Comments on the manuscript entitled “A General Comprehensive Evaluation Method for Cross-Scale Precipitation Forecasts” by Zhang et al. submitted to GMD

The authors have made revisions to the manuscript, including the addition of new sections and figures. However, I still have some minor comments for the authors’ consideration:

1. L82: Please consider removing “shortly after”.

Response: Thank you for alerting us to this. The statement has been modified as follows:

P4 L82-83: “*Gilbert (1984) proposed two scoring methods, namely, the ratio of verification and the ratio of success in forecasting.*”

2. L115-116: Could you clarify the intended meaning of “The reasons for which, although varied, are worthy of attention, but include its objectivity and practicality”?

Response: Thank you for alerting us to this. The statement is not accurate and has been modified as follows:

P6 L115-116: “*Although the reasons for this are varied, its objectivity and practicality merit attention.*”

The objectivity of the TS refers to the objectivity of its calculation method, while practicality refers to its simplicity, ease of use, and suitability for promotion. The sentence aims to convey that despite numerous factors contributing to the TS's consistently advantageous position, these two notable characteristics merit our attention. Consequently, in devising new scoring methods, it is essential to account for the advantages of the TS scoring method.

3. L127: Could you provide clarification regarding the spatial range of medium and small-scale

systems? Are you referring to “medium-scale” or “meso-scale”?

Response: Thank you for alerting us to this. The statement is not accurate, and it refers to the mesoscale, which has been changed from "modium" to "meso" in the text.

P6 L127: “meso- and small-scale”.

P6 L129: “meso- and small-scale”.

P7 L143: “meso- and small-scale”.

P8 L173: “small- to meso-scale”.

4. L135: Please elucidate the meaning of the sentence ‘when a forecast that “looks good” is not as good as one that “looks bad”’.

Response: Thank the expert for the comment. With the development of high-resolution numerical weather forecasting, some meso- and small-scale phenomena have been portrayed by models. However, even if the spatial characteristics of forecasts are similar to those of observations, when there is a small deviation in the timing and location of events between a forecast and an observation, both “false alarms” and “missed alarms” will occur, leading to a score lower than the subjectively expected result.

As an example mentioned in Ahijevych et al. (2009) (Fig. 1), for the comparison of geom001 and geom005 predictions, although the geom001 forecast is consistent with the observation in terms of morphology, there is a deviation of displacement, and no overlap between the forecast and observation, then the TS score is 0. On the other hand, the geom005 forecast and observation are quite different, but due to the overlap between the forecast and observation, the TS score is 0.11. Of course, the article also points out that even if modelers and other users believe that geom005

predictions are very poor, a hydrologist might actually prefer geom005.

In conclusion, a larger TS value does not necessarily indicate a better overall forecast. There will be situations where the forecast looks good but the TS score is poor.

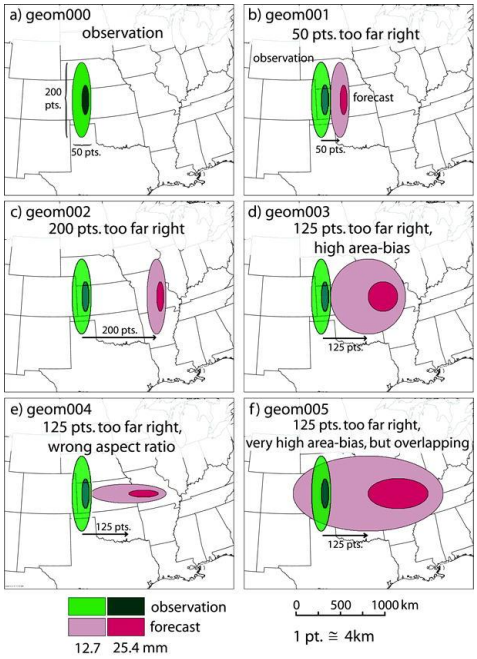

Figure 1. (a)–(f) Five simple geometric cases derived to illustrate specific forecast errors. The forecasted feature (red) is positioned to the right of the observed feature (green). Note, in (f) (geom005), the forecast and observation features overlap. One grid box is approximately 4 km on a side.

References:

Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E.: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts, Weather Forecast., 24, 1485–1497, https://doi.org/10.1175/2009WAF2222298.1, 2009.

5. L159-163: Could you specify the context in which you are referring to “using the traditional skill score”?

Response: Thank the expert for the comment. Simple upscaling method adjusts the high-resolution forecast and observation information to a larger scale, and then using the traditional skill score. Here “using the traditional skill score” specifically refers to the threat score (TS), and also includes

equitable threat score (ETS), Frequency Bias (BIAS), False Alarm Ratio (FAR), Missing Ratio (MR),etc.

6. L198-199: What is meant by “base rate independence”?

Response: Thank the expert for the comment. The base rate is a descriptive statistical term used in categorical events for deterministic forecasting, representing the probability of an observed event occurring. The base rate, alternatively referred to as observational probability or climatological probability, serves as a fundamental descriptive index. It is not a performance indicator. Independent of the base rate is an ideal attribute of scoring indicators, because the scoring does not change with climate change (Yule, 1912). The designed indicators should be suitable for both the rainy and dry seasons within the same region. Also they can be applied to various regions, including arid and semi-arid areas, as well as semi-humid and humid areas.

The concept of "base rate independence" in the manuscript, which is not clearly expressed in the current context, has been changed to "independent of climatological probability" to enhance readability.

P9-10 L198-200: “*The designed scoring performance indices should possess ideal attributes such as fairness, independent of climatological probability, suitability for extreme events, and boundedness as much as possible.*”

References:

Yule, G. U.: On the methods of measuring association between two attributes, Journal of the Royal Statistical Society, 75, 579-652, 1912.

7. L198-199: What is meant by “The devised scoring method should be easy to promote”?

Response: Thank the expert for the comment. “The devised scoring method should be easy to promote” means “The new scoring method should be conceptually clear, pragmatic, and operational, with broad applicability across various regions and seasons.”

8. L216-221: Please consider rephrasing the lengthy sentence for better readability.

Response: Thank the expert for the suggestions. The statement has been modified as follows:

P10-11 L216-221: “*To address the issues of “distorted scores due to the division of precipitation thresholds and increased subjective risks brought about by the setting of the neighbourhood spatial verification method” in traditional and improved precipitation scoring methods, this study refers to the verification method for heavy rainfall forecasts based on predictability (Chen et al., 2019) and combines the advantages of relative and absolute errors. A GCEM is constructed by directly analyzing the proximity of forecasted precipitation to observed precipitation.*”

9. L240, Eq.3: The authors argue that EPI indicates the excessive precipitation index. Thus, u should not be equal to x. However, I disagree. EPI equals 0 when u=x, indicating no excessive precipitation. Including the case u=x in Eq.3 might be beneficial. Assuming one wants to evaluate precipitation forecast using observed precipitation data with missing values, the EPI will be missing in locations/times where no observation is available. EPI is also missing when u=x if Eq. 3 does not include the case of u=x. How do you propose to distinguish between these cases when EPI is a missing value?

Response: Thank the expert for the comment. Through the expert analysis, we have learned that the expert have different perspectives on our definition of EPI. Indeed, as the expert has pointed out, in the current situation ($u \neq x$), when EPI is missing, it cannot be determined whether it is caused by $u=x$ or because u itself is missing. In fact, in the current situation, when EPI is missing, it can only indicate cases that do not belong to ($0<u<x$), which may include cases where u is missing, $u=x$, or $u>x$. If Eq. 3 is set to $u=x$, then Eq. 2 should also be set to $u=x$. This will result in duplicate calculations, where $u=x$ needs to be calculated both in EPI and in IPI. Another reason is that we want to calculate the degree of over forecasting and under forecasting when there are biased forecasts. Therefore, we do not define $u=x$ within the calculation range of EPI or IPI.

10. L246, 250: Eq. 4 combines Eqs. 2 and 3. However, Eqs. 2 and 3 do not include $u=x$, whereas Eq. 4 does. Additionally, I still believe it may not be necessary to define IPI and EPI. It seems analogous to already having mean error; hence, defining a positive mean error index and a negative mean error index may be unnecessary. However, I am open to retaining IPI and EPI if the authors find it necessary. Nevertheless, removing IPI and EPI could help in sharpening the focus of the paper.

Response: We extend our gratitude to the expert for agreeing to retain IPI and EPI in the case of different opinions.

11. Figs. 1-4: The font size of the x- and y-axes is too small, especially for Fig. 3. Please consider using a larger font size.

Response: Thank you for alerting us to this. Figures 1-5 have been adjusted appropriately.

12. Code and data: I suggest saving the source code and data separately instead of combining them into one large file. This would facilitate easier downloading of the source code. Additionally, the data should be saved in individual files rather than being divided into parts within a single RAR file. This current setup requires users to download all files before the data can be uncompressed, which could be inconvenient.

Response: Thank the expert for the suggestions. The code and data have been stored separately, and the data is stored in individual files.