827. DOI: 10/f7z6jx.

[3] Beilei Zan et al. "Solving the Storm Split-Merge Problem—A Combined Storm Identification, Tracking Algorithm". In: *Atmospheric Research* 218 (Apr. 1, 2019), pp. 335–346. ISSN: 0169-8095. DOI: 10/gf52zh.