Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC3
https://doi.org/10.5194/hess-2021-614-AC3, 2022

Reply on RC3

Roberto Bentivoglio et al.

Author comment on "Deep Learning Methods for Flood Mapping: A Review of Existing Applications and Future Research Directions" by Roberto Bentivoglio et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-614-AC3, 2022

Reply to anonymous Reviewer #3

We thank the Reviewer for carefully reading our paper and providing very useful comments for its improvement. Here we provide answers to the issues raised along with details on the amendments to the original manuscript to be featured in the revision. Unless otherwise specified, reported line numbers refer to the updated version.

My concerns are that

1)-the review is currently not always very precise in distinguishing the contexts in which different approaches are relevant

We thank the Reviewer for raising this point. We have added comments on each approach throughout the paper, e.g., line 109 (referred to inundation maps): "These maps can then be used also as calibration data for other applications such as flood susceptibility or flood hazard mapping."

Another example is in lines 114-115 (referred to susceptibility maps): "However, it can provide reliable information when no quantitative data is available."

We also illustrated the main limitations of each flood map in lines 558-563: "Nonetheless, each of the presented maps has its own limitations. Susceptibility maps provide only qualitative results and rely on recorded flood events. Therefore, limited recorded data may lead to incorrect predictions. Moreover, it is important to design an appropriate model to integrate heterogeneous environmental information. Inundation maps mostly consider real events, thus they suffer from the acquisition method's problems. For example, satellites struggle to extract information below clouded areas (e.g., Meraneret al., 2020). Hazard maps, instead, are limited by the accuracy of the underlying numerical simulator."

2)-it lacks an assessment of which techniques were used over time (some now popular techniques were not available 2 years ago)

We thank the reviewer for this observation. We included temporal context in section 3.2.1 to specify why certain methods were applied only in the recent years: lines 275-277 "The late use of convolutional and recurrent models is motivated by their recent popularization and development, along with a rise in awareness of the ML advancements, contrary to fully-connected layers, that have a longer application history."

3)-it could occasionally be better at explaining concepts with a focus on the hydrological target group

We thank the Reviewer for pointing out this aspect. In this light, we amended relevant text throughout the paper, adding more examples and comments reflecting a hydrologic/hydraulic perspective. For instance, we added in lines 189-193 "We explain the concept of invariance and equivariance with an example. Consider a picture with a flooded area in its top-left corner and one with the same area but shifted to be in the bottom-right corner. An invariant model would predict that there is a flooded area in both images, while an equivariant model would also reflect the change in positions of the flood, i.e., identify that the flood is in the top-left corner in one case and in the bottom-right corner in the other. In this case, invariance and equivariance are associated to a spatial translation, but the same principle applies to other transformations, such as temporal translation."

Following this example, we also added a similar consideration in lines 668-672 (section 5.1.1): "Symmetries result in inductive biases, which address the curse of dimensionality by decreasing the required training data (e.g., Wang et al., 2020a) and enabling the processing of different data types, such as meshes. From a flooding perspective, symmetries can be understood and motivated by referring to the example in Section 2.2.1. For instance, analogously to translation, the rotation of a domain should result in an equivalent rotation of the predictions."

Another hydrologic explanation is in lines 385-287 "For example, if 90% of an area represents non-flooded areas, a model which assumes that there are no floods will have 90% accuracy."

4)-generalization of deep learning predictions to locations / events outside the training data is a key aspect that deserves a more prominent place in the paper. Currently this topic is raised in several subsections. It might be useful to provide an overview on what are actually the needs, which can then be used to discuss whether different approaches are conceptually able to meet this (and if this was/was not implemented in current research)

We agree with the Reviewer that the generalization of deep learning models should deserve a more prominent place as it is a valuable and difficult gap. Thus, we created a corresponding subsection in the knowledge gaps section (section 4.2) in our updated manuscript as follows:

"Generalization refers to the capacity of a model to extrapolate from a training dataset into unseen testing data. This means that a DL model can correctly predict scenarios unused in its development. This property is particularly relevant because training requires data, model development, and time. In the context of flood modelling, there are two main generalization objectives: (i) boundary conditions, i.e., different rainfall events, and (ii) topographical changes, i.e., different case studies. However, the transference between different areas is challenging for DL models because of the difference in input and output data. In fact, except for flood inundation mapping, most reviewed papers focused on generalizing different boundary conditions (e.g., Guo et al., 2021; Berkhahn et al., 2019). Instead, only a few papers tested the model on areas not considered during training. Löwe et al. (2021) could generate flood hazard maps for unseen areas within the same study region as the training dataset, as there was little variability of inputs and outputs. Zhao et

al. (2021b) instead pre-trained a model for flood susceptibility on an urban area and then used it for another similar area. They showed that pre-training improves predictions with respect to a model trained from scratch, both in cases of low and high data availability. These works show that such approaches are in their infancy and have been tested on limited datasets. A DL model which cannot generalize to new areas has to be trained every time for a new study case. Thus, it may have limited advantages over a hydraulic model, since it requires more effort, data, and time. Instead, a general DL model which can generalize to new areas could emphasize the advantages over numerical models. This concept was experimented also for rainfall-runoff modeling where DL models outperformed state-of-the-art alternatives in the prediction of ungauged basins in new study areas (Kratzert et al., 2019b)."

5)-comparisons of scores across papers need to be interpreted more carefully than what is currently the case. Scores are not necessarily computed in the same manner. In particular, non-flooded areas are not handled consistently in the literature, which has a major impact on the results.

We thank the Reviewer for the valuable comment. We revisited this section and added the following paragraph to identify the issue mentioned. Lines 397-400: "Moreover, since different works generally use different datasets, a comparison across them may not always be meaningful. Instead, our purpose here is to show that, for the same case study, DL tends to outperform traditional models."

Additionally, we thought that the issue of incomparable metrics could also be reflected in the absence of a unified dataset for the different flood applications. Thus, we added a new paragraph in the data availability knowledge gap in lines 643-650: "Another issue, which emerges also from Section 3.2.5, is the lack of a unified framework to compare different approaches with each other. This can be achieved by creating flood-based benchmark datasets for each mapping application. For flood inundation, some datasets have been already used across different works (e.g., Bonafilia et al., 2020). However, works on both flood susceptibility and hazard mapping consider different datasets, focusing on different geographic areas or flood types. One possibility could then be to unify different case studies in a single dataset, for each application, allowing to assess the validity of a model more objectively. For flood susceptibility, case studies with the same input availability could be merged in a dataset with many flood types, scales, and geographical areas. A similar reasoning could be made for flood hazard mapping, selecting, for each case study, initial and boundary conditions for specific return periods."

I think all of these issues can be adressed in a revision. I have provided detailed comments below.

Detailed comments

6)line 51: the automatic discovery of representations is "to some extent" possible. We are still dealing with an input output model. It is quite a common misunderstanding that deep learning can find "any representations", while many relations in hydrology are highly nonlinear and require careful consideration of the data.

We thank the Reviewer for this clarification. With appropriate data, deep learning can learn any representation, as a consequence of the universal approximation theory results (Hornik et al. , 1989). However, it is indeed relevant to choose the data with careful selection. Thus, we changed the sentence in line 52 of the original manuscript to avoid confusion and we further clarified that data still require accurate pre-processing and selection. Lines 52-53: "Nonetheless, data must be carefully selected according to the task

at hand."

7)line 126-137: I don't think the detailed overview of modelling approaches is needed in this review.

We thank the Reviewer for the suggestion. We reduced the length of this paragraph and described numerical models briefly. The section has been modified as follows: "Flood hazard maps are carried out by numerical models, which simulate flood events by discretizing the governing equations and the computational domain. We distinguish between one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) models with increasing complexity and, generally, accuracy (e.g.,Horritt and Bates, 2002; Teng et al., 2017)."

8)line 190 to 205: this text is somehow misplaced in this section. It is more an assessment of the properties of different techniques and it would probably make more sense to place it after the different layer types were introduced.

We thank the Reviewer for the observation. We placed here this section to introduce the necessity of models like convolutional or recurrent networks to overcome the shortcomings of fully connected layers. While writing the paper, we tried placing this section elsewhere (e.g., before and after introducing the different layers), but ultimately we agreed that the current configuration was more suitable, even if it may hinder the reading flow. Thus, we would prefer to keep it as it is.

9)Figure 2: I think it would help many readers if the figure illustrates that the convolutional kernels map many pixels to one. Also in the text (line 210), a simply explanation of the kernels (spatially weighted average where the weights are learned during optimization) may be helpful.

We thank the Reviewer for the precious comment. We added an explanation of what kernels in lines 199-200 : "A kernel represents a spatially weighted average which is applied to the input and where the weights are learned during optimization."

10)Table 2: I believe the correct citation for the work of Guo et al. is 2021, not 2020

Thank you for the observation. We modified the citation as correctly pointed out.

11)Section 3.2.4: The review is generally missing a section that discusses under which conditions a deep learning network can generalize, i.e. predict flooding in different locations

We thank the Reviewer for the comment. With the new section on generalization, we tried to discuss what are the current issues and limitations. However, the problem of generalization is complicated and minimum conditions needed to guarantee it are still an ongoing research even in the machine learning community.

12)Section 3.2.5: A key issue when assessing flood predictions (inundation and hazard) is the large number of zeros (often >95% of the dataset) which implies that, for example, accuracy scores almost per definition are in the order of 80% and above. This issue needs to be explained here. In addition, binary scores such as CSI are very vulnerable to double penalty issues.

Thanks for the valuable comment. We modified section 3.2.5 and added further clarification of the problems related to classification metrics, addressing the issues of imbalanced datasets and appropriate metrics for flood mapping. We did not address the issue of CSI since very few papers use that metric.

Section 3.2.5 has been modified as follows: "In supervised learning, we distinguish between regression and classification problems, depending on whether the target values to predict are continuous (e.g., water depth) or discrete (e.g., flooded vs non-flooded area), respectively. Depending on the task, we employ a different set of metrics to evaluate model performances.

Regression metrics are a function of the differences, or residuals, between target and predicted values. The most common metrics include the root mean squared error (RMSE), the coefficient of determination (R^2), and the mean average error (MAE). RMSE and MAE improve as they approach zero, while R^2 improves as it approaches one. In general, RMSE is preferred to MAE since it minimizes the standard deviation of the errors, thus decreasing the presence of extreme outliers. However, since these metrics are averaged on a domain, their comparison across different works requires careful attention.

Classification tasks can be either binary (e.g., building a model to predict flooded and non-flooded locations) or multi-categorical (e.g., classifying between permanent water bodies, buildings, and vegetated areas), according to the output number of classes. In the following discussion, we focus on the former, with concepts naturally extending to the second case. When computing binary classification metrics, flooded areas are generally represented as positive class, while non-flooded areas as negative class. The most common metrics for flood modelling are accuracy, recall, and precision, followed by other indices such as the area under the receiver operator characteristic curve. Accuracy represents the number of correct predictions over the total. While popular and easy to implement, this metric is inappropriate for imbalanced datasets, where some categories are more represented than others. For example, if test samples feature an average 90% non-flooded area, a naïve model constantly predicting no flooding will reach 90% accuracy, despite having wrong assumptions. Furthermore, since it may be better to overestimate a flooded area than to underestimate it, one could resort to metrics such as recall that account for false negatives and thus penalize models that cannot recognize a flooded area correctly. However, when used alone, recall can lead to similar issues to those described for accuracy, e.g., yielding a perfect score for a model always predicting the entire domain as flooded. Thus, for an exhaustive understanding of the model's performance, one should also consider metrics accounting for false positives, i.e., where the model misclassifies non-flooded areas as flooded. There are several possible metrics, such as the F1 score, the Kappa score, or the Matthews correlation coefficient, each with their drawbacks and benefits (e.g., Wardhani et al., 2019;Delgado and Tibau, 2019; Chicco and Jurman, 2020). A reasonable choice is the F1 score, which is the geometric mean of recall and precision, and it thus equally considers both false negatives and false positives. Another good example is the ROC (Receiver Operating Characteristic) curve that describes how much a model can differentiate between positive and negative classes for different discrimination thresholds (Bradley, 1997). The Area under the ROC curve (AUC) is often used to synthesise the ROC as a single value. However, the AUC loses information on which parts of the dataset the model performs best. For this reason, one should always interpret these results carefully, especially when comparing different studies. Our purpose here is to show that, for the same case study, DL tends to outperform traditional models."

Following this addition to the text, Tables 3 and 4 have been modified accordingly, prioritizing metrics such as F1 and AUC.

13)Section 3.4: In general, for flood inundation, it is not completely clear to me whether the authors focus on models that can predict flood inundation (in binary form) given some rainfall or on "gap filling" in remote sensing data. This needs to be checked in all related sections.

For most works, flood inundation consists of determining flooded and non-flooded areas from remote-sensing data, i.e., given a picture, the model determines which areas are

flooded. Only one paper does not consider remote sensing data (Dong et al.). We clarified this in the review in line 427: "Only Dong et al. (2021) differ from the other papers by considering sensors in place of flood pictures."

14)Line 472: Due to the 0 problem mentioned above, "slight" increases in accuracy may actually be linked to substantial changes of the quality of a model. The scores therefore need to be interpreted carefully and it is also not guaranteed that all papers computed scores in the same manner.

Thanks for the important observation. As discussed in comment 12), we specified that it is not guaranteed that all works compute scores in the same manner and that comparison requires careful attention.

15)Line 496: Pham et al. assessed flood conditioning factors, Löwe et al. performed a forward selection to identify relevant topographic variables, Zahura et al. tested feature importance in their random forest model

We thank the Reviewer for providing valuable suggestions. Since this section concerns flood hazard we now included Löwe et al. . However, we did not include Zahura et al., as they employ other machine learning models, while the review focuses only on DL.

16)Table 5: It is not clear to me why not all the papers performing hazard predictions where included in this table? In addition, the error scores may not be comparable across papers (0 problem or similar) which should be mentioned. Also speed up is a difficult quantity to compare, because it depends on the assumed number of numerical simulations that should be performed (e.g. if we assume that we have to assess flood hazard for 1000 rain events, then the speed up factor obtained by a neural network will be much higher than when only 10 events are considered). Most certainly, these assumptions are also not the same across papers and therefore not comparable.

We greatly appreciate this comment of the Reviewer and we address its different parts individually.

For flood hazard, as for the other applications, not every paper was included in the provided tables. Many papers do not provide information on computational times of both numerical and deep learning models and, thus, they do not report speed-up metrics. Nonetheless, to provide a general overview of every paper, we now included all papers in all the tables, even those not reporting speed-up metrics or comparison against other machine learning models.

As regards the error scores, we agree that they may not be comparable throughout the works as the scales and resolutions may differ. However, we believe that these errors, along with the study case area, provide a measure of the model's reliability. We mentioned the issues when comparing scores across different studies in lines 543-545: "However, the comparison of speed-up across different papers is often unrealistic since it depends on the number of performed numerical simulations and on the type of numerical model. A similar consideration persists for the error scores, as they depend on the scale of the case study and on its resolution."

Regarding the 0 problem, we agree that the scores may differ depending on the number of zeros in a domain, since most regression metrics are averaged. Thus, we added the following in lines 377-378: "However, since these metrics are averaged on a domain, their comparison across different works requires careful attention."

17)Section 4.2: The discussion on generalization abilities needs to be differentiated a bit more. Both Guo et al. 2021 and Löwe et al. 2021 consider terrain characteristics as input

to their models, and in Löwe et al. 2021 generating predictions outside the training dataset was explicitly the focus of the work. As mentioned by the authors, these approaches are in their infancy and have been tested on limited datasets, but these approaches do consider effects of e.g. the built environment in the form of 2D grids.

We thank the Reviewer for this observation. We expanded the discussion on generalization in section 4.2 by differentiating between generalization on boundary conditions (i.e., different rain events) and initial conditions (i.e., different topography).

Lines 612-609: "Generalization refers to the capacity of a model to extrapolate from a training dataset into unseen testing data. This means that a DL model can correctly predict scenarios unused in its development. This property is particularly relevant because training requires data, model development, and time. In the context of flood modelling, there are two main generalization objectives: (i) boundary conditions, i.e., different rainfall events, and (ii) topographical changes, i.e., different case studies. However, the transference between different areas is challenging for DL models because of the difference in input and output data. In fact, except for flood inundation mapping, most reviewed papers focused on generalizing different boundary conditions (e.g., Guo et al., 2021; Berkhahn et al., 2019). Instead, only a few papers tested the model on areas not considered during training. Löwe et al. (2021) could generate flood hazard maps for unseen areas within the same study region as the training dataset, as there was little variability of inputs and outputs."

18)Section 5.1: While investigating the possibility to consider mesh-based deep learning setups is an interesting direction, the authors present no argument why this should work better than convolutional approaches (which are also used for simulating fluid movements). Other than stated around line 610, they are simply a different data representation with advantages and disadvantages (mesh generation) and may or may not improve performance.

We thank the Reviewer for this comment. We provided arguments based on recent works suggesting that mesh-based models are better than convolutional neural networks for generalization, accuracy, and stability in fluid dynamics. This is expressed in lines 674-675: "There already exist promising works which simulate fluid dynamics with mesh-based GNNs, with increased generalization, accuracy, and stability, with respect to CNNs (e.g., Pfaff et al., 2020; Lino et al., 2021)."

We introduced the limitations of meshes, in lines 657-658: "Unstructured meshes, nonetheless, inherit similar problems as those typical of numerical models, such as mesh generation and the need of explicitly defining how each node is connected."

19)Line 648: From here on the text no longer focuses on meshes (which is the Section heading) but on physical conditioning.

We thank the Reviewer for this observation. We added a section named physics-based deep learning that includes physics-based neural networks and neural operators.

20)Line 656: I think a formulation that will be easier to understand for many readers is that the PINN can only be trained for a specific boundary condition (such as a specific rain event) and it is subsequently only able to simulate this specific event.

We greatly appreciate this comment. We modified lines 693-694 as follows: "However, PINNs can only be trained for a specific boundary condition (e.g., a specific rain event) and can subsequently only simulate that specific event (Kovachki et al., 2021)."

21)Line 656: FNOs need to be mentioned as one approach amongst many. DeepONets are

a widely known alternative and new approaches are constantly developed. The same is true for DGP in the following section.

We thank the Reviewer for the comment. Indeed, there are many possibilities and alternatives as regards those approaches. We added DeepONets and clarified that there are several approaches which can be used in lines 698-699: "While many approaches have been proposed, such as DeepONets (Lu et al., 2019) or multipole graph neural operator (Li et al., 2020), Fourier neural operators (FNO) have currently achieved the best results (Li et al., 2021)."

Moreover, we included a section related to Bayesian neural network in the probabilistic deep learning section. Lines 720-725: "Along with those related to the model's input, uncertainties are also present in the model's prediction. To account for this kind of uncertainty we can use Bayesian neural networks (BNN). BNNs are models with stochastic components trained using Bayesian inference. They assign prior distributions to the model parameters to provide an estimate of the model's confidence on the final prediction (Blundell et al., 2015). If, for different parameter sampling, the output is unvaried, then the model has a good confidence on the prediction and vice versa if different parameters give different results. Jacquier et al. (2021) used BNNs to determine the confidence intervals in flood hazard maps, providing a measure of the model's reliability."

22)Section 5.3: I don't see how GANs fix data scarcity issues (line 680). They are indeed an interesting approach for e.g. gap filling or the generation of rainfall scenarios, but they do not be trained and do not relieve us of the problem that e.g. flood observations are hardly available. The discussion in the first parts of this section goes in a very different direction than the transfer learning approaches (which focus on training models with few data), which creates confusion.

We thank the Reviewer for the comment. We understand the doubts issued with GANs and VAEs. We have now mentioned that they do not solve the issue of a complete lack of data but can be useful in many situations where little data is available. For example, we could use the data of the GAN as augmentation to the few real data we have; then these new GAN data can provide more training samples for larger DL models, which would be more challenging with fewer data. GANs can also be used to generate floods in areas never experienced before, without necessarily being fed to a DL model afterward.

Following those considerations, we clarified the issue and modified section 5.3 as follows.

Moreover, we added a section on new data sources, which was previously introduced in the knowledge gaps (section 4.4). We also decided to remove the paragraph on transfer learning to avoid confusion.

"Even though remote sensing and measuring stations provide noticeable amounts of data, several parts of the world still lack enough data to deploy deep learning models. New satellite missions and added sensor networks throughout the world increasingly provide new data sources (e.g., van de Giesen et al., 2014). The flexibility of DL partially overcomes data scarcity by facilitating the use of a wider variety of data sources. For instance, several papers already employ cameras to detect floods and measure the associated water depth (e.g., Vandaele et al., 2021; Jafari et al., 2021; Moy De Vitry et al., 2019). Structural monitoring with cameras can provide reliable data where it was previously hard to obtain, such as in urban environments. Social media information, such as tweets or posted pictures, can also be used to identify flood events and flooded areas (e.g., Rossi et al., 2018; Pereira et al., 2020). In this case, the quality of the retrieved information must be further validated before its use for real applications. Moreover, the heterogeneity of the sources of these data needs to be carefully taken into account when deploying a suitable DL model.

Another approach can be to generate artificial data to supplement scarce data. This can be done using generative adversarial networks (GAN), which create new data from a given dataset (Goodfellow et al., 2014). GANs are composed of two neural networks, named generator and discriminator, whose purpose is, respectively, to generate new data and to detect if a given data is real or fake. A trained GAN can produce new fake but plausible data, facilitating data augmentation, i.e., providing more training samples. Interesting applications of GANs could overcome some limitations of satellite data (Lütjens et al.,2020, 2021), predict flood maps (Hofmann and Schüttrumpf, 2021) or meteorological forecasts (Ravuri et al., 2021), and create realistic scenarios of flood disasters for projected climate change variations (Schmidt et al., 2019). GANs could also be used to generate a plausible urban drainage system or topography for cities that do not have any sewers construction plan or in areas where only low-resolution data is available (e.g., Fang et al., 2020b).

However, GANs are difficult to train (Goodfellow, 2016). Variational autoencoders (VAE) are another type of generative model, which can overcome this issue. Differently from standard autoencoders, VAEs model the latent space with probability distributions that aim to ensure good generative properties to the model (Kingma and Welling, 2013). Once the model is trained, new synthetic data can be generated by taking new samples from the latent distributions. Nonetheless, because of the model's definition, the predictions are less precise than GANs. As such, VAEs and GANs offer a trade-off between the reality of the prediction and the availability of training data."

Conclusions

23)First bullet - this conclusion could be more clear about the methodological preferences being the current status which is developing rapidly.

We appreciate the comment and have clarified for each application which are the methodological preferences, based on the methods proposed so far (and thus excluding future direction models).

Lines 743-751: "Flood inundation, susceptibility, and hazard mapping were investigated using deep learning models. Flood inundation considers as the main data images of floods, mostly taken via satellite. The main and most accurate deep learning models were CNNs. In flood susceptibility, deep learning models consider several inputs, the most important being slope, land use, aspect, terrain curvature, and distance from the rivers. The main deep learning model used were MLPs, often in combination with other statistical techniques, but CNNs generally provided more accurate results. So far, flood hazard maps estimate the water depth in a study area by using deep learning as a surrogate model for numerical simulations. For this application, there are no deep learning model preferences. However, RNNs are preferable for spatio-temporal simulations. Regardless of the application, results show that deep learning solutions outperform traditional approaches as well as other ML techniques."

24)Line 724 - I would say DL for hazard mapping so far relies on numerical simulations, this may change.

Thank you for the suggestion. As mentioned, indeed, flood hazard models may also consider real flood events for training. We added "so far" in line 747, as shown in question 22, to indicate this.

Further comments are presented in section 3.5.2, e.g., in lines 514-515 "Even though observed data were not employed, they could be used in future research to corroborate

the transferability of such methods."

25)Line 731-736 - Some of the existing architectures do enable generalization but this certainly requires more research and testing. Meshes are one way forward amongst others.

Thank you for the observations. We changed "cannot" with "struggle to".

26)Line 737-741 - Physics-informed learning is not only relevant in a warning context but for virtually all kinds of flood simulations. FNOs and DGPs are potentially interesting approaches, but there are others. Your are overstating the ability of geometric DL which (to my knowledge) has not been tested in the flood context.

We thank the Reviewer for the comment. As discussed before, we modified the Future Research Directions section to include more possible models. Indeed, as regards physics-based learning, many models can benefit from it. We rewrote this section clarifying those issues as follows: "Physics-based deep learning provides a reliable framework for flood modelling since it considers the underlying physical equations. Probabilistic hazard mapping can take advantage of deep Gaussian processes or Bayesian neural networks to determine the uncertainties associated with the model and its inputs."

While geometric DL has not been used yet in flood context, based on recent findings (e.g., Pfaff et al. 2020, Lino et al. 2021, Wang et al. 2021), we believe that it may be a valuable tool for flood modelling.

27)Line 742-745 - As mentioned before, there is some logic here that does not make sense, because the GANs need to be trained against observed data. Once we have a GAN, what would be the point of training another deep learning model that only learns to emulate the output of the GAN?

We greatly appreciate this comment. Following the comments previously addressed (comment 22), we modified the paragraph in the conclusions as follows (lines 784-787): "DL necessitates large quantities of data which are difficult to collect in several areas of the world. New data sources such as camera pictures and videos, or social media information can potentially be used thanks to deep learning models. Moreover, generative models, such as GANs and VAEs, can be employed to produce synthetic data for such data-scarce regions, based on training data collected elsewhere"

References:

Löwe, R., Böhm, J., Jensen, D. G., Leandro, J., & Rasmussen, S. H. (2021). U-FLOOD – topographic deep learning for predicting urban pluvial flood water depth. Journal of Hydrology, 603, 126898. https://doi.org/10.1016/j.jhydrol.2021.126898

Pham, B. T., Luu, C., Phong, T. Van, Trinh, P. T., Shirzadi, A., Renoud, S., Asadi, S., Le, H. Van, von Meding, J., & Clague, J. J. (2020). Can deep learning algorithms outperform benchmark machine learning algorithms in flood susceptibility modeling? Journal of Hydrology, 592(July 2020), 125615. https://doi.org/10.1016/j.jhydrol.2020.125615

Zahura, F. T., Goodall, J. L., Sadler, J. M., Shen, Y., Morsy, M. M., & Behl, M. (2020). Training machine learning surrogate models from a high-fidelity physics-based model: Application for real-time street-scale flood prediction in an urban coastal community. Water Resources Research, 56(10), e2019WR027038. https://doi.org/10.1029/2019WR027038

References:

Hornik, K., Stinchcombe, M. and White, H., 1989. Multilayer feedforward networks are universal approximators. Neural networks, 2(5), pp.359-366.

Dong, S., Yu, T., Farahmand, H. and Mostafavi, A., 2021. A hybrid deep learning model for predictive flood warning and situation awareness using channel network sensors data. Computer□Aided Civil and Infrastructure Engineering, 36(4), pp.402-420.

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. and Battaglia, P.W., 2020. Learning mesh-based simulation with graph networks. arXiv preprint arXiv:2010.03409.

Lino, M., Cantwell, C., Bharath, A.A. and Fotiadis, S., 2021. Simulating Continuum Mechanics with Multi-Scale Graph Neural Networks. arXiv preprint arXiv:2106.04900.

Wang, R., Walters, R. and Yu, R., 2020. Incorporating symmetry into deep dynamics models for improved generalization. arXiv preprint arXiv:2002.03061.