Reviewer 2

In this work, author use deep learning algorithms to develop a framework including CNN, BiLSTM, and AM for convective weather forecasting, called CNN-BiLTS-AM method. The approach is novel. The NCEP FNLS analysis datasets were exploited as inputs in which several factors related to meteorology, convective physical quantities, and geographical variables were selected by experts. In addition, station measurements were collected and used for model developments and validation purpose. The training and testing datasets were from 2015 – 2020 in which data of randomly selected 72 days use for testing and the remained data used for training. In order to evaluate the proposed method's performance, different machine learning methods such as KNN, RF, GBDT, SVM, and a physical model (WRF) were implemented, and results were compared in testing dataset and an individual case. The comparison results showed that the CNN-BiLSTM-AM overperformed other algorithms in the test cases.However, I have major comments as follows:

Response: We gratefully thank you for taking the time to provide constructive comments and helpful suggestions for our manuscript, which have significantly raised the quality of the manuscript and enabled us to improve the manuscript. Each suggested revision and comment was accurately incorporated and considered. Below the comments are response point by point.

Comment 1: The problem statement is not clear. Convection weather forecasting is related to many events. As I understand, evaluation is mainly focused on forecasting of precipitation rate (section 4.2, 4.4, 5.1) and precipitation occurrences (section 4.3). In fact, it is two different problems: classification and regression. In this work, please state clearly what is output?

Response: Thank you very much for your detailed comments.

By using the ERA5 hourly data and observation data, our study employs DL algorithms to establish a forecasting model for severe convective weather (SCW) called CNN-BiLSTM-AM. The model is capable of generating 0-6 hour short-term precipitation prediction, which is a regression problem. In the following time, we will carefully review the full text and correct any ambiguities.

Comment 2: It is also not clear that author develop one, two, or many models to solve the given problems and how to build models. There is no problem with a general deep learning framework as presented in section 3.1. However, loss functions, which are very important in DL, depend on output designs and number of models. How many models were developed and how were corresponding outputs and loss functions in this work?

Response: Thank you very much for your detailed comments.

The loss function selected for minimization during training was mean squared error (MSE). The formula is as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i' - y_i)^2 \quad (1.3)$$

Where m represents the training sample size, y_i represents the actual value, y_i' represents the predicted value. Statistical measures, including correlation coefficient (r), standard deviation (σ_n), and root-mean-square error (RMSE) were used to assess model performance.
Next, we will introduce the current well-established deep learning network models such as ConvLSTM, Predrnn++, CNN, FC-LSTM, LSTM, to compare with the CNN-BiLSTM-AM model, and re-conduct the experiments and analysis. The loss functions of these models will be elaborated in subsequent research work.

Comment 3: There is also inconsistence in construction of training and testing datasets which is key points for deep learning/machine learning algorithms. In section 3.2, authors only mentioned to classification problem. In addition, it is not clear what is used as ground truths to label a grid as positive/negative sample. If all 2400 station measurements in China were collected, the model should be built for all China. However, in experimental area is Henan region (section 2.3) with only 12 stations. Were authors use station precipitation measurements to label positive/negative samples? How many stations were used and its location and distribution? Please report also detail number of training and testing samples for this case? (line 223-225)

Response: Thank you very much for your detailed comments.

In this study, observation data were used to label the predictors. These observation data were collected from 2,400 ground stations across the country, but in this study, in order to improve operational efficiency, we only selected the observation data of 119 stations related to the experimental area of Henan for the construction of the database.

Comment 4: The dividing of training and testing datasets are not suitable for this problem. Forecasting is for future, so the use of one- or two-year data for testing is more independent. For example, data from 2015 – 2020 can be divided into training dataset (2015 – 2019) and testing dataset (2020). The current division based on randomly selection can bring very good results on modelling but not correct for future and independent datasets.

Response: Thank you very much for your detailed comments.

We will reconstruct the training and testing datasets using real-time data from January to December between 2015 and 2020, as well as ERA5 hourly data, with 2015-2019 as the training dataset and 2020 as the testing dataset , and then re-experiment and analyze.

Comment 5: In section 3.2, the questions mentioned above are also rising for training and testing datasets for regression problem (precipitation estimation). Author needs to add this information to this section.

Response: Thank you very much for your detailed comments.

The model is capable of generating 0-6 hour short-term precipitation prediction, which is a regression problem. In the following time, we will carefully review the section and correct any ambiguities.

Comment 6: In section 4.1, many evaluation indexes for regression were not defined. Example: R2, RMSE.

Response: Thank you very much for your detailed comments.

We will add the missing evaluation indexes as soon as possible.

Comment 7: In section 4.2, the test dataset is from 2015 – 2017 (line 243). It is different from the description of training/testing datasets in Section 3.2 (2015 – 2020, line 222). Also, please point out the predicted parameter is hourly (or daily) precipitation and forecast duration. The observation data is collected from how many ground stations and where they are.

Response: Thank you very much for your detailed comments.

In order to maintain the consistency of the article, we will unify the training dataset and the testing dataset in this section. The prediction parameter is hourly precipitation for the next 0-6 hours. Since the experimental area of this study was selected in Henan, China, the observation data used were from 119 ground stations in Henan.We will add the above information timely.

Comment 8: In section 4.2, in Figure 8, the boxplots were used to present predicted precipitation. However, this description cannot point out the outperformance of the proposed method and RF because they are likely to each other. Moreover, there is confliction of results in Figure 8 and Figure 5. In Figure 5, the model is underestimated precipitation while in Figure 8, the model is overestimated precipitation. Please explain why.

Response: Thank you very much for your detailed comments.

Figure 5 presents a comparative analysis of predicted and observed precipitation on convective weather test dataset from 2015 to 2017. Figure 8 presents a comparative analysis of predicted and observed precipitation at 12 stations in Zhengzhou on July 22, 2022.The difference between the two is that the time scale is different, the object of the analysis is different, and the results obtained are different. From the above results, it can be seen that the performance of the model in long-term forecast timeliness is different from that in short-term forecasts.

Comment 9: In figure 9, it is not clear that diurnal variation is calculated at one station, or averages on 12 stations in Henan. If the dataset is the same as in Figure 8, there is disagreement on results of boxplot and line chart which present performance of CNN-BiLSTM-AM. In Figure 9, error of the proposed method is almost equal to 0 while in figure 8, there is difference of prediction and observation presented by boxplot. Please explain why.

Response: Thank you very much for your detailed comments.

Figure 8 presents a comparative analysis of predicted and observed precipitation at 12 stations. Figure 9 depicts the diurnal variations offered by diverse models in July 2022. The main difference between the two is the time scale. Figure 8 depicts the distribution of total precipitation at 12 stations over the process of the event on July 22, 2022, while Figure 9 depicts hourly variation over the process, so the results of the analysis will be different.

Comment 10: In Figure 10, the RF (b) should have lower error than CNN-BiLSTM-AM (a) based on color scales, and it is conflict with above conclusions. Following this concern, why RF is not used for stability analysis in Section 5.1.

Response: Thank you very much for your detailed comments.

In Figure 10, RMSE value of the CNN-BiLSTM-AM model is mostly between 0.11mm and 3.87mm. RMSE value of the RF model is mostly between 0.10mm and 4.25mm. Judging by the results, the performance of the CNN-BiLSTM-AM model is better than that of RF, and is not conflict with above conclusions. Secondly, we have used RF for stability analysis in Section 5.1.