COMPARING RECURRENT AND CONVOLUTIONAL NEURAL NETWORKS FOR PREDICTING WAVE PROPAGATION

Stathi Fotiadis, Eduardo Pignatelli, Anil A. Bharath
Department of Bioengineering
Imperial College London

Mario Lino Valencia, Chris D. Cantwell
Department of Aeronautics
Imperial College London

Amos Storkey
Institute for Adaptive and Neural Computation
The University of Edinburgh

ABSTRACT

Dynamical systems can be modelled by partial differential equations and a need for their numerical solution appears in many areas of science and engineering. In this work, we investigate the performance of recurrent and convolutional deep neural network architectures to predict the propagation of surface waves governed by the Saint-Venant equations. We improve on the long-term prediction over previous methods while keeping the inference time at a fraction of numerical simulations. We also show that convolutional networks perform at least as well as recurrent networks in this task. Finally, we assess the generalisation capability of each network by extrapolating for longer times and in different physical settings.[1]

1 INTRODUCTION

Many physical systems in science and engineering are described by partial differential equations (PDEs). This study investigates the performance of recurrent and convolutional deep neural networks to model such phenomena. Accurately predicting the evolution of such systems is usually done through numerical simulations, a task that requires significant computational resources. Simulations usually need extensive tuning and need to be re-run from scratch even for small variations in the parameters. With their potential to learn hierarchical representations, deep learning techniques have emerged as an alternative to numerical solvers, by offering a desirable balance between accuracy and computational cost (Carleo et al., 2019).

Here, we focus on the modelling of surface wave propagation governed by the Saint-Venant (SV) equations. This phenomenon offers a good test-bed for controlled analyses on two-dimensional sequence prediction of PDEs for several reasons. First, in contrast to some physical systems, such as fluid flow, the evolution of the real system is unlikely to enter chaotic regimes. From a representation learning point of view, this makes model training and assessment relatively straightforward. Despite this, the SV equations are strongly related to the Navier-Stokes equations, widely used in computational fluids. Further, computational modelling of surface waves is used in seismology, computer animation, in predictions of surface runoff from rainfall – a critical aspect of the water cycle (Moussa & Bocquillon, 2000) – and flood modelling (Ersoy et al., 2017).

This study provides three contributions. First, we identify three relevant architectures for spatiotemporal prediction. Two of these architectures lead to improved accuracy in long-term prediction over previous attempts (Sorteberg et al., 2019) while keeping the inference time orders of magnitude smaller than typical solvers. Secondly, our comparison between recurrent and purely convolutional models indicates that both can be equally effective in spatiotemporal prediction of SV PDEs. This is in alignment with the findings of Bai et al. (2018) that demonstrates that convolutional models are as effective as recurrent models in one-dimensional sequence modelling. Finally, we evaluate the generalisation of the models in situations not seen during training and indicate their shortcomings.

[1] Code and data available at `github.com/stathius/wave_propagation`

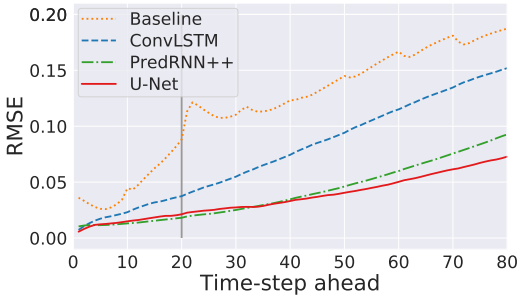

Figure 1: Comparing the long-term prediction of the four models on the test set. The Causal-LSTM and the U-Net significantly outperform the baseline LSTM model. The vertical line indicates the training horizon.

Time-step ahead	20	80
LSTM (baseline)	0.08 ± 0.00	0.19 ± 0.03
ConvLSTM	0.05 ± 0.00	0.15 ± 0.01
PredRNN++	$\mathbf{0.02 \pm 0.01}$	0.09 ± 0.01
U-Net	$\mathbf{0.02 \pm 0.00}$	$\mathbf{0.07 \pm 0.01}$

Table 1: Root Mean Square Error (RMSE) comparison of model prediction at specific time-steps ahead. Best accuracy in bold. The standard errors are across 4 runs with different initialisation.

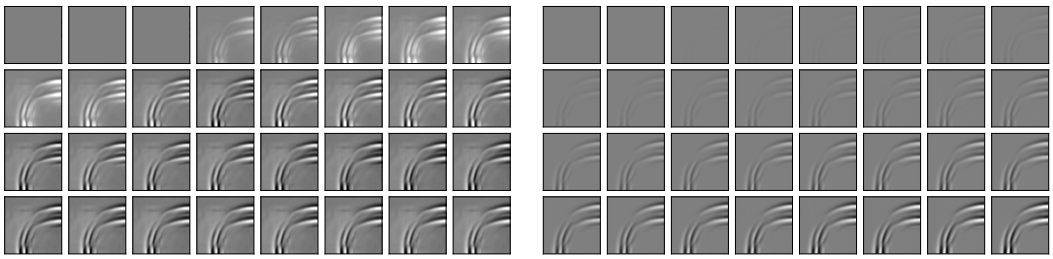

Figure 2: Cumulative reconstruction of the output from the feature maps of the pre-last layer of the U-Net (top) and PredRNN++ (bottom). Prediction corresponds to the 80th time-step ahead. We ordered the feature maps by the absolute value of their weight, from the most important to the least. The PredRNN++ gradually builds up its prediction. The U-Net works differently: the first few feature maps put emphasis on the boundary conditions. Then some of the feature maps focus on the peaks (white colour) and some others on the troughs. All combined build the final prediction.

2 Related work

Deep learning methods have been proposed for spatiotemporal forecasting in various fields including the solution of PDEs. Recurrent neural networks have been proven a good fit for the task, due to their innate ability to capture temporal correlations. Srivastava et al. (2015) use a convolutional encoder-decoder architecture where an LSTM module is used to propagate the latent space to the future. Variations of this technique have been successfully applied to the long-term prediction of physical systems, such as sliding objects (Ehrhardt et al., 2017) and wave propagation (Sorteberg et al., 2019). Convolutional LSTMs (ConvLSTM) use convolutions inside the LSTM cell to complement the temporal state with spatial information. Whilst initially proposed for precipitation nowcasting, ConvLSTMs were also found successful for video prediction (Shi et al., 2015). Wang et al. (2018a) proposed the PredRNN++, featuring spatial memory that traverses the stacked cells in the network and improves the accuracy of short-term prediction over ConvLSTMs.

Feed-forward models have, also, been used in spatiotemporal forecasting. Mathieu et al. (2015) used a CNN to encode video frames in a latent space and extrapolated the latent vectors to the future. Tompson et al. (2017) employed CNNs to speed up the projection step in fluid flow simulations. U-Net has been used for optical flow estimation in videos (Dosovitskiy et al.) as well as in physical systems, such as sea temperature predictions (de Bezenac et al., 2017) and accelerating the simulation of the Navier-Stokes equations (Thuerey et al., 2018). While both recurrent and convolutional models have been successfully applied for the prediction of PDEs, there is a paucity of studies comparing the two categories from a representation learning point of view.

Other architectures for spatiotemporal prediction include Generative Adversarial Networks, for fluid simulations (Kim et al., 2018) and Graph Networks for wind-farm power estimation (Park & Park, 2019). There is also a growing body of research on physics-inspired networks for solving PDEs (Raissi et al., 2017; Perdikaris & Yang, 2019).

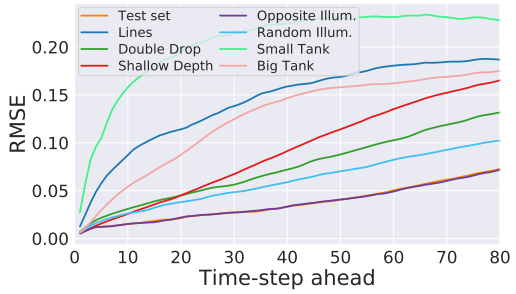

Figure 3: Generalisation in different physical settings for the U-Net. The network copes well with changes to illumination and even with two drop but cannot predict well linear waves or different tank size.

Time-step ahead	20	40	60	80
Test set	**0.02**	**0.03**	**0.05**	**0.07**
Opposite Illum.	**0.02**	**0.03**	**0.05**	**0.07**
Random Illum.	0.4	0.06	0.08	0.10
Double Drop	0.04	0.07	0.10	0.13
Lines	0.11	0.16	0.18	0.19
Shallow Depth	0.04	0.09	0.13	0.16
Big Tank	0.08	0.14	0.16	0.17
Small Tank	0.19	0.22	0.23	0.23

Table 2: RMSE of U-Net across datasets at specific points in time. Performance varies across different physical settings. The model is invariant to an orthogonal phase shift in illumination.

3 Evaluated models

Four different models are assessed in this work. Three of them are recurrent (LSTM, ConvLSTM, PredRNN++) and one is feed-forward (U-Net). A detailed description of all the implementations can be found in Section B of the Appendix. The LSTM model was specifically developed for wave propagation prediction (Sorteberg et al., 2019) and serves as a baseline on which we sought improvement. It is composed of a convolutional encoder and decoder with three LSTMs in the middle. The LSTM modules use the vector output of the encoder as an inner representation and propagate it forward in time. Each LSTM propagates a different part of the sequence (see Appendix).

The other models were selected on the basis of their applicability to relevant tasks. ConvLSTM and PredRNN++ have been empirically shown to perform well at short-term spatiotemporal predictions. The rationale for using them in long-term prediction is that the underlying physics of wave propagation do not change. If a model learns a good representation of short-term dynamics, then the error accumulation should remain low long-term. Both models use convolutions inside the recurrent cell to create a synergy between spatial and temporal modelling. Additionally, PredRNN++ employs a spatial memory that traverses the vertical stack to increase short-term accuracy.

The feed-forward model is based on the U-Net architecture used in spatiotemporal prediction. For example, it has been used to infer optical flow (Fischer et al., 2015) , motion fields (de Bezenac et al., 2017) and velocity fields (Thuerey et al., 2018). In contrast, we train the network end-to-end and conditional on its own predictions; the latter shifts the focus from short-term to long-term accuracy.

4 Results

4.1 Long term prediction: Extrapolation in time

We evaluated how well the models extrapolate in time. Given ground-truth simulations of 100 frames in length, we tested the model predictions up to 80 steps, much more than the maximum of 20 frame sequences that the models are trained upon. The RMSE at each time step is calculated as an average over all the test sequences. Results show that the baseline LSTM gives the worst performance. The RMSE error reaches 0.10 after only 21 frames while the error sharply raises after frame 10 (Figure 1). A probable cause is the usage of three distinct LSTMs, which require more data to train upon. The ConvLSTM offers an improvement: it reaches 0.1 RMSE after only 53 frames. The error trend is also very gradual, almost linear. An even greater improvement comes from the PredRNN++, which provides a very low error over the whole prediction range. Its maximum error at frame 80 is 0.091, substantially lower than the LSTM (0.186) and the ConvLSTM (0.150) (Table 1). This confirms the findings of Wang et al. (2018b), that PredRNN++ is more efficient than ConvLSTM. U-Net is on par with PredRNN++ until frame 34, but has better long-term prediction, reaching 0.071 RMSE at frame 80 vs 0.091 of the PredRNN++. The U-Net decreases the RMSE by 62% compared to the baseline. It is also the faster model, providing a $240\times$ speed-up over the numerical solver that we used (Table 6 in Appendix).

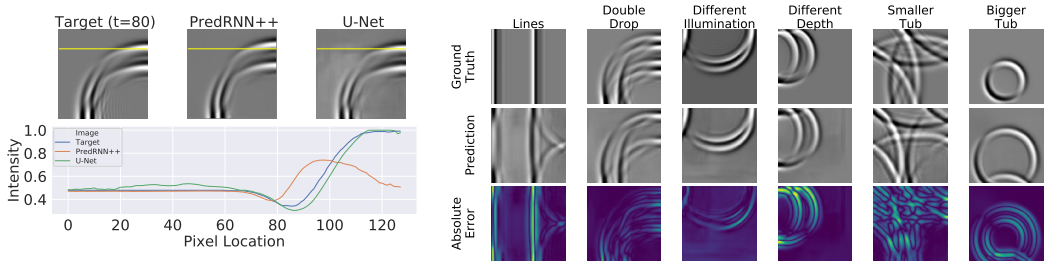

Figure 4: Left: Qualitative comparison between the U-Net and the PredRNN++ on the test set. The intensity profile corresponds to the yellow line (the line with the highest variation). In this particular case, the PredRNN++ has missed the time constant. **Right: Predictions (at time $t = 80$) of the U-Net in the various dataset that have not been seen during training.** In double drop, we see how the model fails to accurately predict the double wave-front. For bigger and smaller tub it misses the time constant.

Qualitatively, it appears that the PredRNN++ propagates its internal representation one step at a time while the U-Net predicts multiple frames in one pass. How the output is reconstructed in the last layer is indicative of the differences (Figure 2).

4.2 Generalisation: Extrapolation in other physical settings

Here, we evaluate the capabilities and limitations of our models by testing under different initial conditions, illumination models and tank dimensions (Table 3 in Appendix). For conciseness, we only present the results of the U-Net but the same conclusions stand for all the models.

The U-Net seems to be quite robust to changes in illumination. The RMSE for opposite illumination angle ($135°$) is indistinguishable to the original test set (Figure 3 and Table 2). This indicates that the learned representation is invariant to a perpendicular phase shift in lighting conditions. Propagation of linear waves appears to be more challenging, RMSE exceeds 0.10 after just 12 frames. The visualisation shows how the morphology of the prediction is qualitatively different, containing circular artefacts, reminiscent of the training data (Figure 4). When two drops are used, the RMSE is fairly low but the two wave-fronts of the predictions are sometimes blurred. We also varied the tank size to study the effect of wave speed. It seems that both cases are challenging with the smaller tank size, or equivalently faster waves, exceeding 0.10 RMSE after just 5 frames. Predictions in Figure 4 demonstrate how the network miscalculates the wave speed, and its predictions are either faster or slower than the ground truth. Please note that direct comparisons between datasets based on the RMSE is not without shortcomings. Each dataset has its own inherent "variation" which affect the RMSE, i.e. waves move faster in a small tank (see Figure 11 in the Appendix for a discussion).

5 Conclusions and Future Work

In this work we investigated the use of deep networks for approximating wave propagation. Using a U-Net architecture, we managed to reduce the long-term approximation RMSE to 0.071 against the previous baseline of 0.186. At the same time, the U-Net is $240\times$ faster than the simulation. Our results suggest that the U-Net outperforms state-of-the-art recurrent models. It is unclear why U-Net models perform so well in this task. It been demonstrated that convolutional networks are effective at modelling one-dimensional temporal sequences (Bai et al., 2018); it might be true for higher-dimensional data. Furthermore, the simulated data are based on few-step solvers. In such a case the memory modules may not offer a significant advantage. Lastly, we extensively assessed how the networks generalise in unseen physical settings and pointed out current limitations.

In the future, we aim to introduce noise in the simulation so the system becomes stochastic. It would be interesting to see if in this case the recurrent models learn the dynamics better than the U-Net. A big shortcoming of the current models is generalisation in other physical settings. We plan to address this by a physics-inspired latent space factorisation and meta-learning.

ACKNOWLEDGMENTS

Eduardo Pignatelli, Anil Bharath and Chris Cantwell would like to acknowledge the support of the Rosetrees Trust. Stathi Fotiadis is supported by a Doctoral Scholarship from the Department of Bioengineering at Imperial College London.

REFERENCES

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. 2019. URL http://arxiv.org/abs/1903.10563.

Nicolas Cellier. Scikit-fdiff / skfdiff. https://gitlab.com/celliern/scikit-fdiff/, 2019. [Online; accessed 11-8-2019].

Emmanuel de Bezenac, Arthur Pajot, and Patrick Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. *arXiv preprint arXiv:1711.07970*, 2017.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Hazırbas¸ Hazırbas¸, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks. Technical report.

Sebastien Ehrhardt, Aron Monszpart, Niloy J. Mitra, and Andrea Vedaldi. Learning A Physical Long-term Predictor. 3 2017. URL http://arxiv.org/abs/1703.00247.

Mehmet Ersoy, Omar Lakkis, and Philip Townsend. A saint-venant shallow water model for overland flows with precipitation and recharge, 2017.

Philipp Fischer et al. Flownet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015.

Byungsoo Kim, Vinicius C. Azevedo, Nils Thuerey, Theodore Kim, Markus Gross, and Barbara Solenthaler. Deep Fluids: A Generative Network for Parameterized Fluid Simulations. 6 2018. doi: 10.1111/cgf.13619. URL http://arxiv.org/abs/1806.02071http://dx.doi.org/10.1111/cgf.13619.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

Roger Moussa and Claude Bocquillon. Approximation zones of the saint-venant equations f flood routing with overbank flow. *Hydrology and Earth System Sciences Discussions*, 4(2):251–260, 2000.

Junyoung Park and Jinkyoo Park. Physics-induced graph neural network: An application to wind-farm power estimation. *Energy*, 187:115883, 11 2019. ISSN 03605442. doi: 10.1016/j.energy.2019.115883.

Paris Perdikaris and Yibo Yang. Modeling stochastic systems using physics-informed deep generative models. 2019.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations. 11 2017. URL http://arxiv.org/abs/1711.10561.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. 6 2015. URL http://arxiv.org/abs/1506.04214.

Wilhelm E Sorteberg, Stef Garasto, Chris C. Cantwell, and Anil A. Bharath. Approximating the Solution of Surface Wave Propagation Using Deep Neural Networks. In *INNS Big Data and Deep Learning*, 2019. doi: 10.1007/978-3-030-16841-4{_}26.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pp. 843–852, 2015.

Nils Thuerey, Konstantin Weissenow, Lukas Prantl, and Xiangyu Hu. Deep Learning Methods for Reynolds-Averaged Navier-Stokes Simulations of Airfoil Flows. 10 2018. URL http://arxiv.org/abs/1810.08217.

Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, and Ken Perlin. Accelerating eulerian fluid simulation with convolutional networks. In *ICML*, pp. 3424–3433. JMLR. org, 2017.

Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. PredRNN++: Towards A Resolution of the Deep-in-Time Dilemma in Spatiotemporal Predictive Learning. 4 2018a. URL http://arxiv.org/abs/1804.06300.

Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. *arXiv preprint arXiv:1804.06300*, 2018b.

APPENDIX

A DATASETS

The datasets were created by simulating the Saint-Venant equations:

$$\begin{aligned} h_t + ((H+h)u)_x + ((H+h)v)_y &= 0 \\ u_t + uu_x + vu_y + gh_x - \nu\left(u_{xx} + u_{yy}\right) &= 0 \\ v_t + uv_x + vv_y + gh_y - \nu\left(v_{xx} + v_{yy}\right) &= 0 \end{aligned} \tag{1}$$

The package triflow (Cellier, 2019) was used for the simulation. The Coriolis force and viscosity terms were neglected, kinematic viscosity was $10^{-6}m^2/s$ which is close to water viscosity at $20°$C, the height H is set to 10 m and the size of the tank is randomly selected in each simulation between 10 and 20m. The initial wave excitation is in the form of a Gaussian droplet at random locations. For rendering, we used $45°$ lighting azimuth and $20°$ altitude. Each sequence is 100 steps long while the time step is 0.01 sec. In total, 3,000 sequences were rendered. The frame size was 184×184 pixels but was subsequently re-sampled down to 128×128. The generalisation datasets were created with the same method by varying the physical properties of the simulation (Table 3).

We also used image normalisation which is known to improve performance on image prediction tasks. Normalising the pixel values to zero mean and standard deviation 1 worked best for us. Note that the normalising values are computed from the training set alone and applied to the validation and test sets. Data augmentation techniques like horizontal and vertical flips were employed on a per sequence basis. From the 3000 sequences of the original dataset, 70% were used for training, 15% for validation and 15% for testing.

Dataset Name	Initial Condition	Height(m)	Tank Size(m)	Illum. Azimuth	Sequences
Training/Validation/Test	Droplet	10	[10, 20]	$45°$	3000
Double Drop	Double Droplet	10	[10, 20]	$45°$	500
Lines	Line wave	10	[10, 20]	$45°$	500
Opposite Illumination	Droplet	10	[10, 20]	$135°$	500
Random Illumination	Droplet	10	[10, 20]	Random	500
Shallow Depth	Droplet	5	[10, 20]	$45°$	500
Small Tank	Droplet	10	[5, 10]	$45°$	500
Big Tank	Droplet	10	[20, 40]	$45°$	500

Table 3: The original dataset was used for training, model selection and evaluation. Models were also trained with the fixed tank dataset to study the effect of tank size. All the other datasets were used to evaluate the generalisation capabilities of the models.

B MODELS

B.1 LSTM

The encoder consists of 4 convolutional layers with 60, 120, 240, 480 feature maps, kernel sizes 7, 3, 3, 3 and padding of 2, 1, 1, 1 pixels. Dimensionality reduction is achieved by using a kernel stride of size 2 in all layers. After each convolutional layer, there is a batch normalisation layer and a *tanh* non-linearity. In the last convolutional layer, dropout is used on 25% of the units, that are chosen randomly in each pass. The final part of the encoder is a fully connected layer of width $\mathcal{L} = 1000$. This is the latent vector input to the three LSTMs. One LSTM is used for the first input, the second LSTM is for predicting the 10th frame (midway) and the third LSTM for all the other frames. The decoder is based on deconvolutions that double the spatial dimensions of the feature maps in each layer until the original 128×128 size is reached. It is a mirror of the encoder in terms of feature map size while the kernel is 3, the padding is 1 and the stride is 2 for all the layers. Figure 6 depicts the architecture.

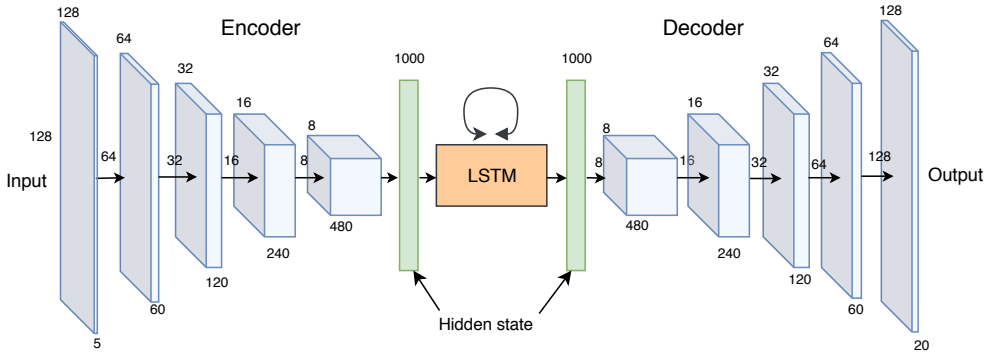

Figure 5: Schematic of the encoder and decoder used in the LSTM model (Sorteberg et al., 2019). Dimensions or layers are left and up, number of channels at the bottom.

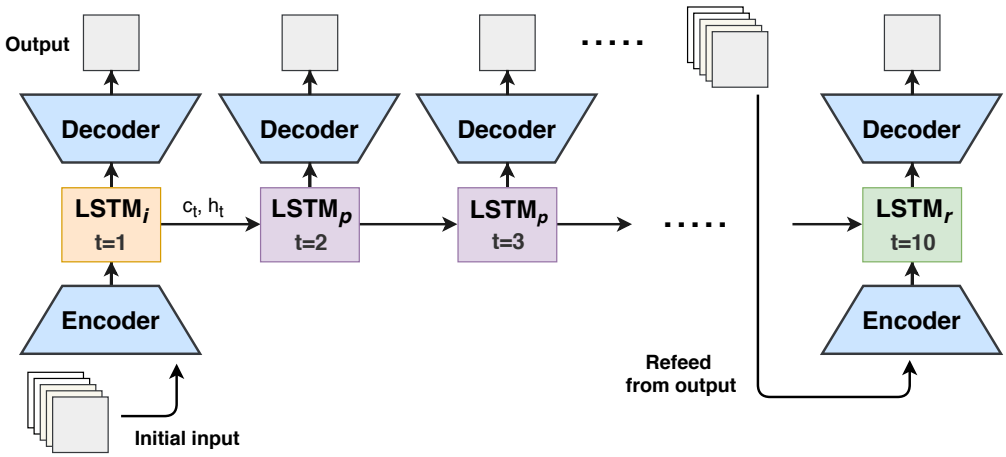

Figure 6: Schematic of the encoder and decoder used in the LSTM model (Sorteberg et al., 2019). Dimensions or layers are left and up, number of channels at the bottom.

B.2 CONVLSTM

Our architecture uses a stack of 3 ConvLSTM cells. Initially, a convolutional encoder with 8, 64, 192 feature maps respectively reduces the spatial dimensions to 31×31. All layers have kernels of size 3, zero padding of width 1 and Leaky ReLU non-linearities with slope 0.2. A stride of 2 pixels is used to reduce the dimensionality. At the final layer, the input is represented by a $16 \times 16 \times 192$

tensor. Inside the ConvLSTMs we use kernels of size 3 and zero padding of 1 pixel to avoid the dimensionality reduction. The decoder uses deconvolutions with stride 2 to double up the pixel dimensions in each layer.

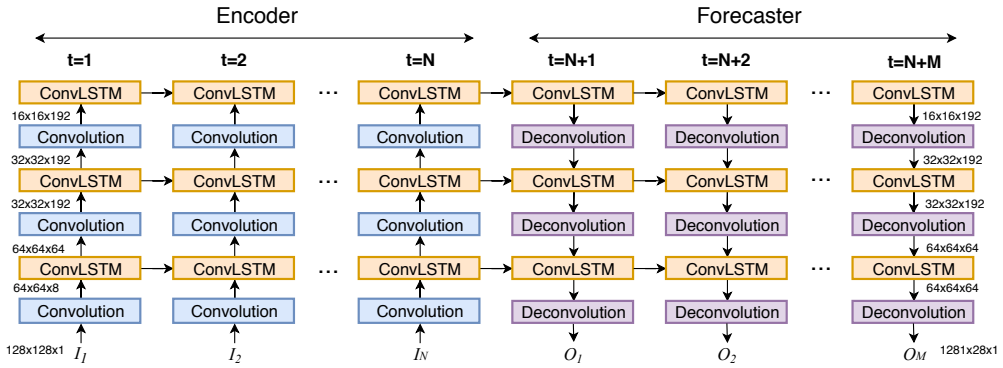

Figure 7: ConvLSTM Model The encoder processes the N=5 input frames one at a time to create an internal representation. The representation gets copied over to the forecaster that uses it to generate M=10 future frames. The feature map dimensions can be seen next to each layer.

B.3 PredRNN++

The unfolding of the model through time is presented in Figure 8. The vertical stack is comprised of one convolutional, one max pooling and four PredRNN++ layers. The convolutional layer has a kernel of size 3, no padding and outputs 8 feature maps. In the original paper, they do not use any dimensionality reduction because their input dimensions are 64×64 per frame. Our input dimensions (128×128) are too big to fit in available GPU memory, so we used max-pooling with stride 4 to reduce the dimensions to 31×31 pixels. Following the original paper, we used 4 PredRNN++ layers but reduced the size of all of them to 64 channels each to meet hardware memory constraints. We used convolutional kernels of size 3. The forecaster uses a deconvolutional layer kernel size 7 and stride 4 to restore the internal state to the original dimensions.

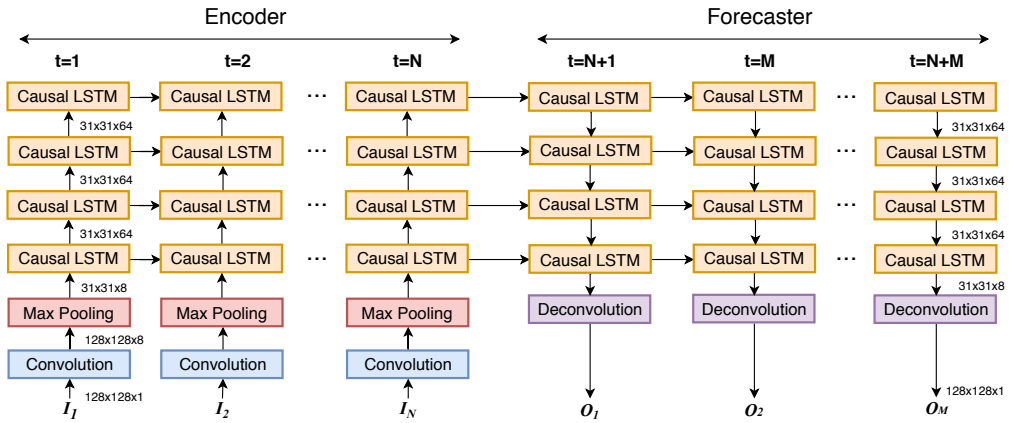

Figure 8: PredRNN++ Model The encoder processes the N=5 input frames one at a time and an internal representation is created in each PredRNN++ cell. These representations get copied over to the forecaster that uses them to generate M=20 future frames. The feature map dimensions can be seen next to each layer.

B.4 U-Net

The encoder is composed of fours blocks each containing two convolutional layers with kernel size 3 and padding 1, followed by ReLU non-linearities. The first three blocks include a max-pooling layer of stride 2 that reduces the size in half. The number of feature maps doubles in each layer. For the expanding part, we use bilinear interpolation with scale factor 2 instead of deconvolutions to

keep the number of parameters low. Skip connections are also employed to copy feature maps from earlier layers but contrary to the original paper we do not reduce the dimensions of the copied feature maps. This way, high-level, coarser feature maps are combined with fine-grained local information of lower layers over the whole domain. The network architecture can be seen in Figure 9.

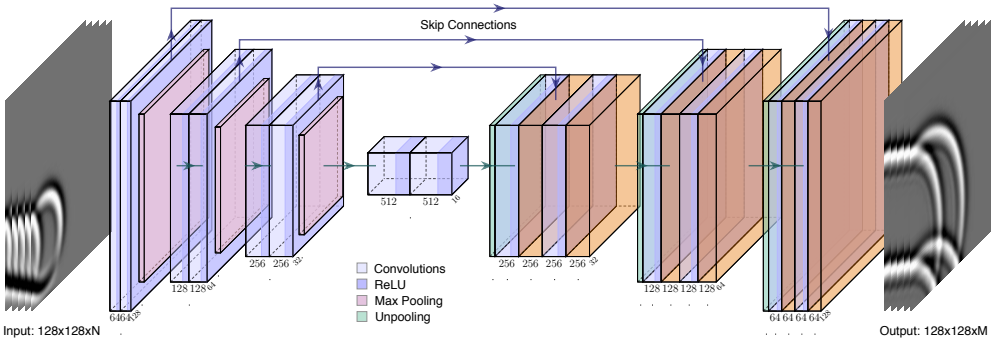

Figure 9: Schematic of the U-Net model. The number of channels is below each layer and its dimensions on the side. Our model has input N=5 and output M=20.

C HYPERPARAMETERS

Assume that N is the number of input and M the number of output frames of the model. For each training iteration, we randomly selecting K sub-sequences of length $N + M$ from each simulated sequence. The models were trained to minimise the MSE over their respective output length M. In each iteration, the weights are updated using an Adam optimiser while a scheduling scheme adjusts the learning rate (LR) by a scaling factor of 10^{-2} if there is no improvement in validation error after a given amount of epochs (patience). The hyperparameters of interest are the input length $N \in \{3, 5, 10\}$, training output length $M \in \{1, 5, 10, 20\}$, samples per sequence between weight updates $K \in \{1, 3, 5, 10\}$, batch size $b \in \{16, 32\}$, LR $\in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and patience $p \in \{5, 7\}$. Grid search was used to find the best set of hyperparameters of each model. The training budget was 24h hours. To obtain an arbitrary long prediction we the output as the next input. The goal is to obtain networks with a low error in long term prediction so, during model selection, we chose the hyper-parameters that gave the lowest validation error over 50 frames regardless of the output size of the model M. The final hyperparameters and model sizes can be found in Tables 4 and 5.

Model	Input Length	Output Length	Samples per Sequence	Batch Size	Learning Rate	Patience
LSTM	5	20	10	16	10^{-4}	5
ConvLSTM	5	10	5	8	10^{-3}	7
PredRNN++	5	20	5	4	10^{-4}	3
U-Net	5	20	10	16	10^{-4}	7

Table 4: Hyper-parameters of the best performing model for each architecture

D MODEL SIZE AND SPEED

Models were implemented in PyTorch and the code is publicly available in GitHub. Models were trained on a GTX 1060 GPU with 6GB of memory. Total training time includes evaluation overhead.

Model	**Trainable Parameters**	**Epoch Time**	**Num. Epochs**	**Best Epoch**	**Total Training Time**
LSTM	88.2M	12m	75	71	24h
ConvLSTM	12.3M	36m	24	18	24h
PredRNN++	2.5M	33m	43	36	24h
U-Net	7.8M	8m	171	166	24h

Table 5: Model size and training times

Method	Time per frame (ms)	Speed-up
Numerical simulator	630.7	-
LSTM	15.0	40x
ConvLSTM	4.5	141x
PredRNN++	9.2	68x
U-Net	2.6	241x

Table 6: Time it takes to compute one frame. Deep learning approximations offer a significant speed-up over numerical simulations.

E RESULTS ADDENDUM

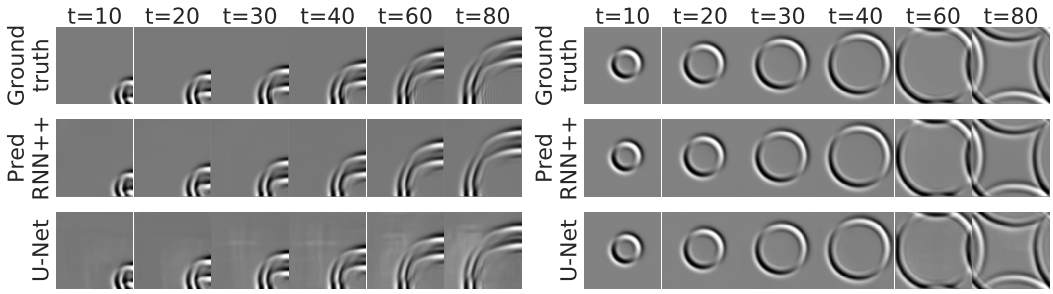

Figure 10: Prediction roll-out from U-Net and PredRNN++. Both sequences are from the test set.

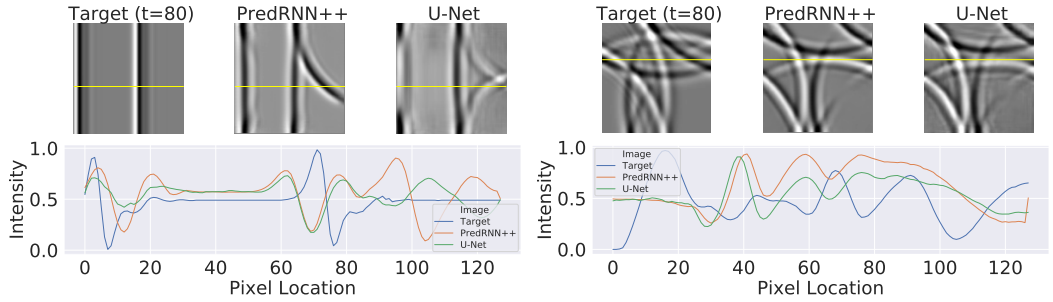

Figure 11: Generalization of U-Net and PredRNN++ in different physical settings. In the left we see linear waves. The networks introduce circular patterns where they don't exist. In the right panel it is the smaller tank, where waves are faster. Both models miss the time constant by being slower than the ground-truth.

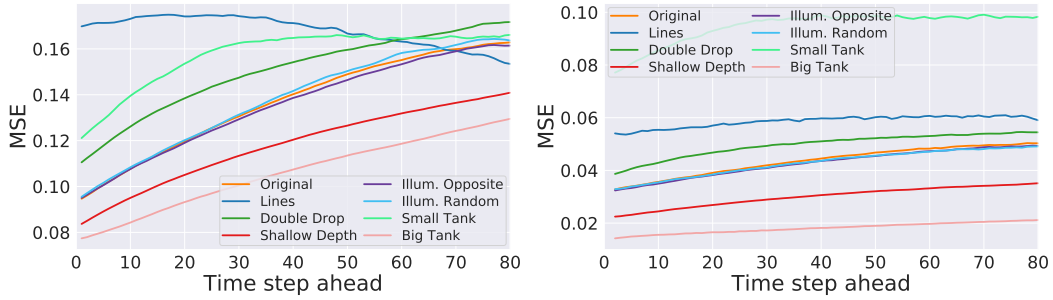

Figure 12: Flat Image (left) and **Previous Frame (right)** baselines compared against ground truth across different datasets. The flat image baseline measures how much the frame is away from the reference height. The Previous Frame indicates how much consecutive frames differ. Datasets are not equally hard to predict and this should be taken into account when assessing the generalisation capacity of a model.

E.1 PREDICTING THE TANK SIZE FROM THE LATENT SPACE

Here we check if the trained U-Net acquired any understanding of the physical properties of the system. We focus on the tank size, or inversely the speed of the wave, for two reasons. First of all, the U-Net failed to extrapolate to different tank sizes. This experiment could provide some insights on why this failure happens. Secondly, tank size information is readily available. Each dataset sequence corresponds to a different tank size, and the tank is always square. In the training and testing dataset we have tank size $s_i \in [10, 20]$ meters. For the smaller tank we used $s_i \in [5, 10]$ and for the bigger $s_i \in [20, 40]$ meters.

The question we try to answer is: does the latent representation of the U-Net capture that tank size information s_i. We take the pre-trained encoder from the U-Net and add some additional layers so that the output is only one number (Figure 13). The system is trained to predict the tank size when given 5 consecutive frame. Only the additional part is updated during training. The weights of the encoder are kept frozen. We compare the pre-trained encoder against a randomly initialised encoder. We, also, compare the models to a dummy regressor that predicts always the mean tank size for each dataset i.e. 15 for the test set, 7.5 for the small tank and 30 for the big tank. Results in Table 6 indicate that the pre-trained encoder can be used to extract the tank size with relatively low error (0.14) while the random encoder gives a much higher error of 2.27, slightly lower to the dummy regressor (2.45). This indicates that the pre-trained encoder encapsulates physically relevant information relating to the tank size. When it comes to the bigger and smaller tanks, both the pre-trained and the random encoders fail to extrapolate and give errors higher than the dummy regressor.

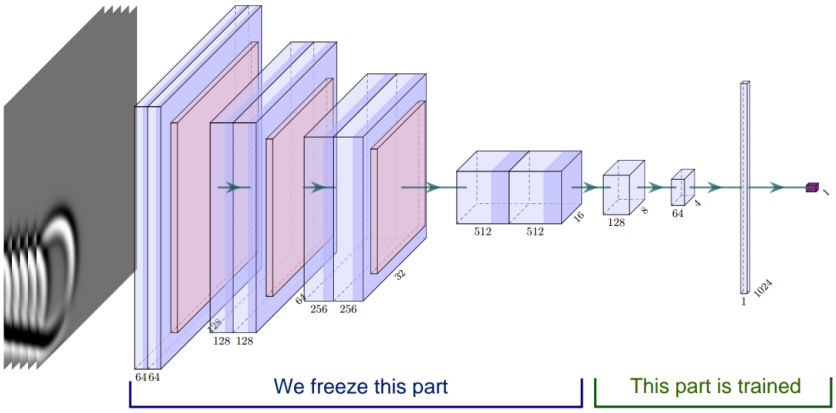

Figure 13: Schematic of the model used for tank size prediction

	Test set	Bigger Tank	Smaller Tank
Pre-trained encoder	**0.14**	6.65	2.23
Random encoder	2.27	14.21	6.35
Dummy regressor	2.45	**5.19**	**1.22**

Table 7: Predicting the tank size from the U-Nets latent space. The pre-trained model learns to identify the correct tank size with a relatively low error. The randomly initialised encoder fails to do so, only marginally improving the error over a dummy regressor. A dummy regressor is one that always predict the mean tank size of the dataset. Both the pre-trained and random encoder fail to extrapolate to the smaller and bigger tanks.