Reply to Referee comment 1

Thank you for your thorough review and valuable comments. Below are our detailed replies to each of your comments. We believe that these revisions have greatly strengthened the quality and readability of our manuscript thanks to your suggestions.

Comment: My first concern is about the temporal partitioning of the CNN model training. The paper mentions the CNN is trained on a 20-year data using cross-validation, yet the specific years assigned to each fold should also be elaborated. Since the final streamflow evaluation covers the period from 2009 to 2012, there is a risk of data leakage if any of those years were included in the training set. It is also unclear why the calibration periods for the GBEHM and ARX components differ from one another and from the CNN training period.

Reply: We sincerely appreciate your insightful comments and apologize for the confusing terminology used in our initial manuscript. We would like to clarify that our use of the term "cross-validation framework" was an inaccurate description of our methodology. Upon careful verification, we found that we actually employed a fixed temporal partitioning approach for the 20-year (1997–2016) UKMO hindcast dataset, rather than k-fold cross-validation. Specifically, the dataset was divided into a training set (1997–2008, 12 years), a validation set (2013–2016, 4 years), and a test set (2009–2012, 4 years). This strict chronological separation completely eliminates any risk of data leakage and ensures a robust evaluation.

The differences in the calibration periods for the GBEHM, CNN, and ARX models are primarily dictated by the characteristics of each model and the availability of their respective input data. For the CNN model, the training period is constrained by the availability of the UKMO hindcast dataset, which provides a 20-year record for a specific model version. By allocating a standard 60% of the total available record for model training, the 12-year period from 1997 to 2008 was designated as the training set. As a physically-based distributed model with high structural complexity and numerous parameters, the GBEHM requires a long-term data sequence to capture various hydrological conditions. Because the GBEHM is calibrated using ground-based gauge observations, we were able to use an earlier and longer calibration period (1990–2005). Finally, the ARX model requires GBEHM-simulated streamflow as input. We avoided using the early stages of the GBEHM simulations to prevent the influence of warm-up instabilities and initial condition uncertainties on streamflow simulation. We finally selected the 2000–2008 period to balance the need for high-quality, stabilized simulation data with a sufficient sequence length for reliable parameter estimation. We are sincerely grateful for these constructive comments and have added a concise explanation into the revised manuscript.

Comment: Second, the CNN generates probabilistic precipitation forecasts via the CSG distribution, which is a well-motivated design choice. However, this statistical information seems to be discarded before feeding to the hydrological model. Is this an intended choice and why did authors choose that? For a system intended for hazard early warning, I think including uncertainty information through the modelling chain would greatly enhance its reliability.

Reply: Thank you for this insightful comment. We appreciate the opportunity to clarify this intended design choice. Converting the probabilistic CSG distribution into a deterministic input is a practical trade-off between maintaining an extended forecast lead time and preserving full probabilistic

uncertainty. The GBEHM is a highly complex, fully distributed model that explicitly solves computationally demanding physical processes, driving it with a full probabilistic ensemble for a 60-day lead time would incur prohibitive computation time, therefore delaying forecast issuance, shortening the effective lead time, and ultimately compromising the value of early warning. However, the statistical information is not completely discarded. By using the mathematical expectation of the predicted CSG distribution as the deterministic input, we maximize the retention of its statistical properties, explicitly preserving the heavy-tailed footprint of extreme events. Nevertheless, we gratefully acknowledge this limitation and have highlighted the development of fully probabilistic streamflow forecasting as a critical direction for future research in the revised manuscript.

Comment: Third, more evaluation metrics targeting hazards (i.e., the topic of this paper) could be introduced. NSE and MSE are often dominated by baseflow conditions and do not necessarily reflect a model performance during extreme events. I recommend the authors select representative flood events from the 2009-2012 evaluation period and present event-scale forecast performance, such as peak flow errors to manifest the model ability for hazard warning.

Reply: We appreciate this highly constructive suggestion. Following your advice, we evaluated peak flow errors for representative flood events during the 2009–2012 period. However, we found that while the framework maintains high overall consistency over a 60-day horizon, its performance in capturing absolute peak flow magnitudes is currently suboptimal. This is primarily because the optimization objectives of our CNN model and ARX model is to minimize global residuals across the entire time series. Consequently, the framework lacks specific optimization mechanisms or weighted loss functions specifically targeted at extreme events. We have acknowledged this suboptimal performance regarding extreme events in the revised manuscript and proposed incorporating hazard-specific objective functions as a critical future improvement.

Comment: Finally, the paper would be much stronger if it provides more technical detail on the GBEHM calibration and its operational feasibility. There is currently little information on which specific parameters were tuned. The study states “In this application, the UYRB is discretized into an 8 km × 8 km grid system and further delineated into 479 sub-basins based on the DEM”, which makes me confused. Is the model grid-based or sub-basin based? More details could be provided. The manuscript would also benefit from a brief discussion of operational feasibility, for example, if the modelling system can be applied in operation, and to achieve that what are the challenges and possible solutions, etc.

Reply: We are grateful for your constructive feedback. We have provided more technical details that you requested point-by-point in the revised manuscript.

1. GBEHM calibration details: We apologize for the omission of the specific calibrated parameters. In the revised manuscript, we list the parameters tuned during the calibration period. These primarily include key parameters governing evapotranspiration, runoff generation, groundwater, and snowmelt, such as the evaporation parameters (C_1, C_2, C_3), soil saturated hydraulic conductivity (K_s), groundwater transmissivity (K_g) and storage coefficient, the snowmelt factor (M_f), and the hillslope shape parameter (f_{ss}). We have expanded Sect. 3.3 in the revised manuscript to list the parameters.

2. Spatial discretization: We apologize for the confusing phrasing. As mentioned in Section 3.3, GBEHM employs a hierarchical sub-grid parameterization scheme. The study region is first

delineated into 479 sub-basins based on the DEM to construct the river network topology for flow routing. Within each sub-basin, the landscape is first discretized into the 8 km × 8 km grid system to integrate meteorological forcing data and capture the spatial heterogeneity of land surface properties, and then further subdivided into hillslope-valley units. Thus, the model solves vertical water and energy balances at the hillslope scale and dynamically aggregates the runoff to the sub-basin scale for lateral river routing. We have rewritten this part in Sect 3.3 to clearly convey this hierarchical structure.

3. Operational feasibility: We appreciate the suggestion to discuss the operational prospects. The proposed framework is practically applicable for real-time forecasting. In operational practice, since the models (CNN, GBEHM, and ARX) are typically pre-trained and calibrated in advance, the total computational time required to process the input data and generate a 60-day forecast is well within one hour on a standard personal computer. The primary operational challenge is the data latency in the public S2S database, where UKMO forecasts have a 21-day delay. To achieve true real-time early warning, operational agencies could acquire data directly through official institutional agreements or substitute UKMO with lower-latency alternative models like CMA. We have added a brief discussion about these aspects into the final paragraph of the conclusions.

Reply to Referee comment 2

Thank you for your thorough review and valuable comments. Below are our detailed replies to each of your comments. We believe that these revisions have greatly strengthened the quality and readability of our manuscript thanks to your suggestions.

Comment: Given the elevation and complex topography of the study area, it would be useful to have a visualization of the gage density of the CGDPA observational product in this region. Are the authors concerned about the accuracy of the GCDPA product in gage sparse regions of the basin? Additionally, the mismatch between the GBEHM resolution (8-km) and the meteorological forcing (25-km) may be one source of error in streamflow forecasts.

Reply: We sincerely appreciate these insightful comments. Regarding the request for a visualization of gauge density, we have noted in Sect. 2.2.1 and would like to direct readers to the previous work by Shen and Xiong (2016), which provides exhaustive spatial distribution maps and station data for the over 2,400 national gauges utilized in the CGDPA product. While we acknowledge inherent uncertainties in high-elevation and gauge-sparse regions, CGDPA remains a highly reliable dataset available for China and has been widely used as reference dataset in the literature (Shaowei et al., 2022; Lu and Yong, 2020; Wei et al., 2019). Furthermore, we fully agree that the spatial mismatch between the 25-km forcing and the 8-km hydrological grid is a non-negligible source of error, as it may smooth out localized precipitation details. We have clarified in Sect. 5.3 that the hydrological model error (MSE_m) incorporates the errors introduced by precipitation data interpolation.

References

Lu, D. and Yong, B.: A preliminary assessment of the gauge-adjusted near-real-time GSMaP precipitation estimate over Mainland China, Remote Sens., 12, 141, https://doi.org/10.3390/rs12010141, 2020.

Shaowei, N., Jie, W., Juliang, J., Xiaoyan, X., Yuliang, Z., Fan, S., and Linlin, Z.: Comprehensive evaluation of satellite-derived precipitation products considering spatial distribution difference of daily precipitation over eastern China, J. Hydrol. Reg. Stud., 44, 101242, https://doi.org/10.1016/j.ejrh.2022.101242, 2022.

Shen, Y. and Xiong, A.: Validation and comparison of a new gauge-based precipitation analysis over mainland China, Int. J. Climatol., 36, 252-265, https://doi.org/10.1002/joc.4341, 2016.

Wei, L., Jiang, S., Ren, L., Yuan, F., and Zhang, L.: Performance of two long-term satellite-based and GPCC 8.0 precipitation products for drought monitoring over the Yellow River Basin in China, Sustainability, 11, 4969, https://doi.org/10.3390/su11184969, 2019.

Comment: The precipitation bias correction is clearly impactful; this manuscript would benefit from a greater understanding of what relationships the CNN is capturing better than traditional statistical models.

Reply: Thank you for this specific comment. Unlike traditional point-to-point statistical models (e.g., quantile mapping), our CNN model excels by capturing two key relationships. First, it extracts spatial dependencies from surrounding atmospheric conditions using a 9×9 grid neighborhood. Second, it models the complex non-linear interactions between 20 multi-level meteorological predictors and local precipitation. We have expanded the discussion on these specific mechanisms in Sect. 5.1 of the revised manuscript to provide better interpretability.

Comment: The following step from Line 230 is unclear and could benefit from further explanation or a figure: “the model generates a deterministic forecast by constructing a large-scale pseudo-ensemble from the predicted CSG distribution at equal quantiles and calculating the ensemble mean.”

Reply: We sincerely appreciate the reviewer for pointing out this ambiguity. The step is essentially a quantile-based discretization of distribution and the detailed procedure is as follows:

1. Based on the three parameters (γ, μ, σ) output by the CNN, the cumulative distribution function (CDF) of the predicted CSG distribution is established.

2.We discretize this CDF into N equal probability intervals (in this study, we set $N = 10000$, corresponding to percentiles from 0.00005 to 0.99995).

3. We apply the inverse CDF (quantile function) at each probability point to extract 10000 discrete precipitation values, thereby forming a "pseudo-ensemble" of 10000 members.

4. Finally, the arithmetic mean of this pseudo-ensemble is calculated to serve as the final deterministic precipitation forecast.

This sampling approach integrates low-probability extreme values from the heavy tail into the ensemble mean to mitigate the smoothing effect. We have expanded the detailed explanation of this procedure in Sect. S1 of the revised supplement to improve clarity.