Response to editor for the revised manuscript:

"CaMa-Flood-GPU: A GPU-based hydrodynamic model implementation for scalable global simulations"

Manuscript ID: egusphere-2025-6500

Dear Editor,

We thank you for the careful reading of our revised manuscript and for the positive assessment that the reviewers' main concerns have been addressed. We also appreciate the constructive editorial suggestions provided before final acceptance. In response, we have carefully revised the manuscript and the accompanying response document following your suggestions.

We believe that the revised version has been further improved and that the remaining editorial concerns have been addressed. We hope that the manuscript is now suitable for publication in *Geoscientific Model Development*.

On behalf of all authors,

Sincerely,

Jiabo Yin and Dai Yamazaki

Reply to editor's comments

"CaMa-Flood-GPU: A GPU-based hydrodynamic model implementation for scalable global simulations"

Manuscript ID: egusphere-2025-6500

Legend

- Editor's comments
- Authors' responses
- Unchanged context from the manuscript
- ~~Deleted text in the revised manuscript~~
- New or changed text in the revised manuscript

Dear Authors, Thank you for submitting revised manuscript and nicely detailed response to reviewers comment. I overall find that the revision has addressed the reviewers' main concerns. The manuscript now contains a clearer description of the multi-GPU domain decomposition and communication strategy, a more transparent experimental setup/design, expanded numerical-stability comparisons, improved figures and captions, and broader context for both CaMa-Flood and GPU-based modeling. I therefore do not think an additional round of external review by reviewers again is necessary. Before final acceptance I request you consider a small number of minor revisions to improve clarity for readers.

To editor: We thank the editor for the careful reading of the revised manuscript and for the positive assessment that the reviewers' main concerns have been addressed. We appreciate the constructive editorial suggestions, which helped us further improve the clarity and internal consistency of the manuscript. We have revised the manuscript and response document accordingly, and the corresponding changes are shown below.

Comment 1. Clarify the terminology around Fig. 6. I think improving this could help reduce confusion. Please make explicit the distinction among unit-catchments, primitive basins, retained basins/basin groups, and bifurcation-formed basin groups. Please also briefly explain why some retained regions in the figure appear spatially large.

To editor: Thank you for pointing out the remaining ambiguity in the Figure 6 terminology. We revised the caption so that the map objects are named more explicitly. The caption now distinguishes unit-catchments, primitive basins, retained basin groups, and bifurcation-formed basin groups, and it explains why some retained regions appear spatially large. We hope this makes the figure easier to interpret without adding unnecessary text to the main body.

Revised manuscript text:

Example of a customized simulation domain constructed ~~from bifurcation-formed basins~~ after ~~the~~ point-of-interest filtering and ~~domain-decomposition step. Each retained basin contains at least~~

~~one point of interest. Coloured patches denote the retained basins, while~~ basin-group domain decomposition. The unit-catchment is the computational element; primitive basins are collections of unit-catchments draining to the same river mouth before cross-basin bifurcations are considered. Coloured patches denote retained basin groups, each containing the upstream unit-catchments needed for at least one point of interest; grey areas mark ~~basins excluded from the selected domain;~~ excluded basin groups. Some retained patches are spatially large because a downstream point of interest can require preserving an entire upstream primitive basin, and cross-basin bifurcation links can merge several primitive basins into one basin group. Markers indicate river-mouth locations and blue lines denote cross-basin bifurcation links.

Comment 2. Clarify the distinction between the performance benchmark and the high-resolution forcing/numerical-stability experiment. The revised manuscript explains that wall-clock benchmark use the 1 degree binary sample runoff, while the numerical-stability comparison includes 0.1° ERA5-Land NetCDF forcing. Please state clearly whether the 0.1° ERA5-Land experiment was used to quantify wall-clock performance overhead from high-resolution forcing input.

To editor: Thank you for noting that the distinction could be clearer. We revised the experimental-setup section to state that neither of the numerical-stability forcing experiments, E2O or ERA5-Land, is used in the wall-clock performance benchmark. They are used to test numerical and input consistency under independent runoff products, while the speed benchmark remains based on the 1° binary sample runoff distributed with the CPU CaMa-Flood release.

Revised manuscript text:

For each numerical-stability comparison, CPU and GPU runs use the same 6-arcmin parameter set, identical forcing, initial states and adaptive sub-step settings, with the bifurcation module enabled in both runs. The E2O and ERA5-Land numerical-stability experiments are not used to quantify wall-clock performance overhead from forcing input; they are used to test numerical and input consistency under independent runoff products.

Comment 3. Reconcile the GPU memory-footprint description. Please check the manuscript, tables, captions, and response documents to ensure that the reported GPU memory requirements are internally consistent, including the stated memory footprint for the 1-arcmin global state. I have confusion about whether 20 or 30 GB memory footprint in total, based on your response document.

To editor: Thank you for catching this possible source of confusion. We rechecked the manuscript and response materials and have standardized the formal value to about 30 GB in total for the 1-arcmin run. To prevent ambiguity, we revised the manuscript wording from "global state" to "aggregate GPU memory footprint". We also clarify that this total includes hydrodynamic state arrays, auxiliary routing and bifurcation buffers, runoff-mapping buffers, and temporary work arrays. The revised text now states that this footprint is distributed by the LPT assignment of bifurcation-formed basin groups, and that cases exceeding the available single-card GPU memory are marked OOM.

Revised manuscript text:

At the 1-arcmin resolution, ~~the global state requires about 30 GB of GPU memory in total.~~ the aggregate GPU memory footprint is about 30 GB in total, including hydrodynamic state arrays, auxiliary routing and bifurcation buffers, runoff-mapping buffers, and temporary work arrays. ~~This footprint is partitioned across ranks by the LPT assignment of bifurcation-formed basin groups, so the per-GPU peak roughly halves with each doubling of the rank count.~~ This footprint is distributed across GPU ranks by the LPT assignment of bifurcation-formed basin groups. ~~Configurations whose per-card GPU memory in Table 1 is below the resulting per-GPU peak are marked OOM in Table 3.~~ Cases exceeding the available single-card GPU memory are marked OOM in Table 3.

Comment 4. Clarify how bifurcation behavior is evaluated in the numerical-stability comparison. I have confusion on this point. The revised MS explain that the bifurcation-enabled configuration exercises both the base routing operations and the bifurcation-specific accumulation pathway. Please make this point clear where the numerical-stability results are discussed, or otherwise clarify the extent to which behavior at bifurcation-routing pathways or points is directly evaluated.

To editor: We agree that this point should be stated where the numerical-stability results are discussed. We therefore expanded the mass-conservation and routing-flux discussion in the numerical-stability subsection. The revised text now clarifies that bifurcation routing is enabled together with the main-channel routing in these runs. Thus, the comparison indirectly verifies CPU–GPU consistency for the complete bifurcation-enabled model configuration, while it is not a separate pointwise validation of individual bifurcation split ratios or bifurcation hydraulics.

Revised manuscript text:

Each outgoing flux is therefore added to the corresponding downstream storage term, with the CPU–GPU difference limited to the floating-point summation order discussed above. ~~The bifurcation-enabled configuration exercises both the base routing operations and the bifurcation-specific atomic-add path.~~ Because bifurcation routing is enabled together with the main-channel routing in the numerical-stability runs, the GPU executes the coupled routing workflow that includes both the base main-channel operations and the bifurcation-specific accumulation pathway. The comparison therefore provides an indirect CPU–GPU consistency check of the complete bifurcation-enabled model configuration, rather than a separate evaluation of individual bifurcation split ratios or point-by-point bifurcation hydraulics.