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 the planned revisions will significantly enhance the quality and readability of our manuscript.

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. To completely eliminate the risk of data leakage and ensure a more robust evaluation, we decide to move away from the cross-validation approach in the revised manuscript. Instead, we will implement a fixed temporal partitioning for the 20-year (1997–2016) UKMO hindcast dataset. Specifically, the training set is 1997–2008 (12 years), the validation set is 2013–2016 (4 years), the test set is 2009–2012 (4 years), thereby avoiding any risk of data leakage.

The differences in calibration periods for GBEHM, CNN, and ARX are primarily dictated by the characteristics of each model and the availability of their respective input data. For 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% ratio of the total available record for model training, the 12-year period from 1997 to 2008 is designated as the training set. For GBEHM, as a physically-based distributed model with high structural complexity and numerous parameters, it requires a long-term data sequence to capture various hydrological conditions. GBEHM is calibrated using ground-based gauge observations, allowing for an earlier and longer calibration period (1990–2005). For ARX model, it requires GBEHM-simulated streamflow as input. We avoid using the early stages of GBEHM simulations (e.g., the 1990s) to prevent the influence of warm-up instabilities and initial condition uncertainties on streamflow simulation. We finally select 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 will add 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 will highlight 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. We will introduce event-scale metrics such as peak flow errors to representative flood events from the 2009–2012 period, explicitly demonstrating the capability of our framework in capturing peak flows across different lead times.

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 will provide 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 will 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}).

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 will rewrite this sentence 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 will add a dedicated paragraph discussing these aspects.