We thank the Reviewer for his careful review and appreciation of our work. We value his insightful comments and suggestions, which we hereby address 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 ML-based approach for two-dimensional hydrodynamic modeling of floodings. In my eyes, the design takes inspiration from Geometric Deep Learning to wisely choose inductive biases so that the presented approach is aligned with knowledge from physical-principles (e.g., interactions follow along a gradient) and numerical modeling (e.g., the depth of the network compensates for time resolution). The result is a graph-based approach that is two magnitudes faster than a comparable numerical model and its performance is as good or better as other baselines — at least in modelland. The ideas are novel, the evaluation is good, and the exposition is clear. I only have small nitpicks and hope that the paper will be published as soon.

We are happy to see the Reviewer appreciates the design inspirations and we thank him for the kind words.

General Comments
1) Ablation Study. I have two problems with the ablation. First: I think that the major part of the Copernicus audience will (sadly) not be familiar with the term and I would thus propose to introduce what an "ablation study" is and what you do there as part of your experimental setup (e.g., after Section 4.3). Second: I will emphasize that I very, very much appreciate that the study did ablations. Ablations have become one of the primary tools in ML research, and I believe that many of the current Deep Learning applications are sourly missing them. Still, I want to attest that the presented ablations are rather unusual. An ablation usually refers to an experiment that removes part of the network. Here, on the other hand, the authors did ablate the loss and the algorithm. I am not sure if it is necessary to deal with this minor idiosyncrasy, but I personally would not call this an ablation at all, but rather an exploration of the importance of hyper-parameter choice (I do admit that both are very similar in intent). On the other hand, if the authors like their ablation-framing, I would like to suggest to also include an ablation that checks what happens if one reduces the model itself (e.g., removing the activations in equation 8 PRELU -> RELU -> no activation; or reducing the inductive bias by not using the h-h part in equation 7).

1) We thank the Reviewer for noticing this possible unfamiliarity with the term. According to both parts of the comment, we changed its name into "sensitivity analysis" and we added an explanation of its purpose in lines 378-380 as:

"Finally, we performed a sensitivity analysis on the role of the multi-step-ahead function (cfr. eq. (14)) and the curriculum learning (Algortihm 1) on the training performance. Sensitivity analysis is a technique that explores the effect of varying hyper-parameters to understand their influence on the model's output."

Regarding the suggested ablations, we already included the effect of removing the h-h part of equation 7 in section 5.1 (the model called SWE-GNN_ng). On the other hand, we believe that changing the nonlinearity would be somewhat tangential to the core message our paper, particularly after clarifying what we meant by ablation. That said, we will gladly consider the effect of different (or no) activation functions in our future work, and we thank again the reviewer for his insights.

2) Future work on speeding the model further up. I think a discussion on further speed ups of the modeling pipeline would be good. Is a two-magnitude speed up already so much that its not worth pursuing even faster models? Can we still expect speedups? How important is the tradeoff between the temporal resolution and the speed mentioned in L. 260? Etc. Here is, for example, a direction I spontaneously see (not saying that this should be included; just for providing inspiration): In theory one can directly output multiple (all) needed timesteps from the network. This would likely speed the process up considerably, since no recurrence over time (as shown, e.g., in Figure 3) would be required anymore. The price one pays then is that this "concurrent output" approach can only model the inundation for a finite time horizon (as a matter of fact, it will always model exactly to that horizon no matter what). One can also think of “in-between solutions' ', in which the model ingests and outputs chunks of time. This is not so easy with graphs, and does break the nice alignment with the physical conception of the problem.

2) We thank the Reviewer for the valuable feedback. Indeed, there are still many possibilities to speed-up the current model. While techniques like the one suggested here have already been applied (e.g., Brandstetter et al. 2022), we believe that it would not entirely solve the issue, as predicting more steps at the same time would still imply using more GNN layers, which are the true bottleneck of the model. Accordingly, assuming that the key factor is the number of GNN layers (for a given prediction horizon), one promising research direction would be to employ multi-scale methods that allow to reduce the number of message passing operations, while still maintaining the same interaction range. Since there is always a trade-off between speed and accuracy, we decided to further expand the suggestion by discussing speed-up in combination with finding a better Pareto front that would also results in better trade-offs.

As such, we included lines 428-431 as:

“Moreover, future works should aim at improving the model’s Pareto front. For improving the speed-up, one promising research direction would be to employ multi-scale methods that allow to reduce the number of message passing operations, while still maintaining the same interaction range (e.g., Fortunato et al., 2022; Lino et al., 2022). On the other hand, better enforcing physics and advances in GNNs with spatio-temporal models (e.g., Sabbaqi and Isufi, 2022) or generalizations to higher-order interactions (e.g., Yang et al., 2022) may further benefit the accuracy of the model.”

Brandstetter, J., Worrall, D. and Welling, M., 2022. Message passing neural PDE solvers. arXiv preprint arXiv:2202.03376.

Fortunato, M., Pfaff, T., Wirnsberger, P., Pritzel, A. and Battaglia, P., 2022. Multiscale meshgraphnets. arXiv preprint arXiv:2210.00612.

Lino, M., Fotiadis, S., Bharath, A.A. and Cantwell, C.D., 2022. Multi-scale rotation-equivariant graph neural networks for unsteady Eulerian fluid dynamics. Physics of Fluids, 34(8), p.087110.

Specific Comments

3) L. 16. I think I know what the authors want to say here, but in general I think that it is not clear what "bigger domains" and "longer periods" of time mean here. Bigger and longer than what?

3) Following the comment, we modified lines 16-17 as:
“Moreover, it generalizes well to unseen breach locations, bigger domains, and over longer periods of time, compared to those of the training set, outperforming other deep learning models”

4) L. 71. Maybe it would be good to say "two orders of magnitude" instead of "up to 600 times speed-ups" to align the contribution with the abstract.

4) We modified line 71 as suggested by the Reviewer as:
“We show that the proposed model can surrogate numerical solvers for spatio-temporal flood modelling in unseen topographies and unseen breach locations, with two orders of magnitude speed-ups”

5) L.121. I would suggest to remove the phrasing "... well-known 'curse of dimensionality', ...", because (1) it might not be as well known as you think, and perhaps more importantly (2) the term "curse of dimensionality" refers to a plethora of phenomena and thus readers might associate something different with it. Instead, I would propose to write something direct like "For MLPs the number of parameters increases exponentially with ..."

5) We thank the Reviewer for the relevant observation. As suggested, we modified lines 120-121 as:
“For MLPs, the number of parameters and the computational cost increase exponentially with the dimensions of the input.”

6) L. 126. Bronstein (2021) as a sole reference is probably a bit unfitting here, since LeCun and Bengio (1995) already discussed the importance of shared weights in CNNS (according to ideas lined out by LeCun in 1989). References:
- LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.
- LeCun, Y. (1989). Generalization and network design strategies. Connectionism in perspective, 19(143-155), 18.

6) We included LeCun et al. (2015) as a further reference in line 126. We opted for this reference instead of the suggested ones, as it is more recent.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), pp.436-444.

7) L. 131. The word "avoid" might be misleading here, as they just don't include these specific physical inductive biases. Maybe use "include" instead.

7) Indeed, we agree with the Reviewer to changing the term “avoid” into “do not include” as it fits better in this context.

8) L. 167f. "Either" here would suggest that both node-features need to be non-zero. I don't see that. Is this a formulation thing or am I missing something?

8) In this sentence, with “either of the interfacing node features is non-zero” we mean that at least one of the two needs to be non-zero. To avoid confusions, we changed the term “either” with “at least one”.

9) L. 174. More a thought than a suggestion: For me the W(l) is like an additional MLP layer (that maps to R1) without an activation. It might thus also be interesting to do an ablation on W(l).

9) We agree with the Reviewer that the weight matrix W could be replaced as well with a more complex function such as an MLP, as proposed as well in other works (e.g., Battaglia et al., 2018). Since preliminary results showed comparable results, we decided avoiding adding this component in favour of the proposed one, as the latter given more interpretability to what each component represents.

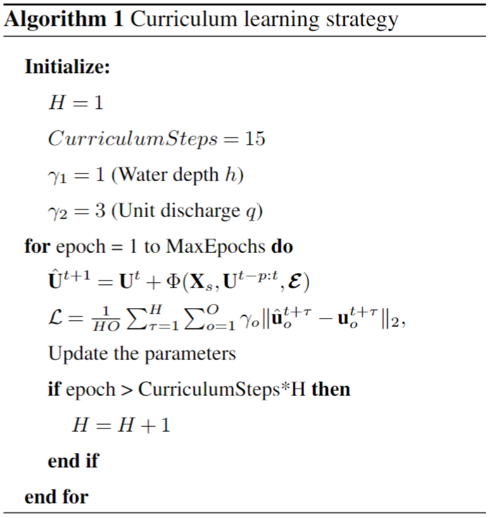
Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. and Gulcehre, C., 2018. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

10) L. 191. Can you clarify this sentence? I understand that you don't want areas with constant values in it, but in my opinion the sentence is formulated a bit vaguely.

10) As suggested as well by Reviewer 1, we modified this line as:
"Both the MLPs in the dynamic encoder and the decoder do not have the bias terms as this would result in adding non-zero values in correspondence of dry areas that would cause water to originate from any node."

11) Algorithm 1. I think you should at least initialize the weighting coefficients and the CurriculumSteps variable since they are used so explicitly (technically the other variables would need to be chosen too, but for them I think it's less crucial).

11) We thank the Reviewer for the suggestion. We modified Algorithm 1, including the initialization of weights and curriculum steps, as recommended.



Algorithm 1 Curriculum learning strategy

```
Initialize:
    $H = 1$
    $CurriculumSteps = 15$
    $\gamma_1 = 1$ (Water depth $h$)
    $\gamma_2 = 3$ (Unit discharge $q$)
for epoch = 1 to MaxEpochs do
    $\hat{\mathbf{U}}^{t+1} = \mathbf{U}^t + \Phi(\mathbf{X}_s, \mathbf{U}^{t-p:t}, \boldsymbol{\mathcal{E}})$
    $\mathcal{L} = \frac{1}{HO} \sum_{\tau=1}^{H} \sum_{o=1}^{O} \gamma_o \|\hat{\mathbf{u}}_o^{t+\tau} - \mathbf{u}_o^{t+\tau}\|_2,$
    Update the parameters
    if epoch > CurriculumSteps*H then
        $H = H + 1$
    end if
end for
```

12) L. 266. This should refer to the ablation study to show that the choice of H=8 is not arbitrary.

12) We thank the Reviewer for the suggestion. We included a reference to the sensitivity analysis in lines 266-267 as:

“We used a maximum prediction horizon H=8 steps ahead during training as a trade-off between model stability and training time, as later highlighted in Section 5.4.”

13) L. 267f. The sentence is a bit peculiar and maybe I missed it: I see that discharge is weighted with 3, but what is the weighting factor for the water depth (I assume 30)?

13) We understand that the considerations on the weighting factor are discussed rapidly. The motivation for weighting the different output components is that there in no normalization performed as pre-processing step and, as such, the values of water depth and unit discharge generally differ in magnitude by a factor of 10. Thus, logically, we would weight the discharge by a factor of 10 to achieve a similar goal as by normalizing them. However, we also consider that for application purposes water depth is more relevant that discharge. Consequently, we opted for a smaller weighting factor whose value of 3 was selected after some preliminary analysis.
Hence, we modified lines 267-268 to clarify this issue as:
“There is no normalization pre-processing step and, thus, the values of water depth and unit discharge differ in magnitude by a factor of 10. Since for application purposes discarge is less relevant than water depth (Kreibich et al., 2009), we weighted the discharge term by a factor of γ2 = 3 (cfr. eq. (14)), while leaving the weight factor for water depths as γ1 = 1.”

14) L. 277. MAE: I guess technically this is the mean of the mean absolute errors per variable, since you calculate the error in u_o (i.e., the Mean of the L1 for each hydraulic variable) and not in u (i.e. the L1 norm over the vector that spans all hydraulic variables). For me your choice makes more sense anyways. More importantly: What is of interest to me here, would be to see how the MAE changes if you include the weighting factors of the loss in the evaluation and get an weighted MAE (so that water depth is more important).

14) We thank the Reviewer for pointing this out. There is a mistake in how the testing MAE was presented: as highlighted later on in the results sections, all test RMSE and MAE are computed independently for each hydraulic variable. Accordingly, during evaluation, the MAE of the different hydraulic variables are never computed together and, as such, there is no need to perform a weighted average among them.

We corrected lines 283-288 as:

“We evaluated the performance using the multi-step-ahead RMSE (eq. (14)) over the whole simulation. However, for testing, we calculated the RMSE for each hydraulic variable o independently as:

$$RMSE_o = \frac{1}{H}\sum_{\tau=1}^{H}||\hat{u}_o^\tau - u_o^\tau||_2 \qquad (15)$$

Analogously, we evaluated the mean average error (MAE) for each hydraulic variable o over the whole simulation as:

$$MAE_o = \frac{1}{H}\sum_{\tau=1}^{H}||\hat{u}_o^\tau - u_o^\tau||_1 \qquad (16)$$”