We thank the Reviewer for the helpful comments and suggestions. We hereby address them individually. In this document we indicate the Reviewer's comments in *italic dark grey*, while text that was changed in the paper in blue.

This paper presents a deep-learning based approach for 2D surface flood modelling that builds on a previously existing model, SWE-GNN. It proposes an alternative model that overcomes the need of a numerical solver to determine initial conditions and includes other novelties to improve the speed and generalization of the model. The paper demonstrates that the model can benefit from fine-tuning to generalize to new case studies. In my opinion, the research is interesting and is very well presented. I am therefore recommending the paper for publication, after minor revisions.

General comments

The paper proposes a model for flood modelling and is applied only to a specific case of floods; dike-breach floods. It would be nice to have a section in the discussion on how the model could apply to other kind of floods, such as fluvial and pluvial floods for example.

We thank the Reviewer for the valuable point addressed here. We added a paragraph in the discussion section that indicates how to deal with other types of floods. This is written in lines 390-394 as:

"While the current model framework can work for dike-breach floods, we did not evaluate it for other types of floods. For river and costal floods, the model should work as well without any changes since the inputs are of the same type as dike breach floods, e.g., upstream discharge hydrograph or sea water levels. On the other hand, pluvial floods require precipitation as a further input. Assuming rainfall as a spatially distributed variable, it could be added as a dynamic forcing; this could work in a similar way as for static features, but changing at each time step, independently of the predicted output. For urban floods, the drainage system should also be included. This could be done, as in numerical methods, by coupling the overland flow, predicted by the mSWE-GNN, with a 1D model for the sewers, possibly with another learned GNN as in Garzon et al., 2024."

Specific comment

L84 – L85: For reproducibility, I would suggest clarifying how a mesh is classified as being too small and how the resolution of the fine mesh is chosen.

We thank the Reviewer for pointing out this concern. While addressing it, we also noticed a writing mistake, since "mesh edge" should be "flow edge", which are defined in the MeshKernel library as "the edges connecting faces circumcenters". Hence we corrected the corresponding sentence and also expanded on the definition of small mesh and mesh resolution.

Lines 83-86 now read as:

"For the same numerical constraints, after the orthogonalization, all elongated elements get removed, resulting in a mixture of triangular and quadrilateral elements. We define elongated elements as those whose line connecting barycentre and edge middle points is 0.1 times smaller than the other lines in the same element."

L93: It might be better to define what epsilon is here rather than in L104.

We thank the Reviewer for spotting this mistake. The epsilon in line 93 is not the same as the one in line 104. For this reason, we removed the first epsilon and rephrased the corresponding sentence as: "...if edge (i, j) exists."

Eq. 3 and 4: I think that it might be useful to have the clarification of what h_di and h_si are.

We added in line 114 a clarification of what these embedding mean:

"The encoded variables Hs, Hd, and E' represent a higher-dimensional version of the original inputs that is more expressive."

We also clarified what h_di and h_si are in lines 135-136 as:

"where ψ(·) : R^5G → R^G is an MLP, ⊙ is the Hadamard (element-wise) product, h^(ℓ)_di is the embedding of the dynamic inputs at node I and layer ℓ, h_si is the embedding of the static inputs at node i, and W^(ℓ) ∈ R^G×G are learnable parameter matrices."

L190: I would suggest replacing 'training simulations' by 'training data' here for clarity.

We replaced the term as suggested.

L193: You should explain that water level is the sum of the water depth and elevation of the cell as this is might not be straightforward.

We clarified how water levels are calculated in lines 193-194 as:

"… and w_i^t its water level, given by the sum of the elevation and water depth at time t."

Section 2.5: Clarify that O is the output dimension.

We clarified the meaning of the variable O in lines 206-207 as:

"… H is the prediction horizon, O the number of output hydraulic variables, and gamma_o are coefficients used to weigh the influence of each hydraulic variable to the loss."

Table 1: To which resolution of the mesh does the edge length refer?

We included an additional explanation line in the caption of Table 1 that reads as:

"All geometric variables refer to the properties of the finest mesh in each dataset."

Section 3.1: I believe a concise explanation of how the Manning's n coefficients are defined for the synthetic dataset would be helpful. I'm curious to know if these coefficients are spatially variable as the model might face challenges in extracting meaningful insights from them if they were spatially uniform.

We thank the Reviewer for the interesting point. In our experiments, we kept the same, spatially-uniform, roughness coefficients for all simulations, using a value of 0.023 $s/m^{1/3}$ everywhere, despite the model can potentially work with different values of it. We included this clarification in lines 219-221 as:

"For the Manning's roughness coefficient m, we used a spatially uniform value of 0.023 m^1/3 s^-1, which is kept the same throughout all simulations."

In the discussion section, we also added a further paragraph discussing possible implications of using a spatially varied distribution of roughness values (lines 402-403):

"We employed a constant and spatially uniform roughness coefficient, meaning that we did not assess how the model generalizes to different values and spatial distributions. This might lead to different dynamics that, following the same reasoning as for the different speeds of propagation, the model should still be able to capture."

L270-275: You provide a nice and clear explanation of CSI. However, you consider 2 thresholds (0.05 m and 0.3 m) while you explain in L273-275 that the flooded and non-flooded cells are distinguished from another in TP, FP and FN. However, this isn't *entirely accurate as cells with water depths below 0.3 m could still be considered as flooded. I would try to be more accurate and write for example: 'TP are true positives, i.e. number of cells where both model and simulations predict water levels above the threshold value'. The same applies for the description of FP and FN.*

We adapted the suggestion to lines 276-278 as:

"where TP are the true positives, i.e., the number of cells where both numerical and deep learning models predict water depth above a given threshold, FP are the false positives, i.e., the number of cells where the deep learning model wrongly predicts water depth above a given threshold, and FN are the false negatives, i.e., the number of cells where the deep learning model does not predict water depth above a given threshold."

L285: I would suggest adding a sentence stating that you are assuming the enhanced SWE-GNN also outperforms the other models, given that the original SWE-GNN outperforms them and despite the added modifications.

We thank the Reviewer for the comment. We added a sentence to clarify that the enhanced SWE-GNN model should also outperform other models, in lines 288-289:

"We also did not compare against other baselines as the SWE-GNN performs better than them (Bentivoglio et al., 2023), so we assumed the same holds for the enhanced version."

L305: Do you know why the MAE of h increases? Could it be due to error propagation throughout the simulation? If you have any insights into the cause of this increase, it might be worth adding it.

We thank the Reviewer for the valuable suggestion. We added the main insight on why errors tend to increase in time, in lines 308-309:

"The main reason for the increase in water depth MAE over time is that as the flood progresses, it covers a greater spatial extent, increasing the number of cells where prediction errors can occur."

Fig. 9 and Fig. 10: Consider using different color scales as it is difficult to visualize the differences in water levels as it is. Also, could you clarify which mesh (e.g. finest mesh) is shown in the figures?

We clarified in figures 9 and 10 that the plots refer to the finest mesh by adding the following sentence in the figure captions:

“All plots represent values only on the finest mesh.”

Regarding the color scales, we believe that the employed scales provide a good combination of colors to highlight the results of water depths, flood arrival times, and their differences. We recognize that some colors might not be visible because of the overlap with the mesh so we improved the figures by reducing the width of the computational mesh edges, which allows to better see the colors.

We report the change in Figures 9 and 10 below.

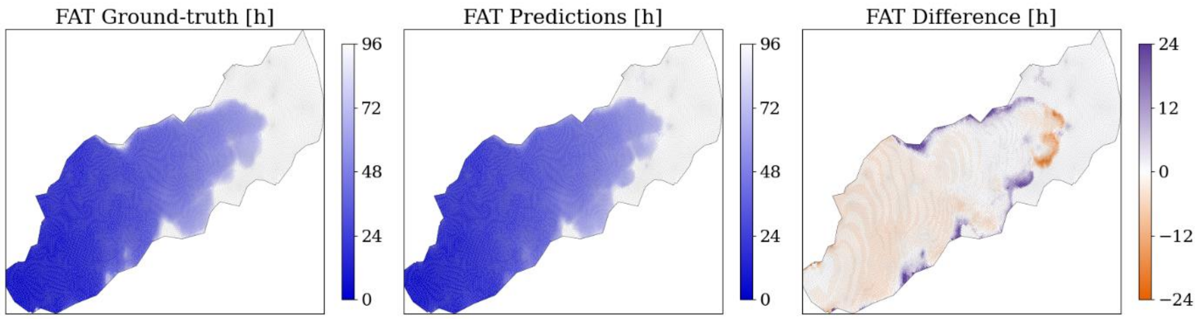

Figure 10

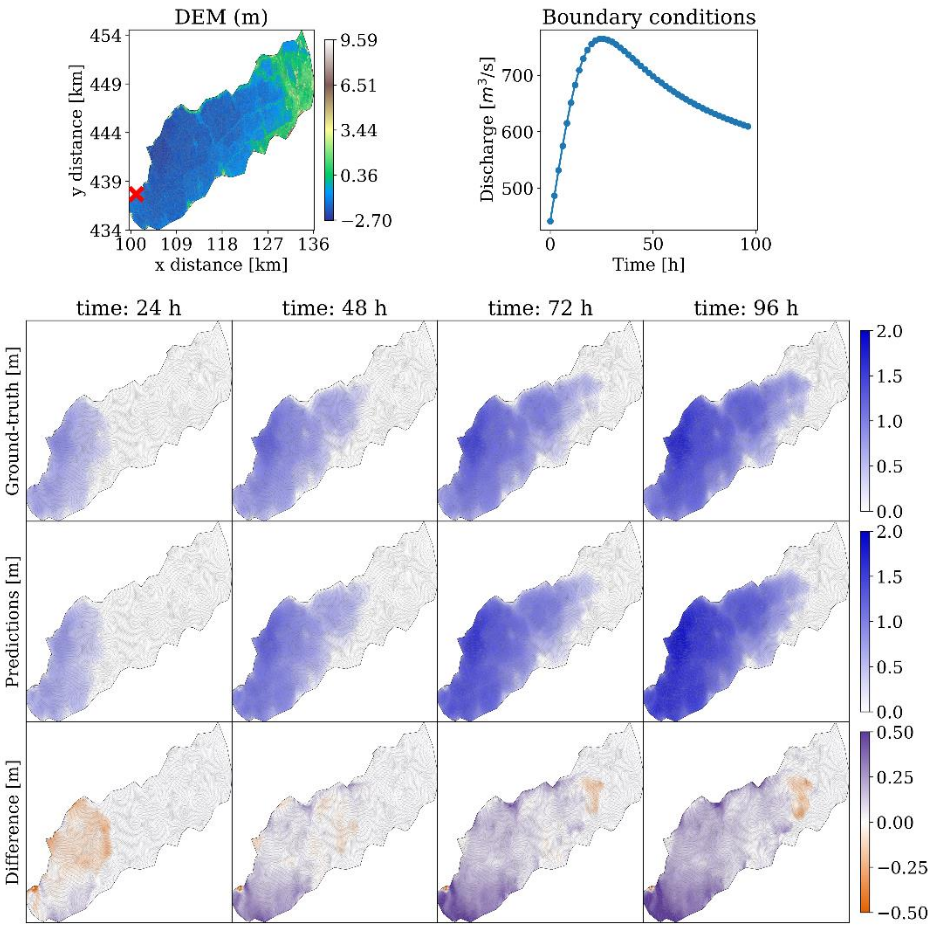

Figure 9

Fig. 10: It is really nice that you detail what the negative and positive values correspond to in the captions of Fig. 9 and Fig. 10. However, I would recommend revising the caption in Fig. 10 to enhance clarity, e.g. 'positive values indicate that the model estimates later arrival times than the numerical simulation, while negative values indicate that the model predicts earlier arrival times'

We thank the Reviewer for the suggestion. We changed the caption as recommended.

Technical corrections

We thank the Reviewer for spotting these issues. We agree with all of them and have modified them as suggested. In case of modifications that are not exactly as suggested, we clarified why and how after the comment.

Throughout the manuscript (e.g., L14, L51, L223, caption of Fig.5): Uncapitalize 'the' in 'The Netherlands'

L178: 'via edges directed towards' rather than 'via directed edges towards'

L292: Clarify 'is comparatively faster than the numerical model'

We clarified the sentence by removing the term "comparatively" which indeed was adding unnecessary confusion.

Throughout the manuscript: Add a spacing between the numbers and the 'm' for meters

L302-303 and 305: Replace the four 'in correspondence of' with 'occur at', 'occur simultaneously to', 'near' and 'at' respectively

We replaced all four 'in correspondence of' as suggested, except the second one which was changed to "close to".

Fig. 8: You might want to keep the notation of CSI0.05 and CSI0.3 like in the manuscript, i.e. add the 'm'

Table 2 and 3: Write the units as [10-2 m] and [10-2 m2/s]

L330: Typo in 'dataset'

L365: Typo in 'dependent'

L393: 'analyze' rather than 'analyse' to keep consistency with the previous sentence

L466: Remove 'the' in 'comes from the an increase'

References:

Garzón, A., Kapelan, Z., Langeveld, J. and Taormina, R., 2024. Transferable and data efficient metamodeling of storm water system nodal depths using auto-regressive graph neural networks. *Water Research*, p.122396.