Zhang et al. introduced the CNN-BiLSTM-AM model for convective weather prediction in China's Henan region. Their findings indicate that CNN-BiLSTM-AM outperforms traditional machine learning models like Random Forests (RF), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), as well as a physical model like the Weather Research & Forecasting Model (WRF). Additionally, the authors employed the RF algorithm to assess input importance in the CNN-BiLSTM-AM model, finding that the resulting rankings align with meteorologists' subjective understanding.

I have a few major comments on the manuscript as follows:

1. The current structure and grammar of the manuscript make it very hard to follow:
    a. Why are sections 3: Deep learning models (it should be plural) and 4.1: Evaluation method outside of Section 2: Data and method?
    b. The manuscript should be revised by a native speaker to rectify grammar and phrasing errors. For example, the first sentence of the abstract has no meaning: (i) it's either "we developed" or "this work presents," (ii) the second part of the sentence is not relevant.
2. The Introduction section is incomplete:
    a. The authors should cite the recent Google's GraphCast model (https://www.science.org/doi/10.1126/science.adi2336) as well.
    b. Based on my humble understanding, "forecaster" is a human who makes forecasts about different convective storms, right? If so, what tools does the "forecaster" use for forecasting? I couldn't understand the context of the first paragraph.
    c. Physical-based models like NCEP GFS and WRF need discussion alongside deep learning approaches.
3. The current Section 2 is incomplete:
    a. The authors compared CNN-BiLSTM-AM with other ML models and the WRF model. At the very least, those models should briefly be mentioned in Section 2 as well.
    b. The authors should provide the websites or links to all the datasets for reproducible purposes.
    c. The comma sign in Table 1 is not commonly used.
    d. Lines 163-205 and Figures 3 and 4 are too trivial. The sentences are just repeating from the figures. However, information about the forecasting timestep and the training loss is missing.
    e. Line 227: In my experience, 30 epochs for training is very few. Why did the authors not train more? Was the loss converged?
    f. The training data is available in a 6-hour timestep; how could the authors configure the CNN-BiLSTM-AM so that it can predict every hour?
4. The Section 4 is incomplete:
    a. Figure 8: What are the shaded orange graphs at the bottom?
    b. Figures 8, 9, and 10: the text inside the plot is very small.
    c. Figure 10: What are the spatial resolutions of the predictions?

d. Figure 10: Where and how did you get the results from the human-forecast?

5. I was confused about the whole approach of the manuscript. First, comparing a deep learning model (CNN-BiLSTM-AM) with traditional machine learning models is not a fair comparison. Why didn't the author compare their approaches with current deep learning models that they referenced, such as ConvLSTM or Pangu-Weather? Second, results suggest CNN-BiLSTM-AM has better performances than RF. Why did the authors use a less effective model to explore the importance ranking of inputs for the better model?

Based on these serious flaws, especially the last point, I would recommend rejecting the manuscript.