

Learning multi-field representations and physics-informed surrogates for active matter dynamics on The Well Benchmark

Project 01 — Predicting Emergent Dynamics

Jhon Villacís

Jorge Pico

The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning

Abstract

Machine learning based surrogate models offer researchers powerful tools for accelerating simulation-based workflows. However, as standard datasets in this space often cover small classes of physical behavior, it can be difficult to evaluate the efficacy of new approaches. To address this gap, we introduce *the Well*: a large-scale collection of datasets containing numerical simulations of a wide variety of spatiotemporal physical systems. The Well draws from domain experts and numerical software developers to provide 15TB of data across 16 datasets covering diverse domains such as biological systems, fluid dynamics, acoustic scattering, as well as magneto-hydrodynamic simulations of extra-galactic fluids or supernova explosions. These datasets can be used individually or as part of a broader benchmark suite. To facilitate usage of the Well, we provide a unified PyTorch interface for training and evaluating models. We demonstrate the function of this library by introducing example baselines that highlight the new challenges posed by the complex dynamics of the Well. The code and data is available at https://github.com/PolymathicAI/the_well.

Active Matter:

- **Varied Physical Parameters:**

$\alpha \in \{-1, -2, -3, -4, -5\}$ $\beta = 0.8$; $\zeta \in \{1, 3, 5, 7, 9, 11, 13, 15, 17\}$

- **Fields present in the data:**

Concentration (scalar field), velocity (vector field), orientation tensor (tensor field), strain-rate tensor (tensor field). 11 fields in total.

- **Data format:**

225 realizations : each with shape [81, 256, 256, 11]

```
Total number of simulations in the training dataset: 175
Total number of simulations in the validation dataset: 24
Total number of simulations in the test dataset: 26
```

```
Total number of simulations in the whole dataset: 225
```

```
In the training dataset there are 45 pairs of (zeta, alpha) values.
In the validation dataset there are 16 pairs of (zeta, alpha) values.
In the test dataset there are 21 pairs of (zeta, alpha) values.
```

kinetic model

$$\Psi(\mathbf{x}, \mathbf{p}, t)$$

Smoluchowski equation,

$$\frac{\partial \Psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\dot{\mathbf{x}} \Psi) + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}} \Psi) = 0,$$

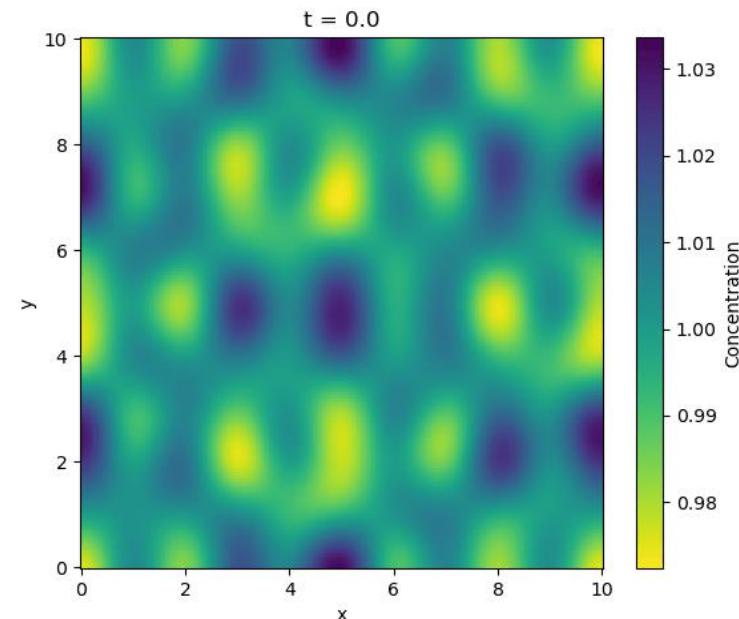
$$\dot{\mathbf{x}} = \mathbf{u} - d_T \nabla_{\mathbf{x}} \log \Psi,$$

$$\dot{\mathbf{p}} = (\mathbf{I} - \mathbf{p}\mathbf{p}) \cdot (\nabla \mathbf{u} + 2\zeta \mathbf{D}) \cdot \mathbf{p} - d_R \nabla_{\mathbf{p}} \log \Psi.$$

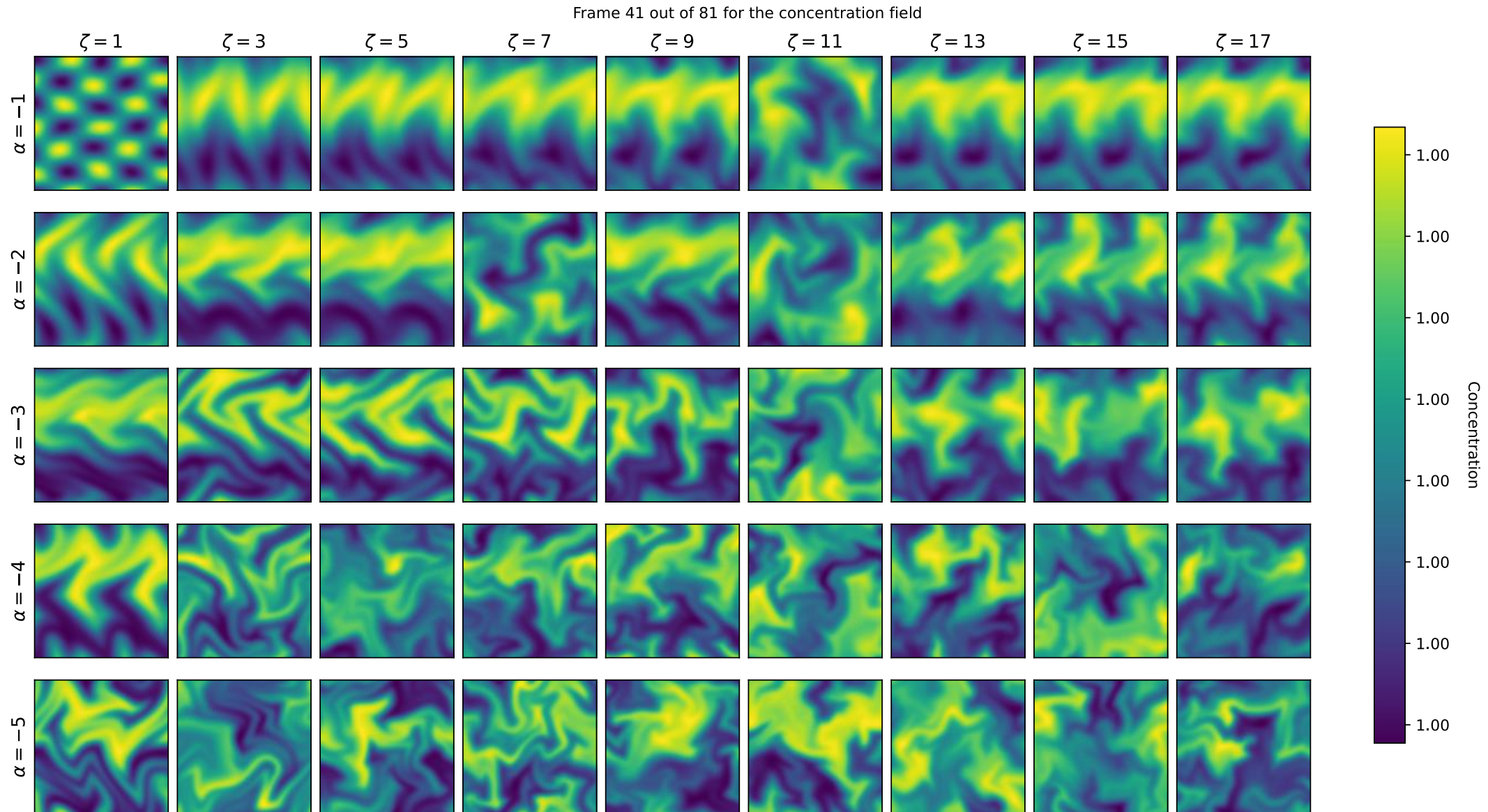
Stokes flow as

$$-\Delta \mathbf{u} + \nabla P = \nabla \cdot \boldsymbol{\Sigma}, \quad \nabla \cdot \mathbf{u} = 0,$$

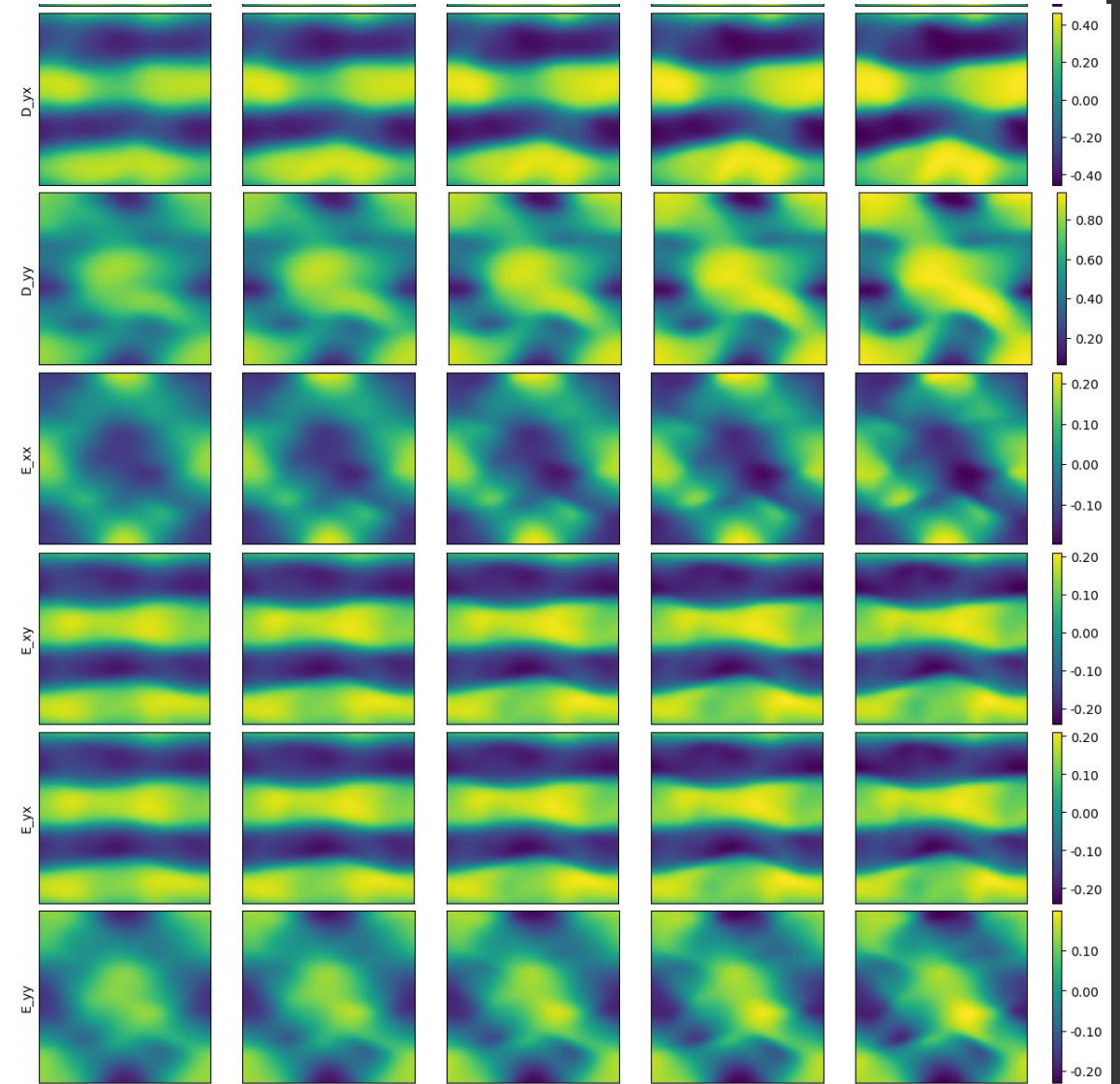
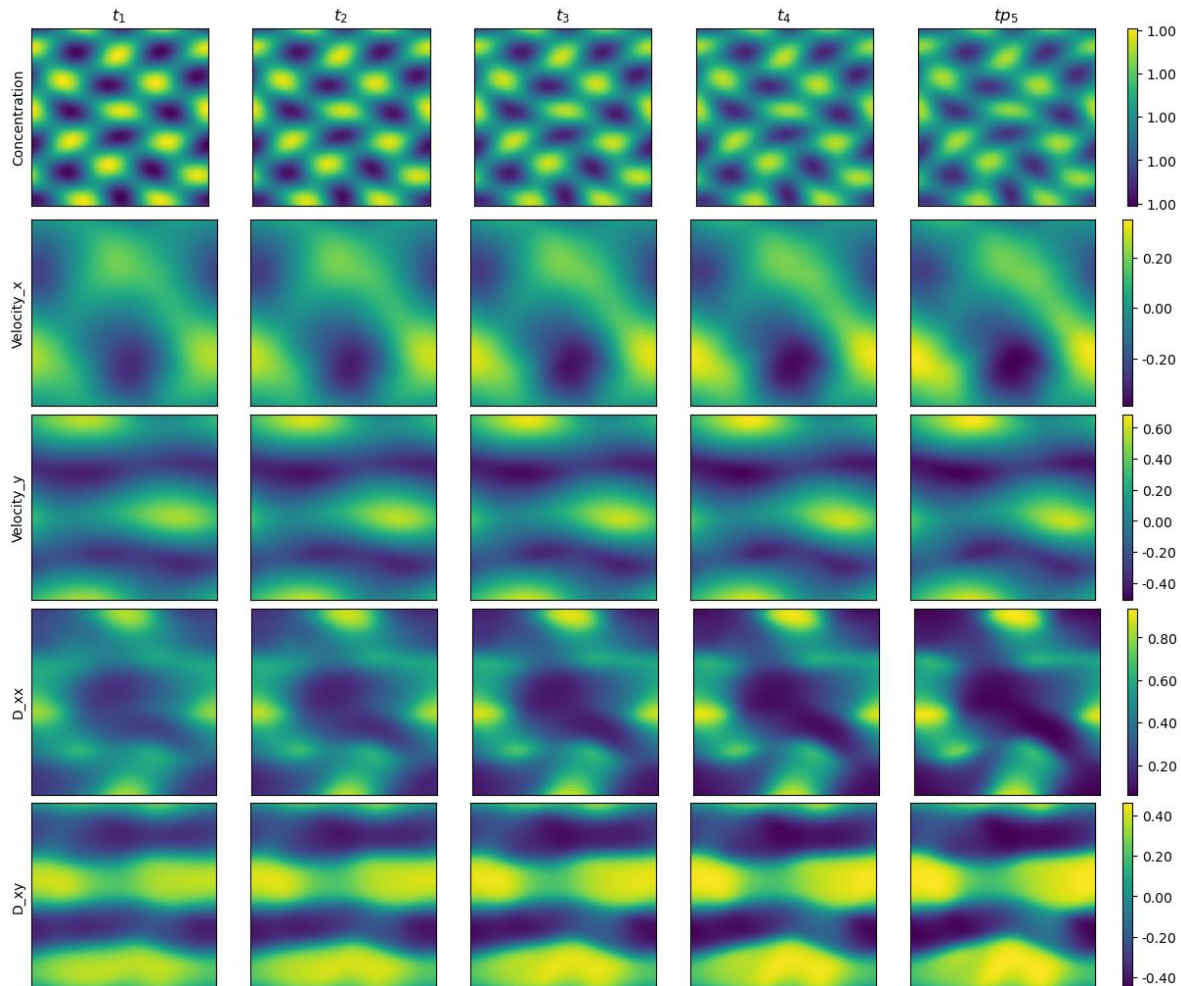
$$\boldsymbol{\Sigma} = \alpha \mathbf{D} + \beta \mathbf{S} : \mathbf{E} - 2\zeta \beta (\mathbf{D} \cdot \mathbf{D} - \mathbf{S} : \mathbf{D}).$$



Active Matter:



The forward problem:



U-NetConvNext:

Implementation of the U-Net model using ConvNext blocks.

```
=====  
Layer (type:depth-idx)                Param #  
=====  
CNextUNet                             --  
├─Conv2d: 1-1                          16,674  
├─ModuleList: 1-2                      --  
│   └─DownStage: 2-1                  --  
│       └─Sequential: 3-1              33,012  
│           └─Conv2d: 3-2              14,196  
│   └─DownStage: 2-2                  --  
│       └─Sequential: 3-3              122,472  
│           └─Conv2d: 3-4              56,616  
│   └─DownStage: 2-3                  --  
│       └─Sequential: 3-5              470,736  
│           └─Conv2d: 3-6              226,128  
├─Sequential: 1-3                      --  
│   └─CNextBlock: 2-4                 --  
│       └─Conv2d: 3-7                  16,800  
│           └─GroupNorm: 3-8           672  
│               └─Conv2d: 3-9          452,928  
│                   └─Conv2d: 3-10      451,920  
│                       └─GELU: 3-11    --  
├─ModuleList: 1-4                      --  
│   └─UpStage: 2-5                    --  
│       └─ConvTranspose2d: 3-12        225,960  
...  
=====  
Total params: 2,785,493  
Trainable params: 2,785,493  
Non-trainable params: 0  
=====
```

- CNextU-net

- Spatial filter size - 7
- Initial dimension - 42
- Blocks per stage - 2
- Up/Down blocks - 4
- Bottleneck blocks - 1

$$\text{VRMSE}(u,v) = (\langle |u-v|^2 \rangle / (\langle |u-\bar{u}|^2 \rangle + \epsilon))^{1/2}$$

VRMSE

Dataset	FNO	TFNO	Unet	CNextU-net
active_matter	0.3691	0.3598	0.2489	0.1034

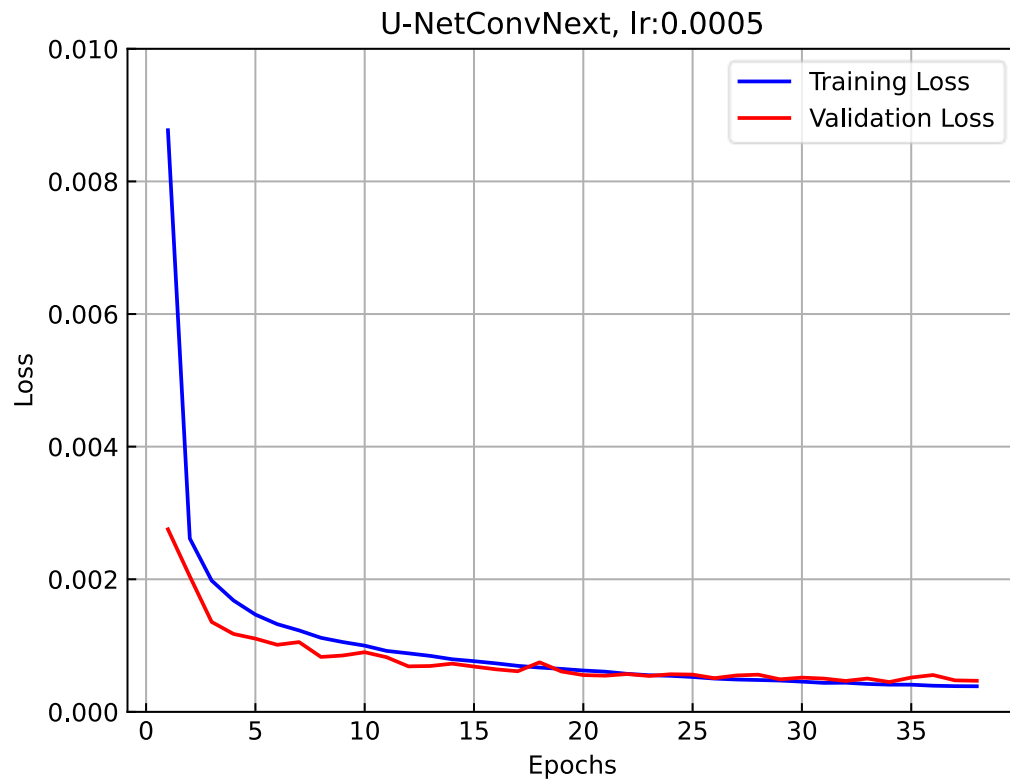
ConvLSTM:

- Hidden dimensions = [16, 32, 64, 128]
- Kernel Size = (3, 3)

```
=====
Layer (type:depth-idx)                Param #
=====
Forecaster                             --
├─ConvLSTM: 1-1                         --
│   └─ModuleList: 2-1                  --
│       └─ConvLSTMCell: 3-1             15,616
│           └─ConvLSTMCell: 3-2         55,424
│               └─ConvLSTMCell: 3-3     221,440
│                   └─ConvLSTMCell: 3-4  885,248
│                       └─ConvLSTMCell: 3-5 3,539,968
├─Conv2d: 1-2                           25,355
=====
Total params: 4,743,051
Trainable params: 4,743,051
Non-trainable params: 0
=====
```

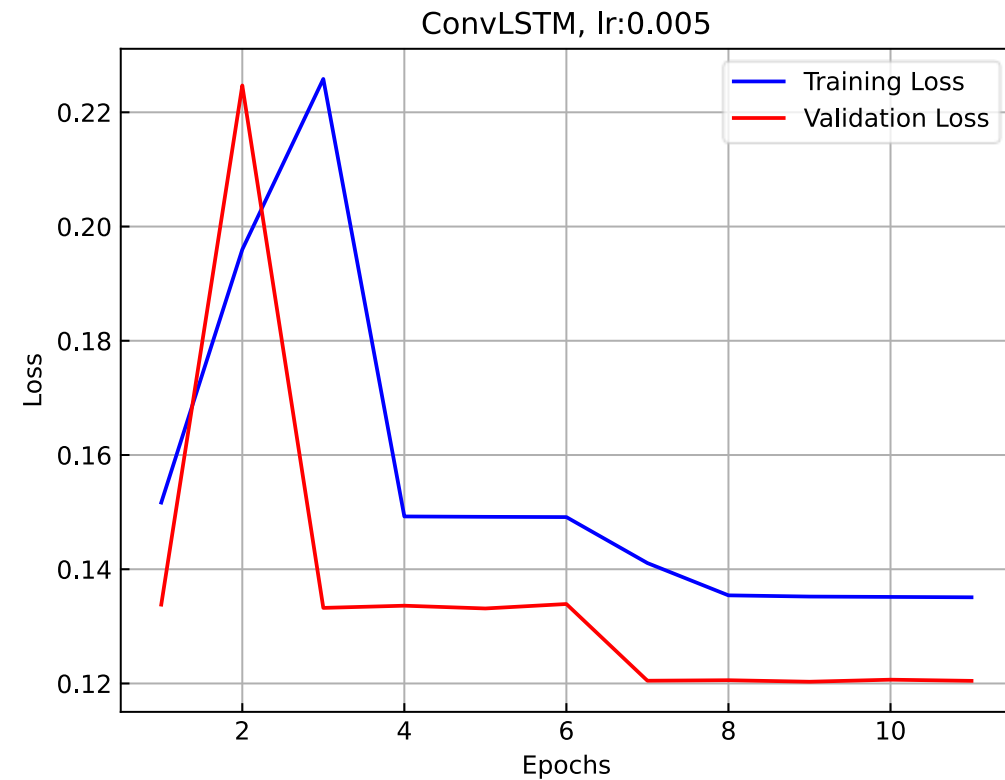
Results:

U-NetConvNext



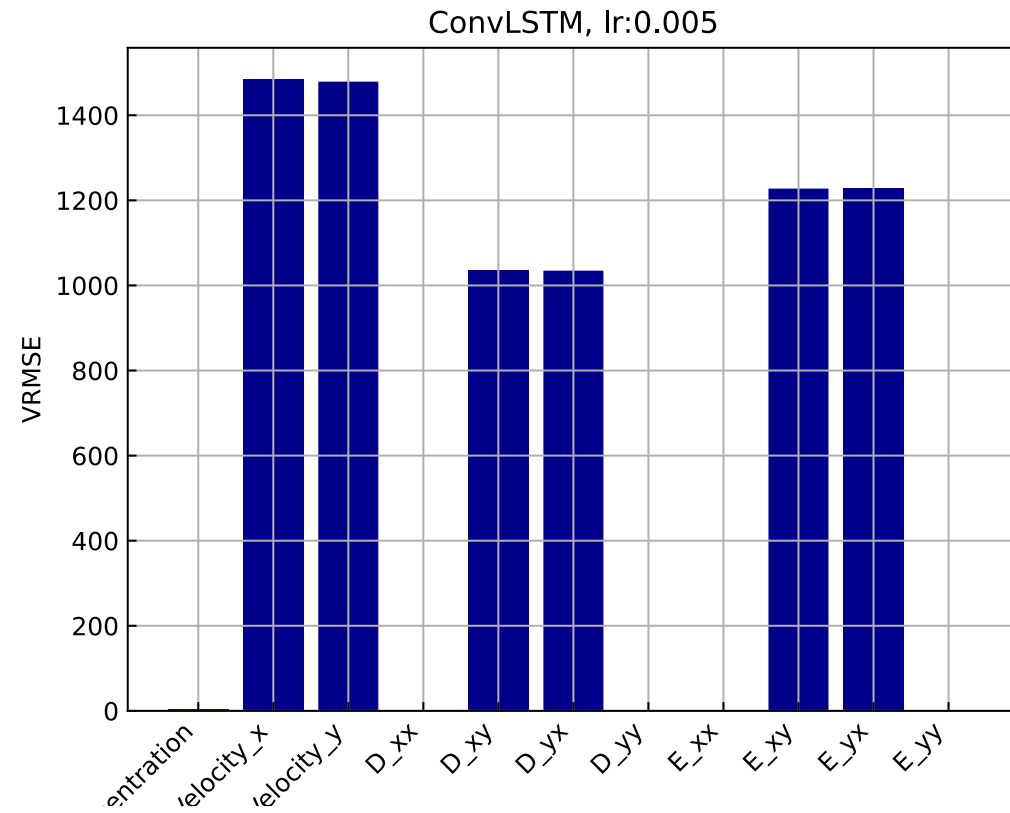
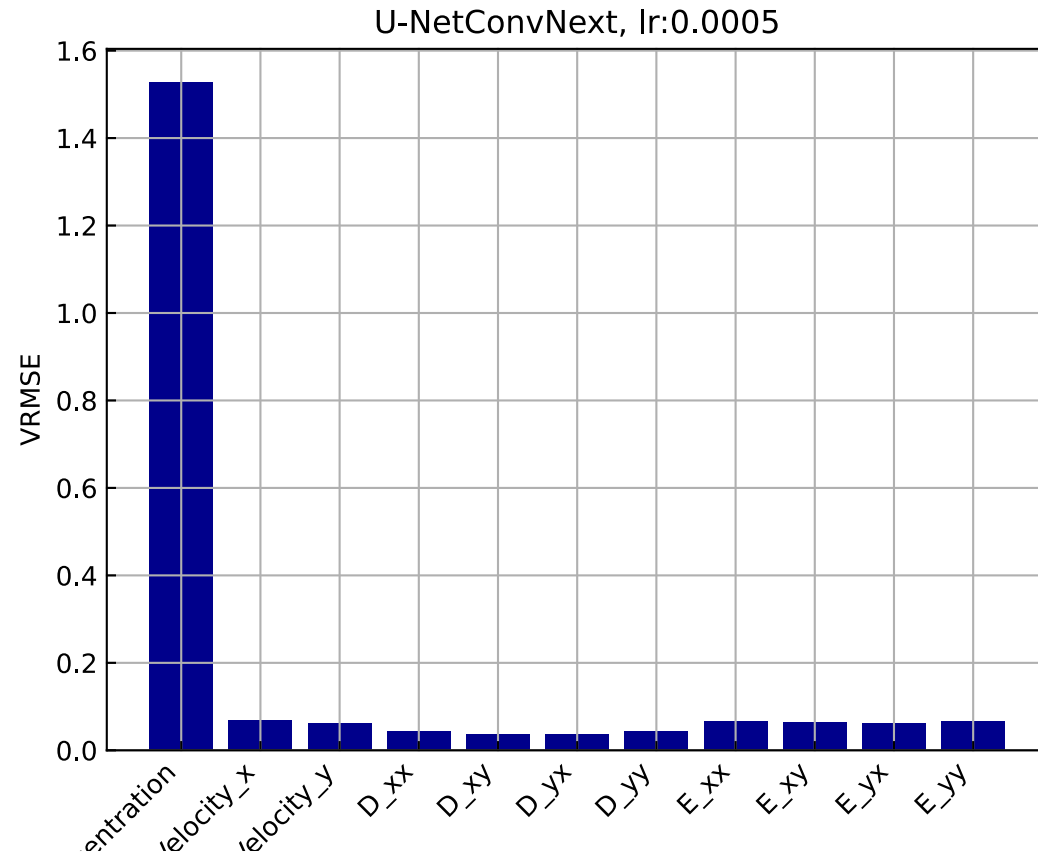
Test loss in VRMSE is: 0.037151

ConvLSTM

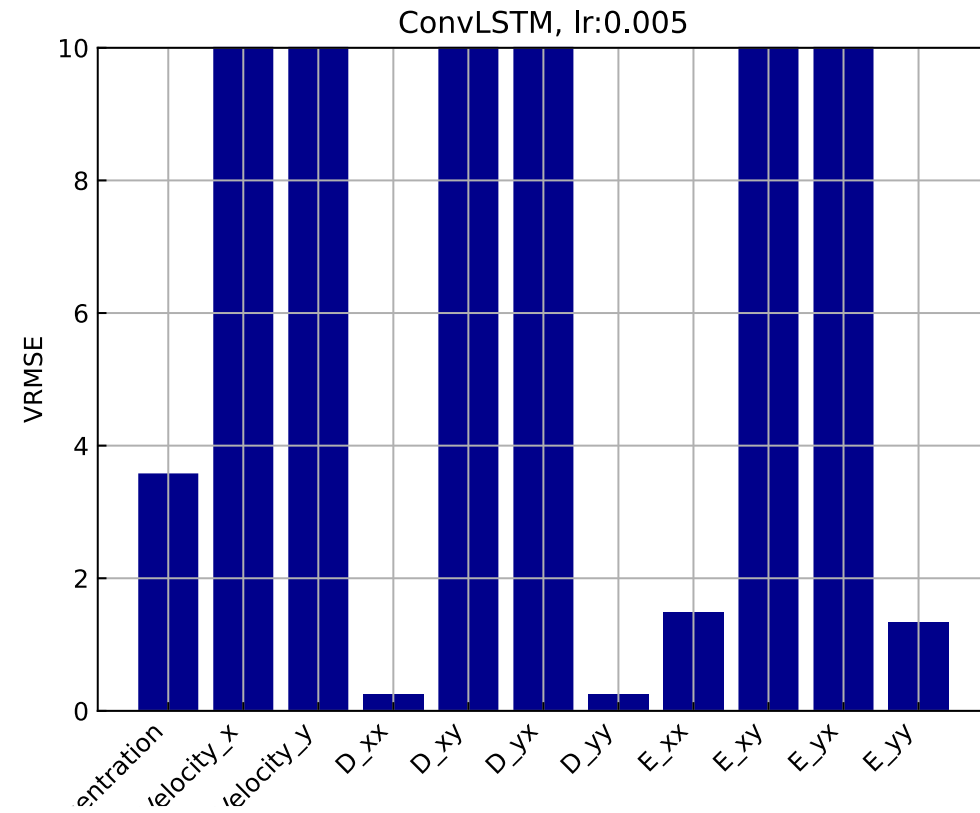
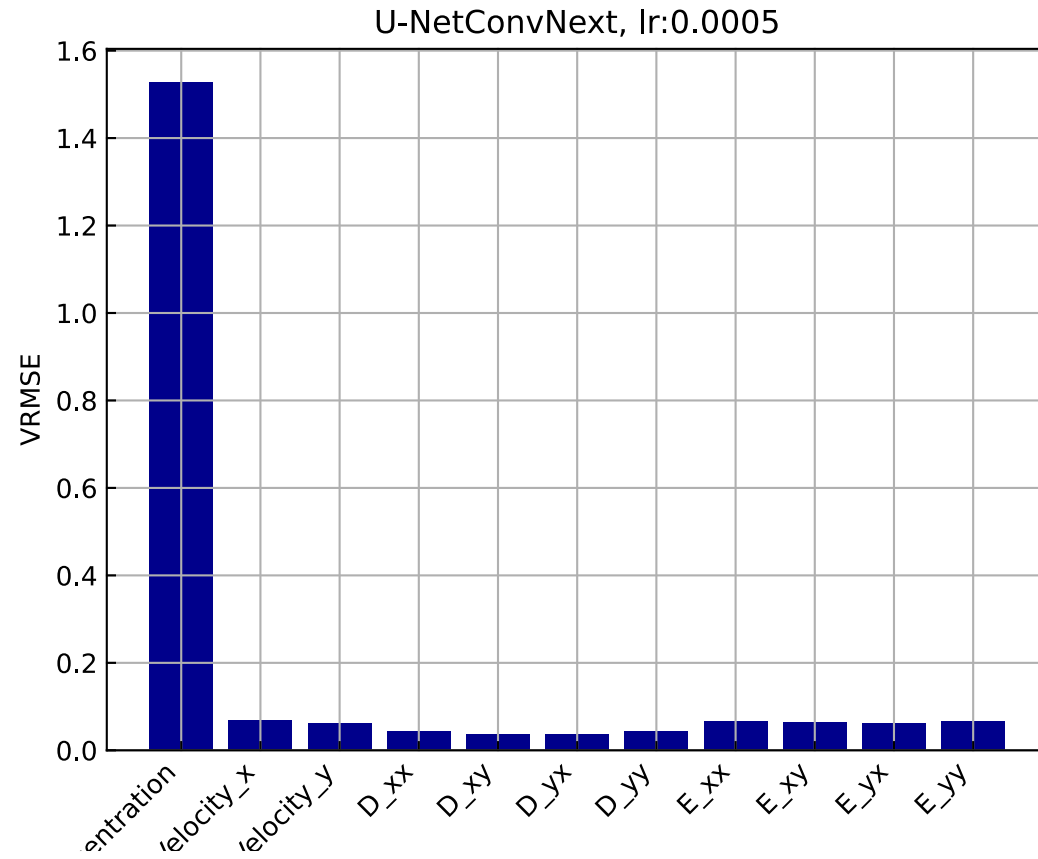


Test loss in VRMSE is: 1.071865

Results:

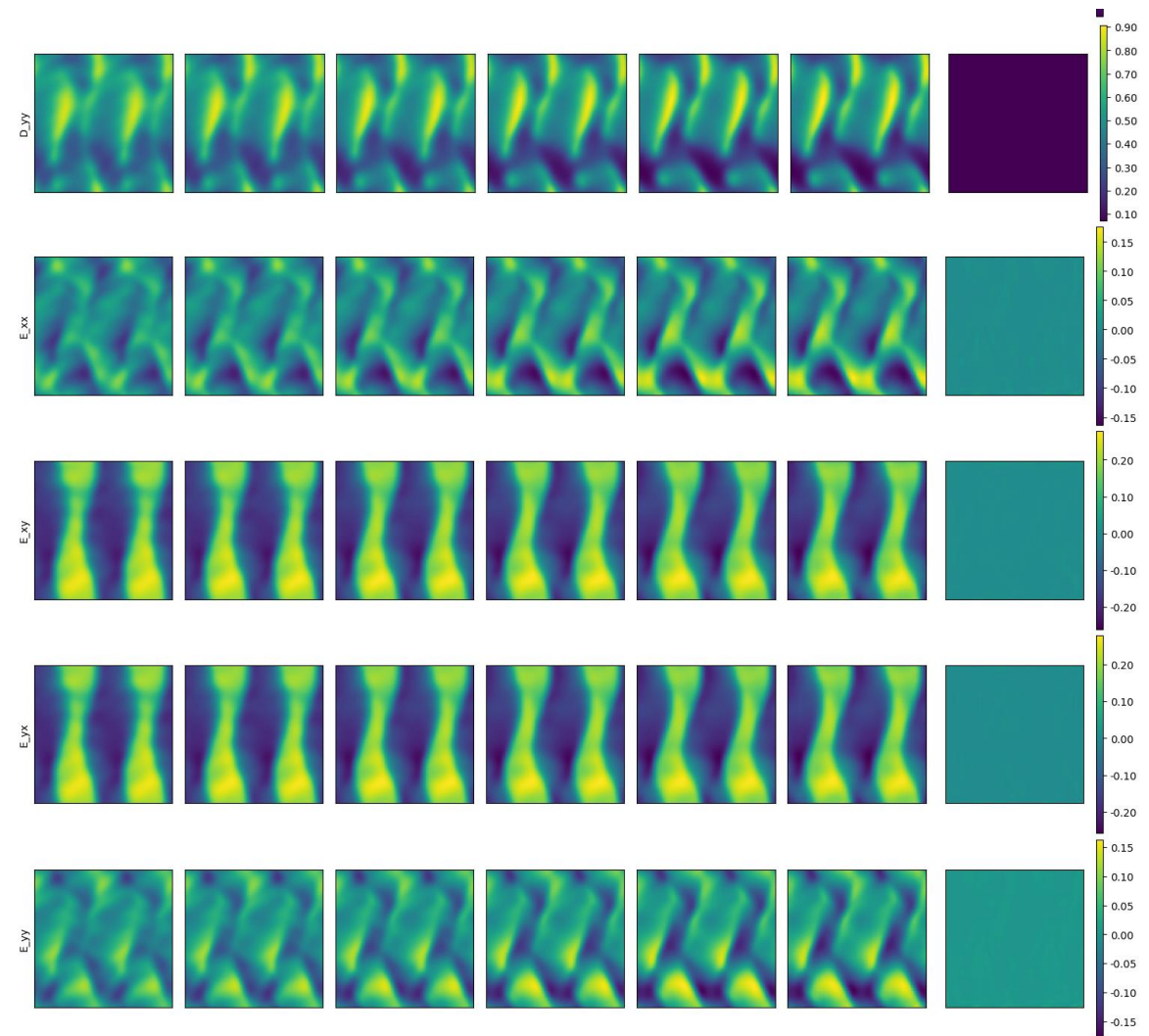
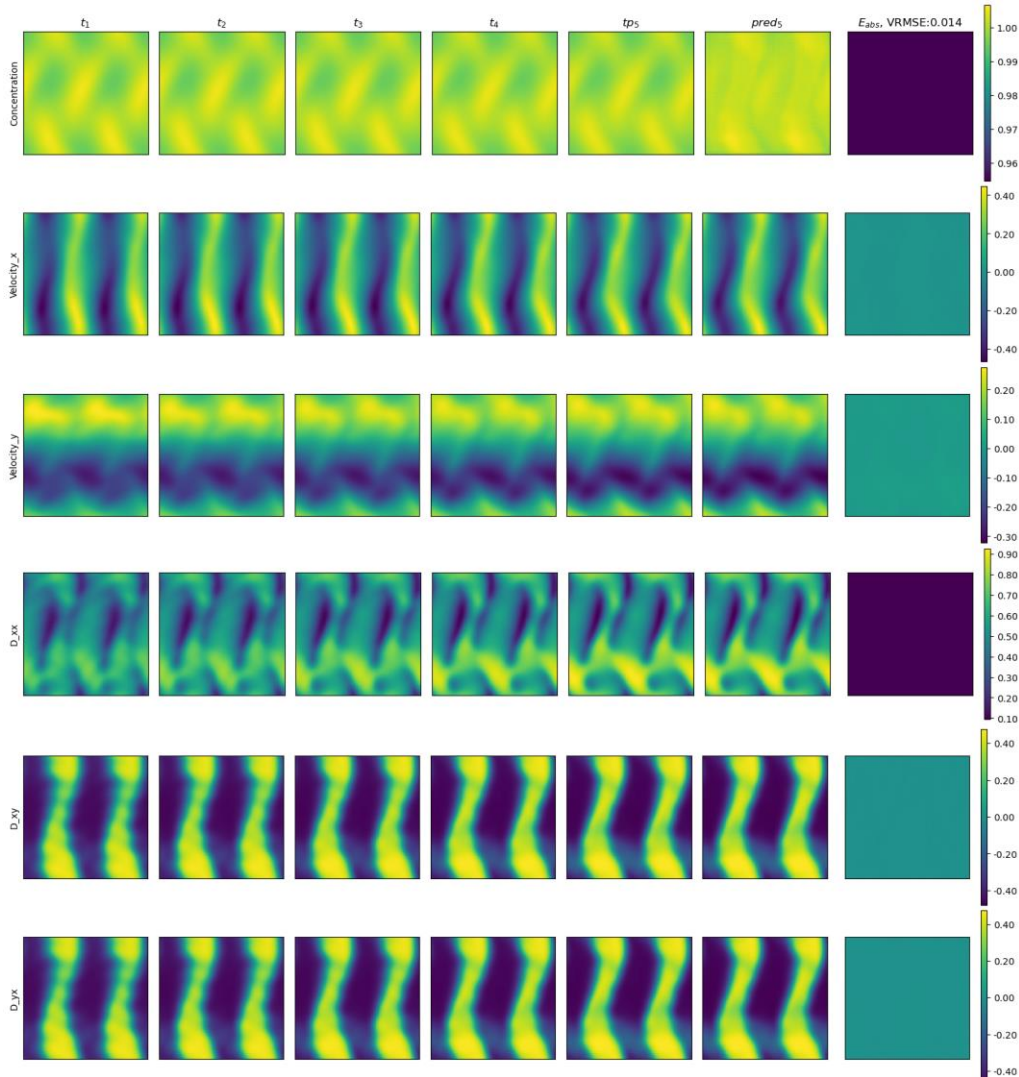


Results:



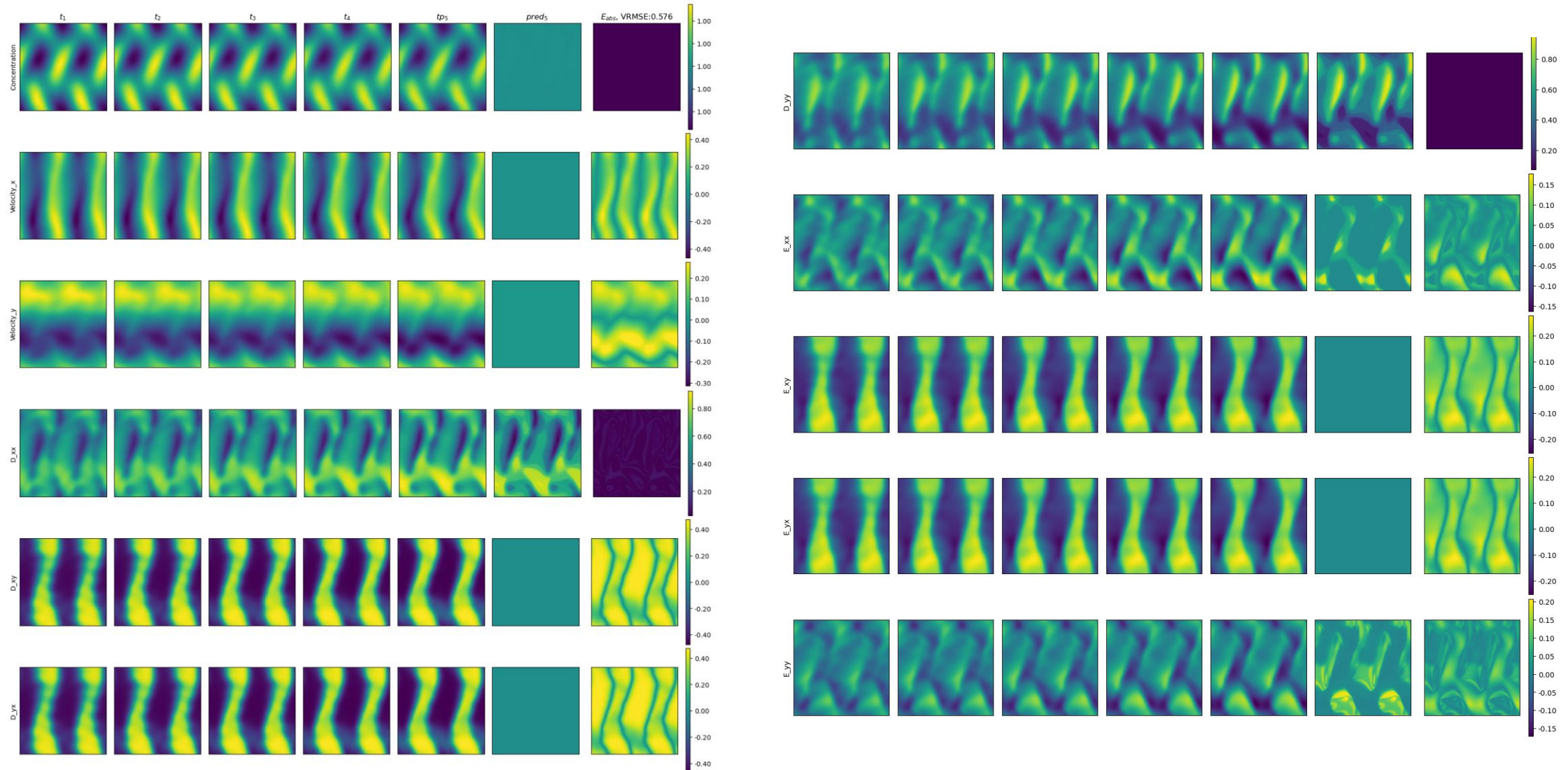
Prediction Visualization:

U-NetConvNext - VRMSE = 0.014674



Prediction Visualization:

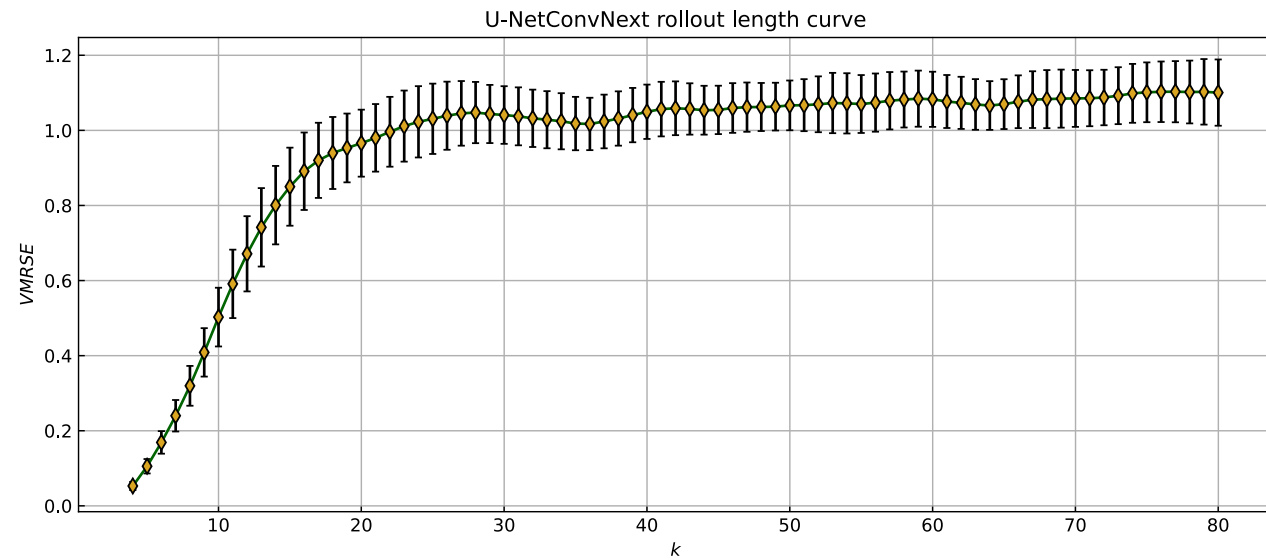
ConvLSTM - VRMSE = 0.576200



Rollout lengths:

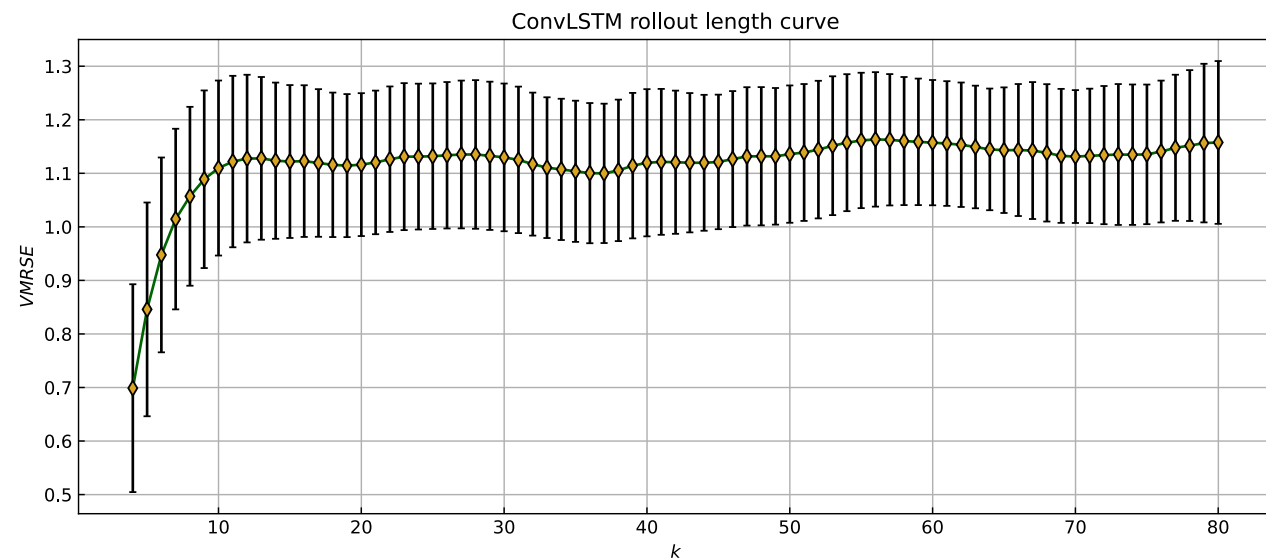
U-NetConvNext

Test loss in VRMSE is: 0.037151



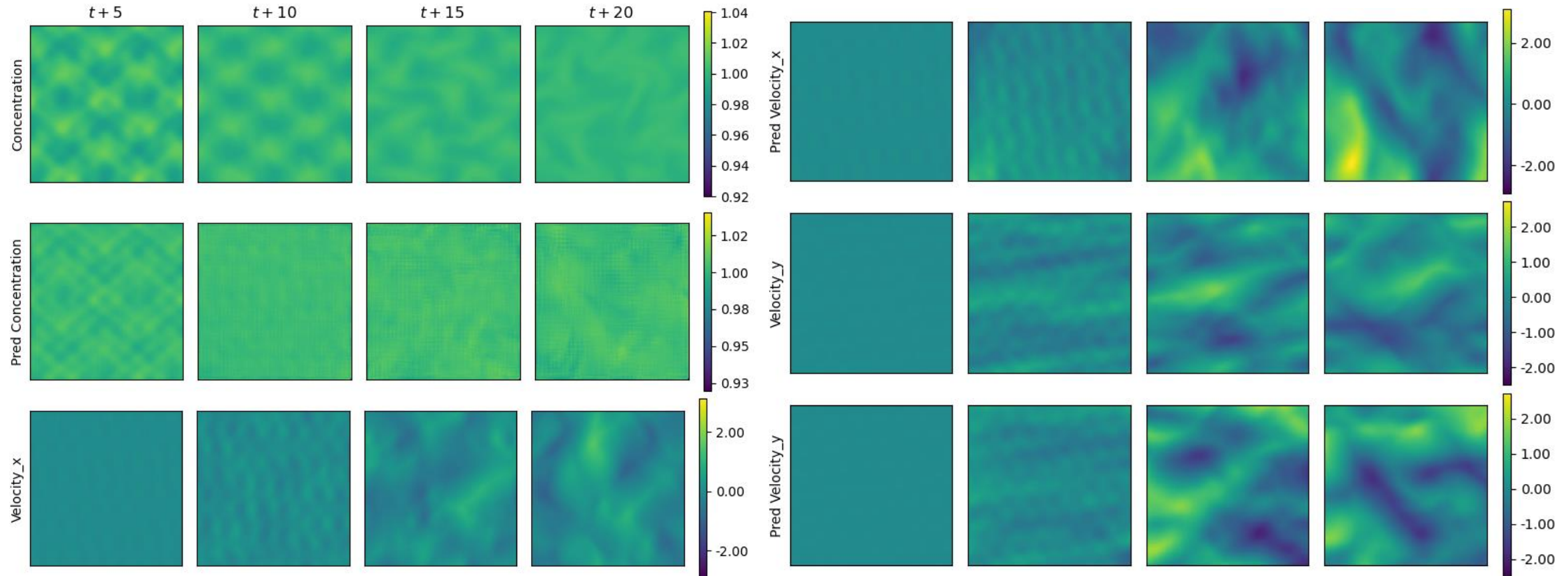
ConvLSTM

Test loss in VRMSE is: 1.071865



Multi-step Rollouts:

U-NetConvNext



Multi-step Rollouts:

ConvLSTM

