Supplementary Material

Catchment	Rain Events		DEM Min Value (m)	DEM Max Value (m)
	Train	Test		
1	3	1	46.079	427.208
2	3	1	-15.436	132.709
3	5	1	15.202	89.664

Table 1. Additional statistics for each catchment. For each catchment, one rain event was reserved for testing. The minimum and maximum DEM values give a measure of the terrain elevation of each catchment.

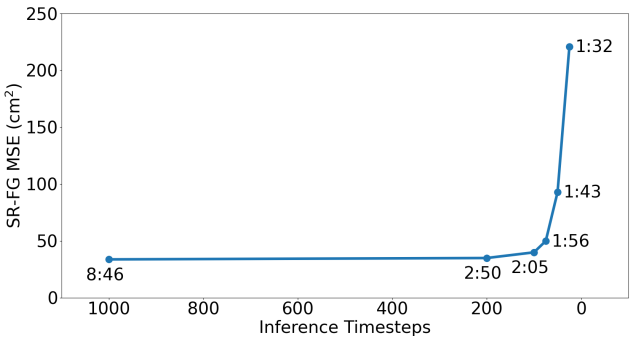

Figure 1. Graph of SR-FG MSE (cm^2) against the number of inference timesteps using the LDM trained on Catchment 1, starting from random noise. The graph annotations indicate the time taken to perform inference on 1000 images at varying timestep counts. The number of inference timesteps can be reduced from 1000 to 200 without significant degradation in performance, lowering inference time from 8 minutes and 46 seconds to 2 minutes and 50 seconds.

A. Data

Table 1 contains additional data statistics for each catchment. Reserving a rain event for testing allows us to accurately evaluate the models on unseen simulations, ensuring the model's ability to perform well on future rainfall events.

B. Inference Time Reduction

Figure 1 illustrates how the SR-FG MSE changes as the number of inference timesteps is reduced when starting the backward diffusion process from random noise. We are able to reduce the number of timesteps by up to 80% before seeing a noticeable decrease in performance.

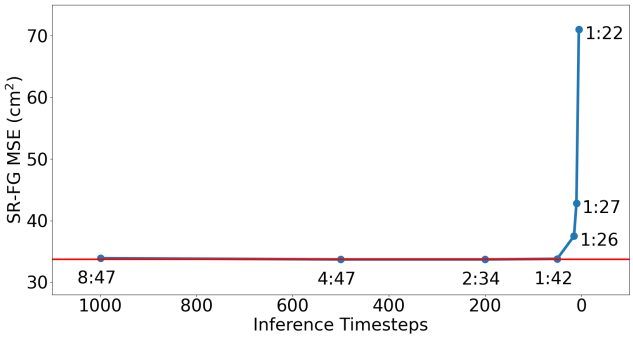

Figure 2. Graph of SR-FG MSE (cm^2) against the number of inference timesteps using the LDM trained on Catchment 1, starting from the noisy coarse-grid image. The horizontal red line shows the SR-FG MSE obtained using 1000 timesteps and starting from random noise. By starting the reverse diffusion process from the noisy coarse-grid image and using 50 timesteps, we are able to reduce the time taken to 1:42, a speed-up of 5×.

Figure 2 shows the SR-FG MSE at different number of inference timesteps when starting from the noisy coarse-grid image. This allows us to achieve a greater reduction in inference timesteps and processing time before any decrease in performance. Overall, we were able to obtain 5× speed-up by starting the backward diffusion process from an intermediate output.

B.1. Further decrease in MSE on Catchment 2

A particularly noteworthy outcome was observed for Catchment 2, as shown in Fig. 3. In this case, initiating the backward diffusion process from the noisy coarse-grid image not only reduced inference time but also led to a lower SR-FG MSE compared to the baseline approach of starting from random noise. Given that Catchment 2 is the most complex among the three, and exhibits the highest CG-FG MSE, this result suggests that the standard 1000 timesteps may be insufficient for the latent diffusion model to adequately reconstruct the fine-grid flood map from random noise. By starting from an input that more closely resembles the desired output, the model is effectively able to bypass part of the denoising process, thereby achieving improved accuracy with fewer inference steps.

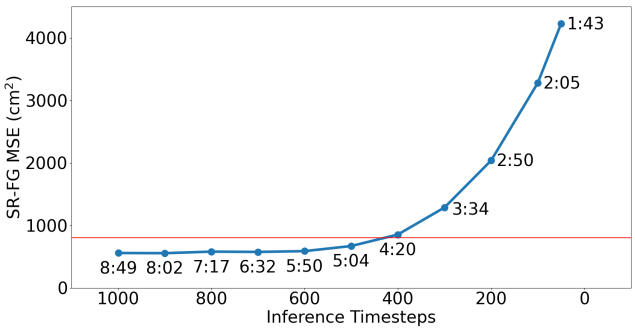

Figure 3. Graph of SR-FG MSE (cm^2) against the number of inference timesteps using the LDM trained on Catchment 2, starting from the noisy coarse-grid image. The horizontal red line shows the SR-FG MSE obtained using 1000 timesteps and starting from random noise. Initializing the backward diffusion process from the intermediate output produces a lower SR-FG MSE compared to starting from random noise.

C. Generalizability

Tables 2 and 3 show the performance of each model on unseen catchments. These tables correspond to the alternate configurations of the generalizability experiment in the main paper, where the models were trained on Catchment 1/2 and evaluated on the other unseen catchments. The overall pattern of the results is the same: diffusion-based models show significantly better performance on unseen catchments; indicative of their superior generalization capability.

Tables 4 and 5 show the performance of the LDM in transfer learning scenarios. These tables correspond to the alternate configurations of the transfer learning experiment in the main paper, where LDMs were initially trained on each catchment before being finetuned on Catchment 1/2. Similarly, the results show that the transfer-learned LDMs are able to obtain comparable results despite only being trained on the finetuning catchment for a much shorter period. This highlights the capability of LDMs to rapidly adapt to new geographical regions by starting from an existing pretrained model.

Training Catchment	Model	Test Catchment 2			Test Catchment 3		
		CG-FG MSE (cm^2)	SR-FG MSE (cm^2)	% change ↓	CG-FG MSE (cm^2)	SR-FG MSE (cm^2)	% change ↓
1	SGUnet [11]	5957.4	5714.2	-4.08	158.7	1895.4	+1094.08
	DM (ours)	5957.4	4535.8	-23.86	**158.7**	**242.8**	**+52.95**
	LDM (ours)	**5957.4**	**4272.8**	**-28.28**	158.7	559.3	+252.33

Table 2. All 3 models were trained on data from Catchment 1 and subsequently evaluated on the unseen Catchments 2 and 3. The diffusion-based architectures outperform the SGUnet by a large margin, showcasing their increased generalization ability in zero-shot settings over CNN-based architectures.

Training Catchment	Model	Test Catchment 1			Test Catchment 3		
		CG-FG MSE (cm^2)	SR-FG MSE (cm^2)	% change ↓	CG-FG MSE (cm^2)	SR-FG MSE (cm^2)	% change ↓
2	SGUnet [11]	344.2	6436.4	+1770.20	158.7	4084.2	+2473.02
	DM (ours)	344.2	2200.9	+539.50	**158.7**	**461.3**	**+190.62**
	LDM (ours)	**344.2**	**1222.2**	**+255.14**	158.7	2886.7	+1718.59

Table 3. All 3 models were trained on data from Catchment 2 and subsequently evaluated on the unseen Catchments 1 and 3. The diffusion-based architectures outperform the SGUnet by a large margin, showcasing their increased generalization ability in zero-shot settings over CNN-based architectures.

Representation Learning Catchment	Finetuning Catchment	CG-FG MSE (cm^2)	SR-FG MSE (cm^2)	% change ↓
1	-	344.2	33.7	-90.20
2	1	344.2	52.0	-84.88
3	1	344.2	56.2	-83.67

Table 4. LDM Performance after being finetuned on Catchment 1 data for 50,000 steps. The first row is the performance of the original LDM trained on Catchment 1 data for 400,000 steps with no transfer learning. The transfer-learned LDMs produced MSE reductions that were close to that of the baseline LDM despite only being trained on Catchment 1 data for one-eighth of the number of steps, showcasing the significant potential of transfer learning in our LDM architecture.

Representation Learning Catchment	Finetuning Catchment	CG-FG MSE (cm^2)	SR-FG MSE (cm^2)	% change ↓
2	-	5957.4	723.0	-87.86
1	2	5957.4	1585.4	-73.39
3	2	5957.4	1438.1	-75.86

Table 5. LDM Performance after being finetuned on Catchment 2 data for 50,000 steps. The first row is the performance of the original LDM trained on Catchment 2 data for 300,000 steps with no transfer learning. The transfer-learned LDMs produced MSE reductions that were close to that of the baseline LDM despite only being trained on Catchment 2 data for one-sixth of the number of steps, showcasing the significant potential of transfer learning in our LDM architecture.