EGUSPHERE-2025-3368 | Research article

Title: An Integrated Deep Learning Framework Enables Rapid Spatiotemporal Morphodynamic Predictions Toward Long-Term Simulations
Author(s): Mohamed M. Fathi et al.
MS No.: egusphere-2025-3368
MS type: Research article

The authors appreciate the peer review process facilitated by the Editor at *ESurf*. We extend our sincere thanks to the reviewers for their valuable feedback, which has significantly contributed to the enhancement of this manuscript.

Reviewer #2

This paper explores the efficiency and accuracy of using Deep Learning (DL) for predicting of geomorphic (river bed) change, in comparison with using a 2D morphodynamic model. The paper is clearly written and for the most part the figures are understandable and illustrative. I have never seen a study like this before, and I think this is a solid contribution to the literature. It represents a first step into using DL for estimating bed change. The study is relatively limited in scope - at least it seems that way to me. That said, I think this is a great introduction, represents an entire study that others could build on, and it is worthy and ready for publication. For reference, I have personally never used machine learning. I am an experienced numerical modeler but the traditional type. I am excited about the potential of studies like this. However, I am not in a position to review the details on the presentation of HDL-GM and whether "reasonable" choices were made in all cases.

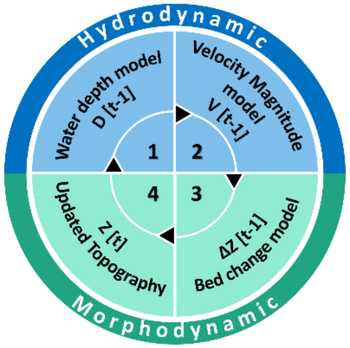
The authors sincerely appreciate the reviewer's time and effort in reviewing this paper. We also value your expertise in identifying missing or unclear points that could have led to misunderstandings, which ultimately enhanced the paper's readability.

The following table contains the authors' responses.

Comment	Responses (Green) and manuscript modifications (Blue)
1. Given this, I did wonder if there was rationale for the applying HDL-GM to 6 flood events. Was there a reason for 6? There is discussion of how error changes when doing continuous modeling, but was there a reason to stop at 6 events? Is it related to the simulation time required for HEC-RAS? Some discussion of how far forward we can predict would be appreciated by me. Even if we don't know, acknowledging this needs to be studied would be helpful for me. Similarly, I wondered why 11 events were used for training. Is there a relationship between	The authors thank the reviewer for raising this important point regarding the number of events used in the training and testing stages. In general, increasing the size of training datasets may have the potential to enhance the ability of DL models to learn hidden patterns and improve predictive robustness. However, in the context of morphodynamic modeling, the generation of long datasets is constrained by the expensive computational cost of physics-based models to generate training datasets. Furthermore, extending the dataset further would elongate the DL training process, and our main aim was to provide an efficient framework. In this study, a constructed hydrograph consisting of 20 flood events was selected to span a broad range of flood magnitudes and hydrograph shapes. Repetition of similar events was intentionally avoided, as it is unlikely to provide additional information to the DL framework while substantially increasing computational cost. The partitioning of these 20 events into 11 training, 3 validation, and 6 testing events follows common practice in data-driven modeling, where approximately 50–60% of the data are used for training, with the remainder reserved for validation and independent testing. The six testing events were therefore not

the number of events used for training and the number of events that can be reasonably predicted?	chosen to represent a predictive limit of the framework, but rather to provide a statistically meaningful and computationally feasible basis for evaluating error accumulation and long-term behavior. The revised manuscript reads as follows: A hydrograph with a broad range of flood events is essential to enhance the capabilities of the DL framework in capturing hidden patterns in the mapping between inputs and outputs. Unfortunately, generating sufficiently long training and testing datasets is hindered by the long processing time using the physics-based model. Additionally, significant portions of the observed hydrograph are baseflow, which corresponds to a period of relatively low flow with minimal sediment transport. To address these challenges, a constructed hydrograph of 20 events was extracted from the observed hydrograph. This event set was designed to encompass a broader range of flood magnitudes while avoiding long periods of baseflow discharge. Repetition of similar events was intentionally avoided, as such redundancy is unlikely to provide additional informative signals for the DL framework. This hydrograph was split into three portions: training, validation, and testing of 11, 3, and 6 events, respectively (Figure 1-b), following common data-driven modeling practice to balance model learning and independent evaluation within computational constraints. At present, the relationship between the number of training events and the maximum number of flood events that can be reliably predicted in a continuous morphodynamic framework remains an open research question. This may be constrained by numerous factors, such as the system size, hydrological responses, sediment transport behavior, and the DL architecture itself. We now explicitly acknowledge that the forward prediction horizon requires systematic investigation of longer-term predictive stability in future work. The revised manuscript reads as follows: These findings highlight the capabilities of the CB-trained framework to predict morphodynamic behavior across a testing dataset encompassing a series of continuous events. Nevertheless, while the HDL-GM framework remains stable, the maximum forward prediction horizon over which robust morphodynamic predictions can be maintained requires further investigation, beyond the scope of this study, and is expected to depend on numerous factors such as hydrologic forcing, geomorphic nonlinearity, and the inductive biases of the DL architecture itself.
2. Another thing I wondered was whether you could use a shorter reach for training but make predictions on a longer reach of river than was used for training. Is that fair? It could be a way to save some training time. Is this a no-no in geomorphology machine learning world?	This is an excellent and timely question. In the present study, training and prediction were performed on the same spatial domain. In fact, this is a critical limitation for such DL algorithms. We acknowledged this clearly in the conclusion section as follows: Despite these promising capabilities, its primary limitation lies in its inability to effectively simulate unseen topographical conditions. Additionally, the current implementation of the HDL-GM framework is for a particular set of hydrodynamic and sediment transport properties, especially Manning n, grain-size distributions, and cohesive properties. To address these limitations, future work is needed to improve the generalizability and performance by integrating essential hydrodynamic and sediment transport physics into the DL framework (Karniadakis et al., 2021; Mohamad et al., 2021).

	Recent advances in temporal DL have enabled models to be trained simultaneously across multiple locations and to generalize predictions to ungauged or unseen sites. Nevertheless, these approaches are predominantly limited to temporal representations and have not yet been extended to fully spatiotemporal modeling frameworks. The field is evolving rapidly, and ongoing efforts, both by our team and in the broader community, are focused on addressing these regionalization challenges.
3. The authors do discuss caveats and limitations, so it's not that those are missing. The points in the two above paragraphs were ones that came up for me as I read, but if the authors feel that addressing these is not useful, I defer to them.	The authors appreciate the reviewer's comment. We have acknowledged this limitation, particularly given that this study is among the first to employ spatiotemporal deep learning frameworks for morphodynamic simulations. We agree that the community is actively working toward more efficient models with improved generalizability to unseen scenarios.
4. L 134,135: What impact does the trapezoidal cross-section assumption make? I did not see a mention of river depth measurements, but wouldn't those be needed to create the cross-sections?	The authors thank the reviewer for highlighting this missing point. The trapezoidal cross-section was introduced due to the absence of bathymetric data in the DEM, following field measurements reported by Costigan et al. (2014). The revised manuscript reads as follows: Due to the absence of bathymetry information in the DEM file, a trapezoidal cross-section profile was burned into the DEM file (Choné et al., 2018), using a channel depth of 1.1 m and a bankfull width of 95 m, informed by field measurements reported by Costigan et al. (2014).
5. Figure 2C: This is a little confusing to me. In C, isn't it the topography at time t-1 that informs the water depth at time t? Based on this diagram, there is just a continual loop at time t. I think, that in the loop from slice 4 to 1, there is an increase in t. Maybe you could show that in the arrow between slice 4 and slice 1?	The reviewer is correct, and we appreciate this careful reading. The framework operates through an integrated four-step loop: (1) hydrodynamic variables are predicted using inputs from topography time t-1; (2) these hydrodynamic predictions are then used to estimate bed change at time t-1; (3) the predicted bed change is applied to update the topography to time t; and (4) the updated topography is subsequently used as the primary input for the next prediction step. The original version of Figure 2C did not sufficiently emphasize the increment in time between successive iterations, which may have led to confusion. The revised figure is as follows:

	c) Models' integration  Hydrodynamic Water depth model D [t-1] Velocity Magnitude model V [t-1] 1 2 3 4 ΔZ [t-1] Bed change model Z [t] Updated Topography Morphodynamic
6. L 225-227: I think the reference dataset is from HEC-RAS and the simulated is from HDL-GM? Maybe you could say that directly, or clarify what they are if I am wrong.	The reviewer is correct in their interpretation: in our terminology, the simulated results correspond to the DL model outputs, whereas the reference denotes the physics-based model results. The revised manuscript now clarifies this distinction explicitly in the relevant sections. To evaluate the accuracy of the proposed model in predicting the hydrodynamic and bed change variables, Root Mean Square Error (RMSE) is used to quantify the average magnitude of errors between HDL-GM predictions and reference results obtained from the HEC-RAS model
7. Equations 1 - 3: I think R and S are bed elevation change. Can you say that explicitly?	Agreed. We planned to use those performance criteria to evaluate the models' performance at both hydrodynamic and morphodynamic levels. That's why we didn't clearly state that those are bed change criteria. However, we agree that explaining this clearly enhances the readability of the paper. where R_i, S_i are the physics-based reference values obtained from HEC-RAS and the corresponding simulated values predicted by the HDL-GM framework, respectively, and $\bar{R}$ is the average value of reference grids. N is the number of active cells within the predicted or reference domains. These performance criteria are applicable to both the hydrodynamic components (water depth and flow velocity) and the morphodynamic component (bed elevation change).
8. Paragraph around line 245, about the EB testing: For the EB scenario, does every event start on the same initial topography? That is event 1 and event 2 are processed in parallel, but starting on the same topography?	Yes. In the EB framework, each flood event is simulated independently and initialized from the same baseline topography. The authors have revised the manuscript to enhance readability and to present this experiment more clearly for all readers. Two distinct versions of the HDL-GM framework were tested. These versions are based on the nature of the training dataset: Event-Based (EB) and Continuous-Based (CB) datasets (Figure 1-b). The EB dataset aggregates discrete events that are simulated independently using the physics-based model, each initialized from the same topographic state. This formulation

	enhances computational efficiency in generating the training dataset by allowing parallel simulations. Conversely, the CB dataset is derived from a single temporally continuous simulation that spans a sequence of events, which demands significantly longer processing time due to its continuous simulation to resolve the full temporal evolution of the system.
9. Figure 3: Is absolute error the absolute value of R-S at a location? And the relative error is relative to what? I don't see equations for these values. Couldn't relative error exceed 100%? Is the color bar saturated at 100% but values go beyond 100%?	Yes. Absolute error is defined as |R–S|. Relative error is computed as the absolute error divided by the reference error at each grid cell. As noted by the reviewer, the relative error can exceed 100%. For visualization purposes, the color bar is intentionally saturated at 100% to improve figure clarity, as values above this threshold uniformly indicate very poor agreement regardless of their exact magnitude. We also added a full paragraph in the methodology section to define how to compute both absolute error and relative error. Although NRMSE and R^2 provide useful global measures of model performance, geomorphic responses are spatially heterogeneous and cannot be fully described by a single performance value. To better capture spatial error patterns, cell-scale absolute and relative error maps were developed, comparing the physics-based HEC-RAS reference results with the HDL-GM predictions. Absolute error represents the absolute difference between the reference and predicted values at cell-scale, with a perfect value of zero and no upper bound. Relative error expresses this absolute error normalized by the reference value at each cell, enabling comparison across regions with differing response magnitudes. While relative error also attains an optimal value of zero, it may exceed 100% in areas where predictions are large relative to the reference bed change values or the reference bed change values are marginal. Performance criteria along spatial error metrics provide a more complete assessment of both the magnitude and significance of prediction errors across the domain.
10. L 312: "95% error criterion indicates that 95% of the active cells exhibit an error below this value." What is "this value"?	The authors appreciate the reviewer's comment. We agree that the original phrasing was unclear. We revised it as follows: The 95% error is defined as the threshold below which 95% of the cell-scale absolute errors across all active cells fall, corresponding to the 95th percentile of the absolute error distribution. Consequently, it offers additional insight beyond mean-based statistics by quantifying the magnitude of errors affecting the vast majority of active cells, defined here as cells exhibiting bed changes greater than 0.02 m.