GMD egusphere-2026-4305 – Author responses to referees

Dear Dr Woldemeskel,

We thank you for your positive assessment and for the insightful and constructive comments. We respond to each point below. Several changes also address comments from RC2, and we cross-reference them where relevant.

1 – Robust assessment of discharge and flood inundation

Referee comment. *While the manuscript focuses on the design and technical developments, a robust assessment of model outputs, particularly discharge and flood inundation, is missing, other than the application for selected case studies. This would be valuable to provide greater confidence in the forecast performance in different regions of the world. Such analysis would also help identify areas of strength and weakness of the modelling/forecasting system.*

Author response. We thank the referee for this comment. We agree that readers need a quantitative context to judge the strengths and weaknesses of the system. We will therefore strengthen the evaluation of the long-term simulation in two ways:

1) For discharge, Sect. 4.1 will be expanded. The evaluation will be based on the reference simulation produced for the 2025 WMO State of Global Water Resources report, at the 708 GRDC stations (out of 898 classified as having no major reservoir or lake influence) with at least 5 years of observations in 1991–2020. The expanded section will show the distribution of the KGE and its components (correlation, bias ratio, variability ratio) at these stations, mapped and summarised by region or climate zone, showing where the system performs well or poorly. We will also discuss the likely causes of lower skill: the lack of calibration, errors in the precipitation forcing, runoff partitioning, the absence of a groundwater reservoir, river parameters, and human water use not covered by the reservoir screening.
2) For flood inundation, we will add a quantitative comparison for the Pakistan case (Sect. 4.4). We will compare the simulated inundation extent with Sentinel-1-derived flood maps for the 2022 event at the available satellite acquisition dates, using hit rate, false alarm ratio, and critical success index (CSI). We will additionally compare the observed and simulated evolution of total inundated area through the expansion and recession of the flood.

Further, as also explained in our response to RC2, the paper documents the development and implementation of river routing coupled within the IFS. The cases in Sects. 4.2–4.4 were chosen to illustrate the range of applications that this coupling now enables, not to quantify forecast skill. This is a first technical implementation, and several parameterisations are still being developed (e.g. runoff generation at high resolution, Sect. 4.6.3). A systematic evaluation of forecast skill for discharge and inundation in the operational configuration across regions and lead times would need its own experimental design and is beyond the scope of this technical paper. It will be reported in future work. We will make this clear in the revised manuscript.

Planned changes to the manuscript:

- Explain the scope of Sects. 4.2–4.4 explicitly: illustrative applications, not an assessment of forecast skill. Mention the planned future evaluation (see RC2).

- Sect. 4.1: expand with the KGE and its components at the 708 selected GRDC stations (1991–2020), mapped and summarised, with a discussion of strengths and weaknesses (see RC2, [M7]).
- Add a short synthesis of regional strengths and weaknesses and their likely causes.
- Pakistan case: add a quantitative comparison of simulated inundation extent with Sentinel-1-derived flood maps for the 2022 event at the available satellite acquisition dates.

2 – "Realistic" flood forecast (L585) and errors in Fig. 7d–f

Referee comment. *In addition, the statement in line 585 that the flood forecast is 'realistic' requires further justification. It appears there is a noticeable timing error in Figure 7d-f. The model also underestimated the subsequent event that peaked around 10/11 February. The authors should discuss these and provide potential explanations for the forecast discrepancies. I think adding the rainfall forcing would shed some light about this.*

Author response. We thank the referee for this comment and we agree that the evidence presented is not adequate to support the word “realistic”. Instead, we will use a quantitative evaluation of the forecasts in Fig. 7d–f, and discuss the differences identified by the referee.

Peak magnitude and timing. For each forecast (29 January, 31 January and 2 February), we will report the error in peak discharge and peak timing at Sellheim, relative to both the observations and the river analysis. The early arrival of the simulated peak was also noted by RC2, and the referee also notes it in the Luangwa case (Sect. 4.3). This suggests a systematic component rather than a forecast-specific error. We can make several working hypotheses:

- the absence of a groundwater reservoir, so that subsurface runoff reaches the channel without delay
- errors in the timing or spatial distribution of the forecast precipitation.
- river channel parameters: in the CaMa-Flood implementation, channel depth is derived from a power-law function of annual mean flow, and riverbed roughness uses a default value that has not been calibrated (L441–443). Both can affect the travel time of the flood wave.
- the runoff generation in ecLand, which controls how quickly runoff reaches the channel.

River model spin-up is an unlikely cause for this event, because the river analysis that initialises the forecasts is a continuous simulation that has run over a longer period.

Underestimation of the 10–11 February event. The second peak, about 5–10 days after the first major peak, is underestimated in the forecasts and, to a lesser extent, in the analysis. Adding the precipitation forcing to the plot will show whether this comes from the precipitation (too little rain, or rain in the wrong place or at the wrong time) or from the hydrological response.

The river analysis reproduces the falling limb of the major peak well, which suggests that the land and river states before the second event are reasonable. The discrepancy is therefore likely due to the rainfall forcing. Because the river analysis is driven by IFS short-range forecasts rather than observed precipitation (Sect. 3.6), an underestimation of rainfall in these forecasts would affect the analysis and the forecasts alike. A contribution from the land or river components cannot be excluded, for example, through the representation of antecedent soil moisture or floodplain storage after the first peak.

Rainfall forcing. We agree that adding the rainfall forcing will help interpret these errors. We will add catchment-averaged forecast and analysis (shortest-range forecast) precipitation to Fig. 7d–f or as a new panel.

Planned changes to the manuscript:

- Replace “realistic” (L585) with the peak magnitude and timing errors at Sellheim for each forecast in Fig. 7d–f.
- Add catchment-averaged forecast and analysis precipitation, together with an observation-based estimate, to Fig. 7.
- Add a paragraph discussing the early peak and the underestimation of the 10–11 February event, with the hypotheses above and a cross-reference to the discussion of the other case studies (see RC2 M8).

3 – Definition of flood categories (Fig. 7)

Referee comment. *Further clarification of how the flood categories (minor, moderate, major etc) are defined in Figure 7 is required. For example, in Australia where the case study is conducted, flood categories are typically defined based on impacts rather than exceedance probabilities. Therefore, a discussion of the rationale behind the approach and how this translates to local emergency management authorities would be beneficial.*

Author response. The categories in Figure 7 are defined by return-period thresholds from a Gumbel distribution fitted to annual maximum discharge from the IFS river reference simulation for 2005–2024 (L567–569; Fig. 7 caption). Using the model's own climatology gives globally consistent thresholds and reduces the effect of systematic model bias. This holds to the extent that the reference simulation (ERA5-forced) and the forecasts (coupled IFS) share the same land and river biases. They are hazard indicators, not impact-based warnings like those from the Bureau of Meteorology. This will be made clear, and no labels that might be confused with the official categories will be used.

Planned changes to the manuscript:

- Move the category definitions from the caption into the text and explain the rationale.
- Rename the classes (e.g. "return-period classes") to distinguish them from BoM minor/moderate/major flood levels.
- Where available, relate model return periods to the BoM flood classification levels at Sellheim
- Note that thresholds above ~20 years are extrapolated from a 20-year record and are therefore uncertain.
- Note that the thresholds come from the ERA5-forced reference simulation, while the forecasts use the coupled IFS, so their biases may differ.

On behalf of all co-authors, Cinzia Mazzetti

Use of AI tools: in preparing these responses, the authors used Claude (Anthropic) to check that all referee comments had been addressed, to identify inconsistencies between the responses and the preprint, and to suggest edits to the wording. All responses and proposed changes were reviewed, revised where needed and approved by the authors, who take full responsibility for their content.