EGUSPHERE-2025-4387

Title: Validation of the Open-Source Hydrodynamic Model SFINCS on Historical River Floods at the Global Scale

Author(s): Tarun Sadana et al.

MS No.: egusphere-2025-4387

MS type: Research article

Dear Dr. Wouter Buytaert,

We would like to express our gratitude for the constructive and helpful comments from you and the two reviewers. They have contributed to substantial improvements in our manuscript.

Please find below the review comments (in black) and replies (in purple). The section, figure and table numbers refer to the original pre-print manuscript version.

With kind regards, and on behalf of all co-authors,

Tarun Sadana

Black (Comment)

Purple (Reply)

Referee 1: Francesco Dottori

Dear Dr. Francesco Dottori,

We would like to thank you for your time and effort to review the manuscript. It is very nice to hear that you think the paper is interesting for researchers working with large-scale flood models.

Please find below the review comments (in black) and replies (in purple). The section, figure and table numbers refer to the original pre-print manuscript version.

With kind regards, and on behalf of all co-authors,

Tarun Sadana

Reply to Francesco Dottori

This paper describes a modelling framework for simulating riverine flooding, focusing in particular on its validation. The authors investigate the sensitivity of the model performance to different geographic conditions, input data and modelling parameters, leveraging a large validation set of flood events at global scale. This allows for drawing general conclusions that in my opinion are interesting for researchers and practitioners working with large-scale flood models.

Overall the paper is clear and well structured, with a detailed analysis of the results and in-depth discusssions of several aspects influencing the skill of the model. Also, I would like to commend the open-source structure of the modelling framework, which gives more relevance to the work.

I have some moderate/minor comments and suggestions that should be addressed to improve the quality of the manuscript before publication .

Main comments:

Thank you for these helpful comments. Please find our reply to the main comments below.

Have you investigated the performance of the modelling framework in different climate zones? Hydrological models, including GloFAS, are usually less skillful in arid and cold regions, and this can negatively affect results. Given the large number of case studies, I think this factor could be included in the analysis.

This is indeed an interesting suggestion. Based on your comment, we have performed an analysis on how the model performs in different climate zones. We propose to include this as a new part (b) in Section 3.1.1 (“Model Performance in Different Climate Zones”), which reads as follows:

"**Section 3.1.1 (b) Model Performance in Different Climate Zones:**

To investigate the influence of climatic variations on model performance, each basin simulation was classified according to the Köppen-Geiger climate classification (Beck et al., 2018) using the centroid of each basin. Individual climate subclasses were grouped into the main climate classes: A (Tropical), B (Arid), C (Temperate), D (Cold/Continental), and E (Polar). This analysis was conducted for individual basins instead of flood events because flood events can occur over very large spatial scales and across multiple climate zones. Thereby, individual basins are weighted equally, meaning that smaller basins contribute to the overall average CSI in the same way as larger basins.

The results show clear variation in model performance across climate zones (Table 4). Performance is generally higher in tropical and temperate climates, where CSI values are 0.43 and 0.38, respectively. Lower performance is observed in arid climates with an average CSI of 0.32, while cold/continental climates show further reduced performance with an average CSI of 0.30. The lowest performance is observed in polar climates with an average CSI of 0.19. It is worth mentioning that most of the events simulated are based in Tropical (n = 193) and Temperate (n =139) climate zones.

Table 4: Model performance grouped per main climate zone. The number of flood events per zone is shown in brackets.

Climate Zone	Description	Average CSI
A (n=196)	Tropical	0.43
B (n=103)	Arid	0.32
C (n=139)	Temperate	0.38
D (n=54)	Cold/Continental	0.30
E (n=7)	Polar	0.19

This pattern is consistent with known limitations of both global hydrological models and satellite-derived flood observations in cold and polar regions, where snow and ice dynamics, frozen soils, and limited river gauges reduce the reliability of both the forcing data (GloFAS discharge; Harrigan et al., 2020) and the flood observations (Tellman et al., 2021). Performance of large-scale hydrological models in arid regions is usually lower compared to temperate and tropical regions (Salinas et al., 2013), as flooding in arid regions is driven by short intense rainfalls that are difficult to capture by hydrological models (Smith et al., 2015; Harrigan et al., 2020). In addition, runoff generation is highly variable due to infiltration and transmission losses in arid and semi-arid river channels, making peak flows harder to estimate (Dottori et al., 2016). Overall, these results suggest that the modelling framework performs most reliably in tropical and temperate climates, and that performance in arid and cold/polar regions is much lower."

Another point is the influence of flood protection structures. To my knowledge, global DEMs are currently not able to capture the presence of embankments along the rivers, and this limitation can heavily affect flood estimates, as observed by Dottori et al (2022) in several European river basins. The authors quicky discuss this issue at the end of Section 3.3 but I think it deserves some more detail. For instance, is there a difference in performance between more protected regions (e.g. rivers on Europe and US) and less protected ones (e.g. in Africa, South America)?

We agree that lacking infrastructure representation (e.g., levees, embankments) is a limitation of global DEMs. Following this suggestion, we have performed an exploratory analysis using the FLOPROS dataset (Scussolini et al., 2016) and the modelled floods. FLOPROS provides protection levels in terms of return periods which can be assigned to a basin cluster. We assigned the protection level to each basin cluster based on its centroid. We analyzed model performance across four key regions (Europe, North America, Africa, and South America), representing relatively highly protected and less protected regions. However, this analysis is exploratory and subject to several limitations, including the coarse spatial resolution of FLOPROS, uneven data availability (some regions have missing data), and confounding factors such as basin size and hydrological forcing. Therefore, to maintain focus and clarity in the main manuscript, we propose to include this analysis in the Appendix rather than as a separate main result section.

We present these results in a new paragraph in Appendices A3. The newly proposed section reads as follows:

"Section A3: Effect of Protection Standards on Model Performance:

A key limitation of global flood modelling frameworks is the inability of global DEMs to capture flood protection structures such as embankments and levees along river channels (Dottori et al., 2022). In this study, the modelled framework makes use of FABDEM (Hawker et al., 2022) and the floodplains are simulated using the subgrid approach. While subgrid methods have been shown to improve the representation of floodplains (Reshma et al., 2024), they still rely on the underlying DEM and therefore do not specifically take into account flood protection structures. To explore the effect of the protection standards on the model performance we make use of the FLOPROS dataset (Scussolini et al., 2016), which provides regional protection levels in terms of return period. Using this dataset, we assign flood protection standards (in return period years) at sub-national administrative level. As flood events can occur over a large area, we look at each basin cluster separately and assign values from the FLOPROS database. Some events occurring predominantly in low-income regions or smaller countries could not be assigned a protection level and were excluded from this analysis. The analysis was conducted in 4 key regions (North America, Europe, Africa, and South America). The results show consistent differences between regions, and the regions with higher protection standards (Europe and North America) show relatively higher bias values (1.12 and 1.26, respectively), indicating a tendency of the model to overestimate flood extent. In contrast, less well protected regions (South America and Africa) show lower bias values (0.96 and 0.89, respectively), suggesting more balanced or slightly underpredicted flood extents.

These differences can be explained by both the absence of flood protection structures in the DEM and regional variations in hydrological forcing. In highly engineered regions such as Europe and North America, levees and embankments are not represented in global DEMs (Dottori et al., 2022), leading to overestimation of flood extent. In contrast, in less protected regions, floodplain processes are closer to natural conditions, which can reduce this source of overprediction.

In addition, differences in hydrological model performance may also contribute to the observed patterns. Alfieri et al. (2013) showed that GloFAS discharge simulations perform well in river systems such as the Amazon and Mississippi but exhibit limitations in semi-arid regions (e.g. the Niger), where processes such as evaporation and infiltration are not fully captured. This can affect the accuracy of simulated floods and may contribute to regional differences in model performance observed in this study.

Table 5: Model Performance grouped per region.

Region	**FLOPROS protection range (return period years)**	**CSI**	**Bias**
Europe	2-10000	0.37	1.12
North America	2-500	0.41	1.26
South America	6.5-357	0.38	0.96
Africa	2-142	0.36	0.89

Minor comments:

Please add the scale on maps where appropriate (e.g. Figure 6 and 9)

Thank you for pointing this out. We revised Figures 6 and 9 to include scale bars in the revised manuscript to improve the readability of the maps.

Lines 47 "While local flood inundation studies have demonstrated promising results..." I would rather say that the use of local-scale flood models is a consolidated approach to map flood hazard and assess flood risk.

We agree with the reviewer and have revised the sentence to better reflect the role of local-scale flood models. The sentence is revised as following:

> "While the use of local-scale hydrodynamic flood models is a well-established approach for mapping flood hazard and assessing flood risk, large-scale or global inundation modelling still faces several challenges, such as limitations in topography and bathymetry data, computational intensity, and the open accessibility of datasets (Wing et al., 2020; Wing et al., 2021; Dottori et al., 2022)."

L 83-84: "Studies conducting event-based validation have not been global in scale, or they have focused on a limited number of events". I would include in the list the work by Risling et al (2024), as one of the few examples of validation studies involving multiple global flood models. The outcomes of the study (e.g. Figure 2) might be also relevant when discussing the performance of the SFINCS modelling framework.

Thank you for this suggestion. The study by Risling et al (2024) has been added to the in-text references. The revised sentence is as following:

> "Studies conducting event-based validation have not been deployed on a global level or have focused on a limited number of events (e.g. Risling et al., 2024, validating eight flood events across large river basins on four continents; Wing et al., 2021, using 35 events in the United States; Bernhofen et al., 2018, using five events in Nigeria and Mozambique)."

Section 3.2.3: My understanding is that all simulations described in this section are based on GloFAS discharge, is it correct?

The simulations performed in this section are indeed based on GloFAS discharge. All boundary conditions including discharge forcing were kept unchanged. Only the DEM input has been changed and the results are evaluated based on this change. We have clarified this in the text as follows:

"Section 3.2.3: The influence of the DEM input

> This section explores how different DEMs influence the performance of the SFINCS model in simulating flood extents. All simulations in this section are forced using GloFAS discharge, while only the DEM input is varied."

L 777-778: If I remember well, DEM for this study was also based on higher-resolution DTMs, can authors provide here more references?

We have clarified this by stating that in the study by Wing et al. (2017), the floodplains are represented by a DEM derived from 1 arc sec (~30m) USGS National Elevation Dataset (NED). We have revised the text as follows:

"In contrast, studies that make use of observed discharge consistently achieved much higher performance. Wing et al. (2017) used the USGS National Elevation Dataset (1 arc sec or ~30m) and evaluated LISFLOOD against high water marks in 35 discrete U.S. events using gauge records and reported a CSI of 0.87 "

Section 4: are you planning to use the SFINCS to produce global-scale flood maps, similarly to what was done by JRC (see for instance the latest update by Baugh et al., 2024)?

The possibility of producing large-scale flood maps using SFINCS, similar to the work conducted by the JRC (Baugh et al., 2024), is indeed an important aspect. We are now exploring the opportunities for upscaling the GEB modelling framework. This is quite a complex task since it also involves upscaling other models such as the agent based model.

Can you include a link to the Global Flood Database and specify whether flood extent data are freely accessible?

The flood extent maps from the Global Flood Database (Tellman et al., 2021) are freely accessible. The dataset can be accessed through the project website (https://global-flood-database.cloudtostreet.ai) as well as through Google Earth Engine (https://developers.google.com/earth-engine/datasets/catalog/GLOBAL_FLOOD_DB_MODIS_EVENTS_V1).
We will include the links mentioned here in the Data Availability Statement. Thank you for pointing out this was missing in the previous manuscript version.

L1240:"Hence, the results were generated in a timely manner". This is rather vague. Can you provide a broad estimate of running times based on the temporal and spatial extent of the events?

We agree with the reviewer that the statement is vague. The simulations were analyzed and a broad estimate was calculated based on the spatial and temporal extent of all the flood events simulated in this study. The following sentence is added to the section "A2.2 Snakemake".

"For example, events were split and executed in parallel across multiple compute nodes equipped with AMD EPYC 9354P processors (64 CPU cores, 96 GB RAM per node). Based on Snakemake log files, the average wall-clock time for a single basin simulation was approximately 7 minutes for a ~1,000 km² domain and a typical simulated flood period of ~30 days.

Each flood event consists of multiple basin cluster simulations (on average ~30 per event), resulting in a large number of individual simulations across the full dataset. These simulations were executed in parallel; however, the overall workflow also includes sequential steps per event (e.g. model setup, forcing update, simulation, and post-processing), as well as I/O and scheduling overhead. As a result, the total elapsed wall-clock time for completing all simulations was approximately 44 days."

References

Risling, A., Lindersson, S. & Brandimarte, L. A comparison of global flood models using Sentinel-1 and a change detection approach. Nat Hazards 120, 11133–11152 (2024). https://doi.org/10.1007/s11069-024-06629-7

Baugh, C.; Colonese, J.; D'Angelo, C.; Dottori, F.; Neal, J.; Prudhomme, C.; Salamon, P. (2024): Global river flood hazard maps. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-floods-floodmapgl_rp50y-tif

References:

- Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, *5*(1), 180214. https://doi.org/10.1038/sdata.2018.214
- Tellman, B., Sullivan, J.A., Kuhn, C. et al. (2021) Satellite imaging reveals increased proportion of population exposed to floods. Nature 596, 80–86 (2021). https://doi.org/10.1038/s41586-021-03695-w
- Harrigan, S., Zsoter, E., Alfieri, L., Prudhomme, C., Salamon, P., Wetterhall, F., Barnard, C., Cloke, H., & Pappenberger, F. 990 (2020). GloFAS-ERA5 operational global river discharge reanalysis 1979–present. Earth System Science Data, 12(3), 2043–2060. https://doi.org/10.5194/essd-12-2043-2020
- Smith, A., Sampson, C., & Bates, P. (2014). Regional flood frequency analysis at the global scale. Water Resources Research, 51(1), 539–553. https://doi.org/10.1002/2014wr015814
- Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., & Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. Advances in Water Resources, 94, 87–102. https://doi.org/10.1016/j.advwatres.2016.05.002
- Scussolini, P., Aerts, J. C. J. H., Jongman, B., Bouwer, L. M., Winsemius, H. C., De Moel, H., & Ward, P. J. (2016). FLOPROS: an evolving global database of flood protection standards. *Natural Hazards and Earth System Sciences*, *16*(5), 1049–1061. https://doi.org/10.5194/nhess-16-1049-2016
- Reshma, R., Devi, N. N., & Kuiry, S. N. (2024). Real-time urban flood modeling: exploring the sub-grid approach for accurate simulation and hazard analysis. *Natural Hazards*, *120*(11), 9609–9647. https://doi.org/10.1007/s11069-024-06579-0

Black (Comment)

Purple (Reply)

Referee 2: Anonymous referee

Dear Dr., Anonymous Referee #2,

We would like to thank you for your time and effort to review the manuscript. It is very nice to hear that you enjoyed reading the article and that you refer to the article as a "landmark study".

Please find your comments below (in black) and our replies (in purple). The section, figures and table numbers refer to the original pre-print manuscript version.

With kind regards, and on behalf of all co-authors,

Tarun Sadana

Reply to Anonymous referee #2

General comments

I enjoyed reading the article by Sadana et al. which presents a landmark global validation of the open-source SFINCS hydrodynamic model using 499 historical riverine flood events. The study demonstrates that while SFINCS achieves a mean CSI of 0.39 using simulated discharges, performance is highly sensitive to input data quality and basin scale. The fully automated and reproducible, open-source workflow, which couples SFINCS with global hydrological models, represents a valuable contribution to the topic.

Overall the article is well written, using grammar and English wording adequate to the standards of a scientific journal. Models and data used are mostly open source, which is appreciated. Main concerns about this work are detailed below, followed by a list of more specific comments.

Thank you for these helpful comments.

We appreciate the reviewer's positive evaluation of the study and the recognition of the contribution of the open and reproducible modelling framework. We also acknowledge the concerns raised regarding the coherence and applicability of some of the input datasets and the framing of the resulting limitations.

The main criticism is related to the appropriateness and coherence of the data used, which lead to results and limitations that are discussed as if they were findings of the work, though many of them could have been well foreseen since the start. In fact some choices taken sound as if the authors did not fully analyzed the range of use of each data source. For example, discharges derived from GloFAS v.4 are, according to the published documentation, meaningful for river basins larger than 500 km^2 (see Grimaldi et al., 2022 and Baugh et al., 2024), due to relatively coarse resolution of the model and even more of the meteorological input data. Here such data is used to model floods in basins in over 60% of cases smaller than 500km^2 (according to Table 3), later on reporting a tendency of GloFAS discharges to underestimate peak flows with selected return periods. This suggests a non optimal knowledge and use of the data. Similarly, I found it sub-optimal the choice to model inundation at 30m resolution in relatively small basins and then aggregate the output to 250m to be evaluated versus MODIS data. The comment made by the authors on using higher resolution data from SAR based products could have been addressed already in this stage of the work, rather than proposing it as an improvement for future works.

We thank the reviewer for this very useful comment. The primary objective of this study is to evaluate the performance of the SFINCS hydrodynamic automated modelling framework under globally available datasets and automated workflows, reflecting the conditions under which large-scale flood modelling systems are often applied. As such, the analysis intentionally includes events across a wide range of basin sizes and geographic contexts, even where the input data may fall outside the optimal applicability range of certain datasets. We acknowledge that this objective was not sufficiently emphasized in the manuscript, which may have created the impression that some limitations were unexpected findings rather than known constraints associated with global datasets. We addressed these limitations and have prepared revisions in text in multiple sections:

- Firstly, Regarding the use of GloFAS v4.0 discharge, we acknowledge that the dataset is primarily intended for river basins larger than approximately 500 km², as noted in Grimaldi et al. (2022) and Baugh et al. (2024). In the revised manuscript we now clarify this limitation much earlier in the methods section and explicitly discuss that this part of the analysis aims to quantify how model performance behaves when global discharge datasets are applied across basins of varying sizes. At the same time, GloFAS v4.0 remains the only open-source global discharge data available that covers the floods in the long temporal coverage required for this study. With respect to the spatial resolution of the inundation simulations, SFINCS was run at 30 m resolution to preserve fine-scale floodplain connectivity and rivers representation in the model grid. The outputs were aggregated to 250 m resolution only to allow comparison with the MODIS-derived flood extents used in the Global Flood Database. We could also have downscaled the flood extents to 30 m from 250 m, but this would have created many “observed” flood pixels outside the actually flooded areas. After discussion of the options (down-scaling the observations or up-scaling the simulations), we opted for up-scaling. We acknowledge that this choice was not clearly explained in the original manuscript and have now clarified it in Section 2.4.2. The additional text is added in Section 2.4.2, which now reads as follows:

  " Note that we resampled the high-resolution SFINCS outputs of 30 m to 250 m using mode resampling to match the resolution of observed flood extents. This approach avoids downscaling the MODIS observations, which would require disaggregation of coarse resolution pixels and could create observed flooded pixels outside the actual flooded areas. By aggregating the model outputs to the resolution of the observations using mode resampling, a consistent comparison between simulated and observed flood extents is ensured."

Furthermore, we agree with the reviewer that SAR-based flood maps could potentially offer advantages over MODIS-derived maps, and we have considered this at the start of the study. SAR-based products, such as those derived from Sentinel-1 (e.g. the Copernicus Global Flood Monitoring product; Salamon et al., 2021; Wagner et al., 2026), provide higher spatial resolution and are less affected by cloud cover. However, these datasets are currently limited in temporal coverage (mainly available from ~2016 onwards) and do not yet provide a globally consistent catalogue of flood events comparable in size to the Global Flood Database (Tellman et al., 2021), which contains over 900 events (of which 499 are used in this study) spanning 2000-2018. Existing SAR-based validation studies typically focus on a much smaller number of events (e.g. tens of events; Risling et al., 2024), rather than hundreds at the global scale.

Therefore, the choice is between using a large number of globally consistent events with coarser spatial resolution (MODIS-based) or using higher-resolution SAR-based flood maps with more limited spatial and temporal coverage. For this study, the availability of a large and consistent event database was essential to enable robust large-sample validation.

In the revised manuscript, we now clarify the tradeoff more explicitly and already in the methodology where we choose the dataset. We also mention the potential benefits of integrating SAR-based inundation datasets for future studies in our recommendations. We expect that as SAR-based datasets continue to expand in temporal coverage, they will become increasingly suitable for large-sample validation studies.

Additions to the following sections are also proposed:

"Section 2.2.2 Hydrological Forcing

GloFAS (LISFLOOD v4.0):

The GloFAS dataset (Alfieri et al., 2013; Grimaldi et al., 2022) is generated using the global hydrological model LISFLOOD-OS, which is a distributed, physically based rainfall runoff model that uses the ERA5 reanalysis data as input to simulate global river discharge at ~5.5 km resolution at the equator. GloFAS discharge is primarily designed for river basins larger than ~500 km² (Grimaldi et al., 2022). In this study, it is applied across a wider range of basin sizes to assess how model performance varies with basin scale when using globally available discharge datasets, including smaller basins. This allows us to evaluate the behaviour of the modelling framework under conditions where the input data may be less reliable. Continued ----"

Lastly, the research questions to be tested in this work could be made clearer. Some side analyses such as the comparison of GloFAS versus USGS observations are a bit unexpected and distract the reader from the main application shown.

Thank you for this suggestion. We agree that the research questions could be stated more clearly. In the revised manuscript we have refined the framing of the research objectives and streamline parts of the analysis to better emphasize the main contribution of the study, namely the large-sample global validation of the SFINCS modelling framework.

In addition, we now explicitly frame parts of the analysis as sensitivity analyses. In particular, the comparison between GloFAS and USGS discharge data is clarified as a sensitivity analysis of the input discharge data, aimed at assessing how the choice of forcing data influences model performance, rather than as a separate or standalone analysis.

We clarify this issue in the manuscript as follows:

> "In addition to the global validation, we perform a sensitivity analyses to assess the influence of key input datasets on model performance. In particular, the comparison between GloFAS and USGS discharge data is used to evaluate the sensitivity of the model to the choice of discharge forcing."

Specific comments

L23: "is" or "was" replaced

Thank you for noticing this. We have revised the sentence to the following:

> "When the (globally simulated) GloFAS data is replaced with observed discharge data for ten events in the US, the CSI improved from 0.39 to 0.67."

L162-170: matching hydrological and hydraulic river networks is far from being a simple task and remains one of the main challenges for operational inundation forecasting at large scales. A matching based only on the upstream area is generally not very accurate, particularly 1) when the resolution of the hydraulic and hydrological model DEMs are substantially different (here ~30m vs. ~5km) and 2) when the 2 underlying DEMs have different sources (here FABDEM and MERIT). I encourage to provide justifications for such choice and validation for selected cases (or summary statistics for a larger sample).

We agree that matching hydrological grids with the river networks used in hydrodynamic models is a challenging task. We also acknowledge the differences in spatial resolution, as hydrological models typically operate at a coarser grid resolution than the river networks used in hydrodynamic simulations. In our implementation, we did not observe issues arising from this matching approach.

In the revised version of our paper, we more clearly explain that linking rivers to the hydrological grid using upstream drainage area matching has been applied by following methods such as by Eilander et al. (2023). This work also applied this method to large-scale flood modelling, in addition, our sensitivity analysis (Section 3.2.1) also shows that using higher-resolution hydrological forcing (e.g., CWatM at 1 km × 1 km resolution) improves model performance.

We have clarified this in the manuscript as follows:

Section 2.1

Headwater and river inflow points:

> "Previous Text ----
>
> This linkage between hydrological discharge and the hydrodynamic river network is based on upstream drainage area matching, a method commonly used in large-scale flood modelling frameworks (e.g. Eilander et al., 2023). The use of a tolerance in upstream area ensures correct placement of forcing points, and the influence of hydrological input resolution on model performance is further evaluated in Section 2.5 Sensitivity Analysis.
>
> Continued ----"

L395: "the river roughness value (0.02) derived from widely accepted Manning's roughness values". River roughness depends on the land use and considering the application to so many different floods/ world regions, one value cannot accurately represent roughness in all context. Why did you not link roughness to the map of land use? If the approach relies on a constant roughness everywhere some additional justifications would be necessary, ideally with supporting references.

We thank the reviewer for this comment and would like to clarify how roughness is represented in the model. In our setup, we did link different land use types to their roughness from the land cover map (ESA-landuse Dataset). However, only for the river channels, a constant Manning's roughness value (n = 0.02) is used.

This approach is also commonly adopted in other large-scale flood modelling frameworks. For example, the JRC modelling framework applies channel Manning's n value of 0.02-0.04 (Dottori et al., 2016; Komi et al., 2017), CaMa-Flood commonly uses a value of 0.03 (Yamazaki et al., 2011), and LISFLOOD-FP subgrid applications uses values of 0.04-0.045 (Neal et al., 2012). These studies demonstrate that representative constant values are widely used in global and continental-scale applications.

In our case, the value of n = 0.02 follows the default parameters in the HydroMT-SFINCS workflow (Deltares 2024) and falls within a similar range of values used in comparable large-scale models. Given the global scope of this study, calibration of roughness values was not feasible.

We have revised the manuscript to clarify this modelling assumption and to distinguish between floodplain and channel roughness. The revised text reads as follows:

> "The rectangular cross-section used in SFINCS is a widely adopted representation in hydrodynamic modelling (e.g. Neal et al., 2012). Floodplain roughness values were spatially varied based on land use data (ESA Land Cover), while a constant Manning's roughness value (n = 0.02) was applied for river channels to ensure consistency across the large number of basins and regions considered. This constant value has been used by other studies (Bril et al., 2026) and

similar approaches using fixed channel roughness values are commonly adopted in large-scale flood modelling studies (e.g. Yamazaki et al., 2011; Dottori et al., 2016; Neal et al., 2012)."

Sect 3.1.1 and Table 3: (following the main comments) here the authors focused on classes of upstream area at and beyond the limits of applicability of GloFAS streamflows. GloFAS is a global system made for large riverine floods and cannot capture the flood dynamics in river basins with small upstream area. Its historical runs have daily resolution and are forced by ERA5 (~30x30km resolution), hence with constant precipitation rates over 1 day and almost 1000km^2, which in the analysis shown is the threshold of the largest basin class. At such scales, actual floods are typically induced by shorter rainfall extremes and at smaller spatial scales. The authors provided some sound justifications (L486-523) for increasing performance in larger basins, and the choice of the classes of upstream area was also driven by trying to have classes with similar sample size. However, GloFAS streamflow are recommended to be used for river basins larger than 500 km^2 (see Grimaldi et al., 2022 and Baugh et al., 2024). I recommend the authors to improve the performance analysis on different classes of upstream area, by focusing on larger limits, where the smaller class includes all basins smaller than 500km^2 , to show that performance deteriorates when using data outside the range of usability of the streamflow input.

Thank you for this very useful suggestion. We agree that the applicability of GloFAS streamflow is generally more appropriate for river basins larger than approximately 500 km², as noted in Grimaldi et al. (2022) and Baugh et al. (2024). The current analysis aimed to evaluate model performance across a wide range of basin sizes, including smaller basins where global hydrological forcing datasets may be less reliable.

In the revised manuscript, we revised the text:

"Section 2.2.2 Hydrological Forcing

GloFAS (LISFLOOD v4.0):

GloFAS discharge is primarily designed for river basins larger than ~500 km² (Grimaldi et al., 2022). In this study, it is applied across a wider range of basin sizes to assess how model performance varies with basin scale when using globally available discharge datasets, including smaller basins."

"Section 3.1.1 Effect of Upstream Area on Performance

Previous Text ----

Notably, the total number of basins varied significantly across the upstream area classes. We note that GloFAS discharge is generally recommended for basins larger than approximately 500 km² (Grimaldi et al., 2022; Baugh et al., 2024). In our dataset, approximately 61% of the analyzed basins fall below this threshold (n = 16,225), while 39% are ≥500 km² (n = 10,367). This allows us to explicitly assess how model performance changes across basin sizes, including those where the hydrological forcing is expected to be less reliable.

> Basins with a large upstream area (≥1,000 km²) were the most represented group (n = 8,834) as compared to basins with a medium-sized upstream area (500 to 1,000 km²; n = 1,533).
>
> Previous Text----"

Previous studies (just like in our study) have also analyzed model performance of large-scale flood models across similar basin size classes (e.g., Wing et al., 2017; Bernhofen et al., 2018). Including smaller upstream basins therefore allowed us to assess how model performance degrades when global discharge datasets are applied outside their typical applicability range. Therefore, in our opinion the results presented in Table 3 and Section 3.1.1 complement existing literature by providing additional evidence on the behavior of global flood modelling frameworks in smaller basins.

L599-608: It sounds like this issue may be addressed to a good extent by merging the MODIS images with the permanent water layer before comparing it with the SFINCS output (following regridding of both maps on a common grid)

Thank you for this comment. We would like to clarify how the post-processing was performed. The merging of the permanent water layer (band = 5) with the flood extents (band = 1) was already part of the original processing workflow for both the validation dataset and the SFINCS output.

The issue addressed in L599-608 is on the meandering or curving of rivers at 250m. At this resolution smaller rivers are not well represented in the flood maps. The issue is addressed again and mentioned in section 4: *"Additionally, the upscaling of SFINCS results from 30 m to 250 m resolution (to match our observed data) caused a loss of details that further impacted model performance."*

Sect 3.2.2: As commented above, the pattern shown in the graph is a result of using modelled discharges from points with upstream area at (and beyond) the lower edge of the range of usability of GloFAS data.

We show in the Section 3.2.2 that some assumptions that have been taken to automate the globally applicable framework have drawbacks. We agree with the reviewer that the problems in the graph (Figure 8) lower edge is being caused by using modelled data. Therefore, we now mention the following in section 3.2.2:

> "As a result, bathymetry derived from GloFAS tends to yield narrower and shallower channels, which reduces modelled conveyance capacity and potentially produces overestimated flood extents (i.e. higher FAs). These results imply that (i) using the 2-year default alone may not capture observed bankfull behaviour in the current fluvial setup and (ii) alternative approaches (e.g. gauge-based bankfull estimates, where available) can improve the bathymetry estimations (Zarrabi et al, 2025; Rad et al, 2024)."

We believe that whoever uses our open-source automated framework should be informed about these systematic biases and have a way to counter them by using open-source river gauge data (e.g., USGS) instead of modelled data (e.g., GloFAS).

L744: the difference in performance related to the use of different DEMs is impressive and would benefit from further investigation. It would be interesting to know if it's related mainly to the resolution or to the product. For instance one could resample the 3DEP DEM to progressively larger grids from 1 to 30m and test the differences in the SFINCS output and related performance.

We thank the reviewer for this suggestion. The section 3.2.3 “The influence of DEM input” mentions the following:

“The resolution of the SFINCS model was kept the same between the two different inputs, although we increased the level of detail (i.e. hypsometry levels, n = 20, default 10) stored in the subgrid tables when using the 3DEP DEM.”

The reviewers suggestion to resample 3DEP is already being done internally in the model. To go into detail, the reprojection of the DEM to the SFINCS model grid is handled internally by the ’Hydromt package’, meaning both the FABDEM (30 m) and 3DEP (1 m) inputs are reprojected to the same SFINCS model grid resolution.

We agree that this aspect was not sufficiently clear in the original manuscript and have now clarified it in Section 3.2.3. We have added the following changes:

“Both DEM inputs (FABDEM at 30 m and 3DEP at 1 m) are internally reprojected to the same SFINCS model grid resolution using hydromt. Therefore, the key difference lies not in the model grid resolution itself, but in the richness of the subgrid representation: higher-resolution DEMs provide more detailed elevation information within each grid cell, which is captured through the subgrid hypsometry tables.”

In addition, when using the 3DEP DEM, we increased the number of hypsometry levels stored in the subgrid tables (n = 20, default = 10), allowing SFINCS to better represent the elevation–volume relationship within each grid cell. This explains the improved performance observed when using higher-resolution DEM input.

L793: “covering approximately three to 35 events”. These numbers are quite specific. Add supporting references.

We have revised this sentence as following:

> *“Our study is unique in the high number (n = 499) of events covered, which is much higher than in existing studies covering approximately three to 35 events (Risling et al., 2024, Wing et al., 2021, Bernhofen et al., 2018).”*

Why the GEB model was not used more extensively, given the better performance? Is it available only for India? Please include some additional detail to clarify.

We agree that it might be unclear why the GEB model was only used for India despite its better performance. While GEB provides higher spatial resolution than GLOFAS, it is currently not available globally. As a result, it could only be applied to a limited subset of case studies rather than to the full set of events analyzed in this study. We have revised the text also as following:

"Section 2.5: Sensitivity Analysis

> *Discharge Data:*
>
> Previous Text ----

> Furthermore, to ensure robustness and avoid reliance on a single hydrological model, we also forced SFINCS with discharge from the GEB model (Section 2.2.2). **However, GEB is currently not available as a global dataset and is still under development for large-scale applications. As a result, its use in this study was limited to a subset of case studies rather than the full global event set.**
>
> Previous Text ----
>
> "

L903: For the Global Flood Monitoring I think that more appropriate citations are Salamon et al (2021) or Wagner et al (2026)

Thank you for the suggestion. We have revised the sentence as following:

> "The findings of this study echo research that calls for better open-source observed flood validation datasets, such as the Global Flood Monitoring System offered by Copernicus Emergency Management Service (Salamon et al., 2021; Wagner et al., 2026; Bates, 2023)."

Again, I think that adding higher resolution SAR based inundation maps would greatly benefit the analysis, thanks to resolutions comparable or higher than that used for the SFINCS model simulations, being less affected by cloud coverage and helping to fill the temporal gaps of days without acquisitions during floods. Indeed, resampling the model output from 30m to 250m sounds like a major limitation in the evaluation of the model performance, as also noted in the text (L579 onwards).

Thank you for this comment. We agree that SAR-based inundation maps provide important advantages for flood validation, including higher spatial resolution and reduced sensitivity to cloud coverage. However, in this study we relied on the MODIS-derived Global Flood Database because it provides a globally consistent dataset covering a large number of historical flood events (n = 499), which was essential for the large-sample validation performed here.

While the resampling of the SFINCS output from 30 m to 250 m may introduce some loss of spatial detail in the evaluation, this approach was necessary to ensure consistency with the spatial resolution of the validation dataset. Furthermore, validation based on MODIS-based flood maps can provide comparable validation metrics to validation based on SAR-based flood maps at large scales. For example, Risling et al. (2024) reported CSI values of approximately 0.45–0.65 when comparing global flood models with SAR-derived flood extents, which are within the range of values obtained in this study.

We agree that incorporating SAR-based inundation datasets could provide an additional perspective on model performance, and we have added this point in the manuscript when discussing the limitations of the evaluation dataset and possible directions for future work.

A summary figure of the distribution of the 499 events would be welcome.

Thank you for this suggestion. We agree with this remark, and have added a new figure to the supplementary section of the revised manuscript. Please see the figure below, which shows the 499 events modelled in this study, and their corresponding Köppen climate zones.

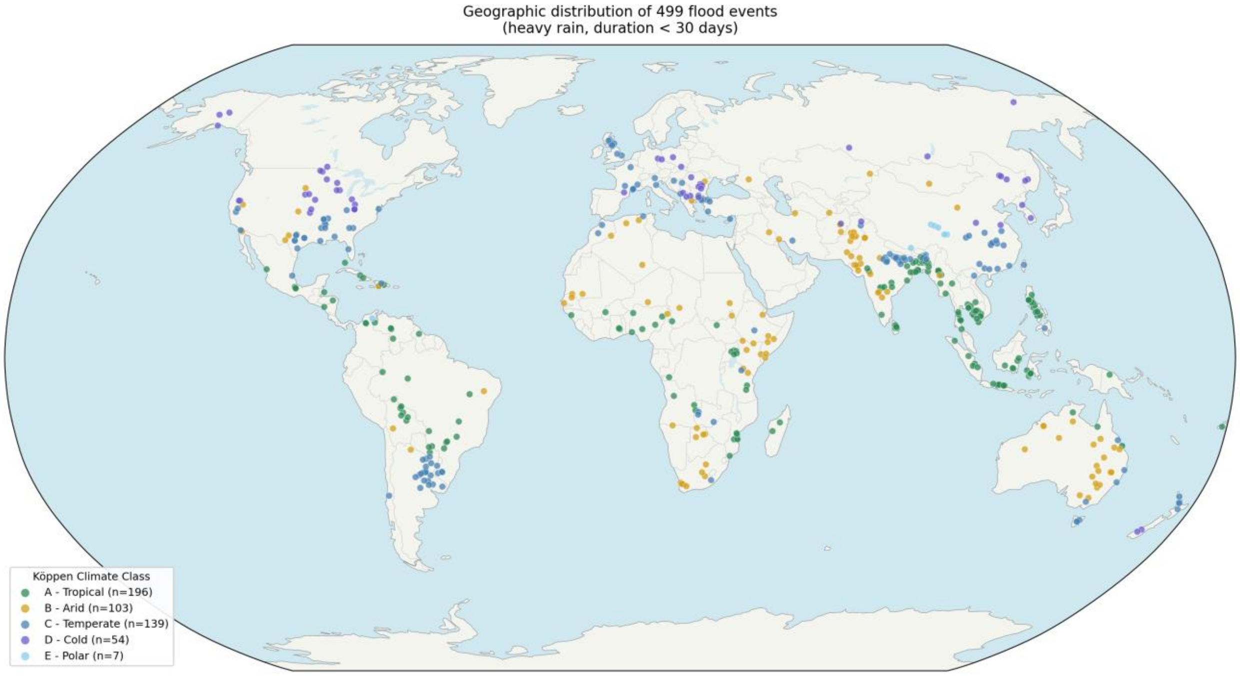

Figure A5. Modelled flood events filtered from the Global Flood Database with respect to their Koeppen Climate zones

References

Baugh, Calum; Colonese, Juan; D'Angelo, Claudia; Dottori, Francesco; Neal, Jeffrey; Prudhomme, Christel; Salamon, Peter (2024): Global river flood hazard maps. European Commission, Joint Research Centre (JRC) [Dataset]

Grimaldi, S., Salamon, P., Disperati, J., Zsoter, E., Russo, C., Ramos, A., Carton De Wiart, C., Barnard, C., Hansford, E., Gomes, G. and Prudhomme, C., GloFAS v4.0 hydrological reanalysis, European Commission, 2022, JRC131349.

Salamon, P., McCormick, N., Reimer, C., Clarke, T., Bauer-Marschallinger, B., Wagner, W., Martinis, S., Chow, C., Böhnke, C., Matgen, P., Chini, M., Hostache, R., Molini, L., Fiori, E., and Walli, A.: The New, Systematic Global Flood Monitoring Product of the Copernicus Emergency Management Service, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 1053–1056, https://doi.org/10.1109/IGARSS47720.2021.9554214, 2021.

Wagner, W., Bauer-Marschallinger, B., Roth, F., Raiger-Stachl, T., Reimer, C., McCormick, N., Matgen, P., Chini, M., Li, Y., Martinis, S., Wieland, M., Kraft, F., Festa, D., Hassaan, M., Tupas, M. E., Zhao, J., Seewald, M., Riffler, M., Molini, L., Kidd, R., Briese, C., and Salamon, P.: The fully-automatic Sentinel-1 Global Flood Monitoring service: Scientific challenges and future directions, Remote Sensing of Environment, 333, 115108, https://doi.org/10.1016/j.rse.2025.115108, 2026.

References:

- Eilander, D., Couasnon, A., Leijnse, T., Ikeuchi, H., Yamazaki, D., Muis, S., Dullaart, J., Haag, A., Winsemius, H. C., & Ward, P. J. (2023). A globally applicable framework for compound flood hazard modeling. *Natural Hazards and Earth System Sciences*, *23*(2), 823–846. https://doi.org/10.5194/nhess-23-823-2023
- Chow, V. T. (1959). *Open-Channel Hydraulics*. McGraw-Hill, New York. applied-hydrology-ven-te-chow.pdf
- Bril, V. C., De Bruijn, J., De Moel, H., Sadana, T., Busker, T., Botzen, W. J. W., & Aerts, J. C. J. H. (2026). Assessing the Effectiveness of Nature-Based Solutions and Building-Level flood risk reduction Measures: an Open-Source Coupled model. Water Resources Research, 62(3). https://doi.org/10.1029/2025wr041436
- Deltares. (2024). Deltares. Retrieved from hydromt_sfincs/hydromt_sfincs/data/lulc/esa_worldcover_mapping.csv at main · Deltares/hydromt_sfincs
- Bates, P. D., Horritt, M. S., & Fewtrell, T. J. (2010). A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. https://doi.org/10.1016/j.jhydrol.2010.03.027
- Neal, J., Schumann, G., & Bates, P. (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. https://doi.org/10.1029/2012WR012514
- Wing, O. E. J., et al. (2017). https://doi.org/10.1002/2017WR020917
- Risling, A., Lindersson, S., & Brandimarte, L. (2024). A comparison of global flood models using Sentinel-1 and a change detection approach. *Natural Hazards*, *120*(12), 11133–11152. https://doi.org/10.1007/s11069-024-06629-7
- Abstract EGU26-20604
- Bernhofen, M., Whyman Ch., Mark A Trigg, P Andrew Sleigh, Andrew M Smith, Christopher C Sampson, Dai Yamazaki, Philip J Ward, Roberto Rudari, Florian Pappenberger, Francesco Dottori, Peter Salamon, Hessel C Winsemius (2018) A first collective validation of global fluvial flood models for major floods in Nigeria and Mozambique, environmental Research Letters, doi.org/10.1088/1748-9326/aae014
- Risling, A., Lindersson, S., & Brandimarte, L. (2024). A comparison of global flood models using Sentinel-1 and a change detection approach. *Natural Hazards*, *120*(12), 11133–11152. https://doi.org/10.1007/s11069-024-06629-7
- Wing, O., Smith, A., Marston, M. L., Porter, J. R., Amodeo, M., Sampson, C., & Bates, P. (2021). Simulating historical flood events at the continental scale: observational validation of a large-scale hydrodynamic model. Natural Hazards and Earth System Sciences, 21(2), 559 –575. https://doi.org/10.5194/nhess-21-559-2021
- Wing, O. E. J., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., & Erickson, T. A. (2017). Validation of a 30 m resolution flood hazard model of the conterminous United States. Water Resources Research, 53(9), 7968–7986. https://doi.org/10.1002/2017wr020917
- Yamazaki, D., Kanae, S., Kim, H., & Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, *47*(4). https://doi.org/10.1029/2010wr009726

- Komi, K., Neal, J., Trigg, M. A., & Diekkrüger, B. (2017). Modelling of flood hazard extent in data sparse areas: a case study of the Oti River basin, West Africa. *Journal of Hydrology Regional Studies*, *10*, 122–132. https://doi.org/10.1016/j.ejrh.2017.03.001
- Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., & Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. *Advances in Water Resources*, *94*, 87–102. https://doi.org/10.1016/j.advwatres.2016.05.002
- Neal, J., Schumann, G., & Bates, P. (2012). A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. *Water Resources Research*, *48*(11). https://doi.org/10.1029/2012wr012514