DEEP LEARNING FOR SHORT-TERM TEMPERATURE FORECASTS WITH VIDEO PREDICTION METHODS

BING GONG, SCARLET STADTLER, MICHAEL LANGGUTH, AMIRPASHA MOZAFFARI, JAN VOGELSANG, MARTIN G. SCHULTZ

I B.GONG@FZ-JUELICH.DE I 2020-05-06

EGU 2020: session ITS4.3/AS5.2 Machine learning for Earth System modelling

DeepRain Intelli AQ JÜLICH Forschungszentrum

ABSTRACT

2m temperature prediction

ERA5

Extraction

YEARS
2016, 2017

Pre-
processing

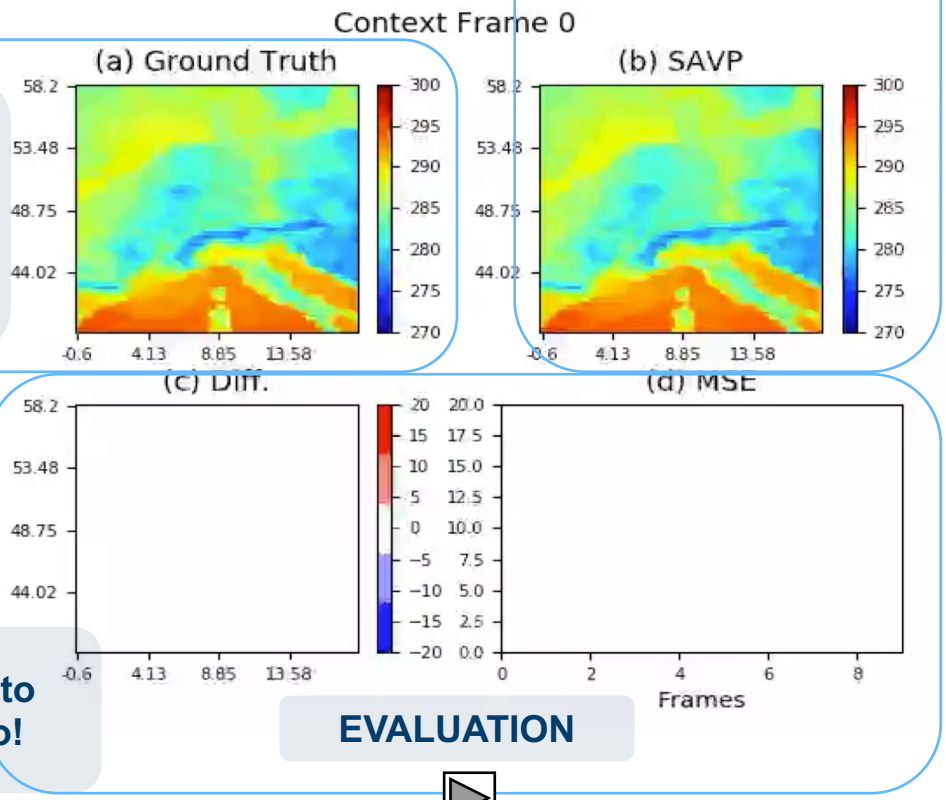
2m T
GEOPOT
SLP
Over Europe

Input

VIDEO PREDICTION
ALGORITHM

Predict

2 m T
FORECAST



Click here to
start video!

EVALUATION

Click to navigate to the corresponding section

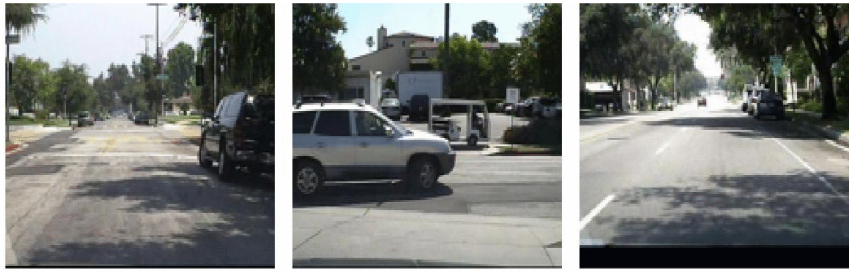
OUTLINE

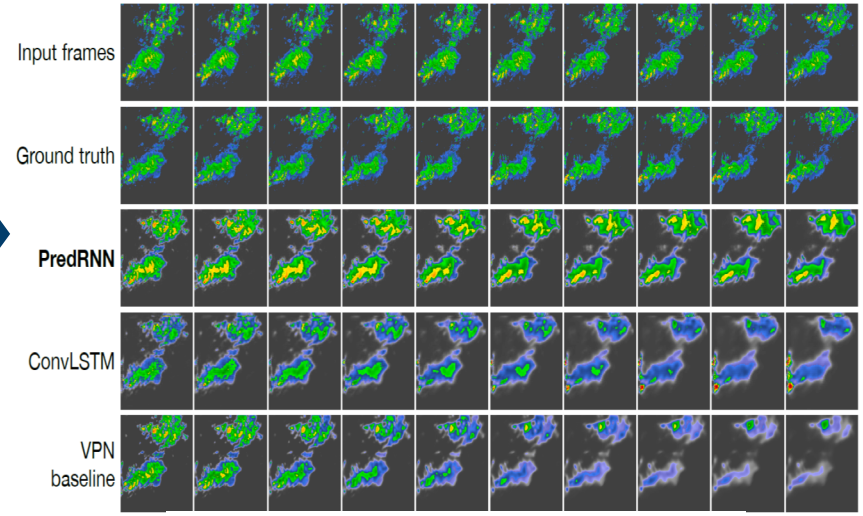
- **Motivation**
- Deep learning architectures
- Parallel deep learning workflow
- Experiment settings
- Results
- Conclusions and outlook

Hint: Click hyperlink to the corresponding section

MOTIVATION

- Video prediction → **New data-driven approach for weather forecasting**



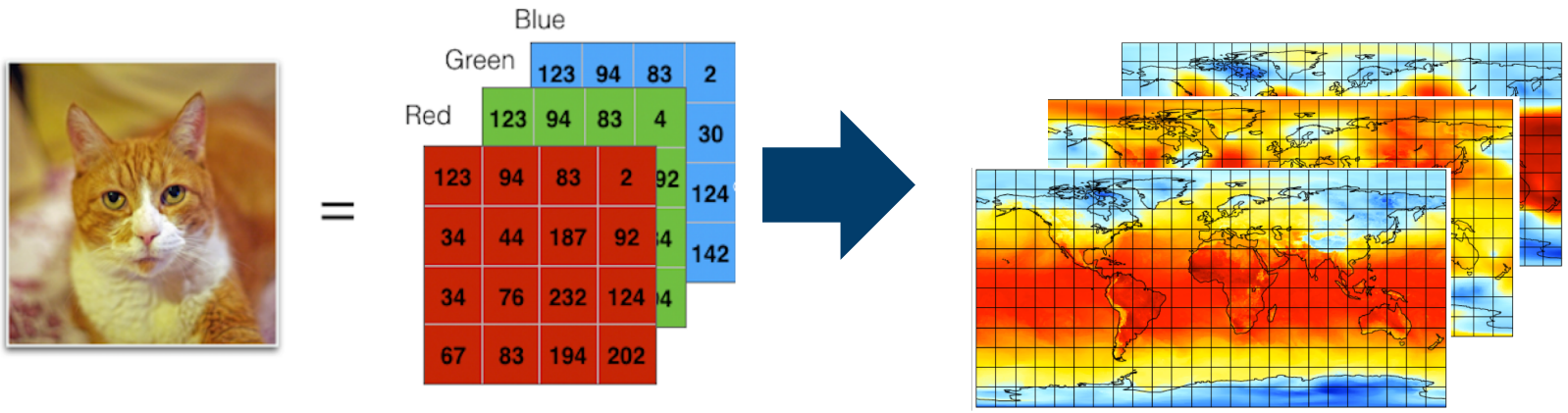


Wang, Y et al. 2017

- In computer vision, video prediction is an essential study area. It has been successfully applied in many applications such as human action prediction (Lee, Alex X. et al. 2018), self-driving cars (Lotter, William et al. 2016), etc.
- There are certain similarities between the deep learning tasks of video prediction and weather forecasting: Explore spatio-temporal patterns from previously observed data to generate the future frames.

MOTIVATION

- Video prediction → **New data-driven approach for weather forecasting**



- The channels of images can be represented by different relevant variables for weather prediction.

MOTIVATION

- Video prediction → **New data-driven approach for weather forecasting**

- Big data in weather forecasting and advanced Deep Learning architecture → **High-Performance Computing (HPC) & Parallelisation**

  - Whenever a deluge of data has to be processed (like in video prediction and weather forecasting applications or for cutting-edge deep learning architectures with enormous parameters), parallel computing is necessary for accelerating the life-cycle of prediction.

- Reproducibility → **End-to-end workflow**

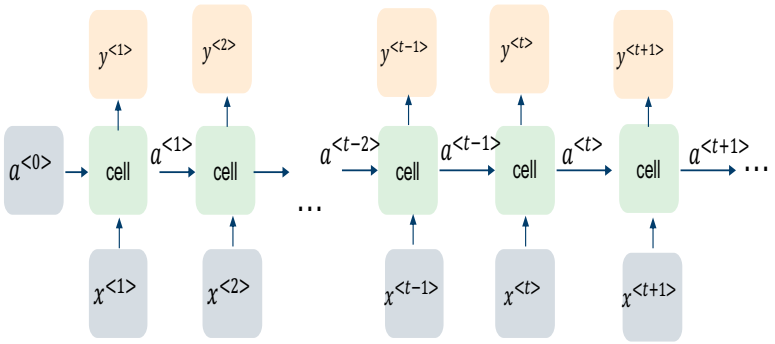
  - For deep learning applications, the data and the selected architectures can change constantly for various experiment settings. Therefore, an end-to-end workflow ensures the reproducibility of these experiments.

OUTLINE

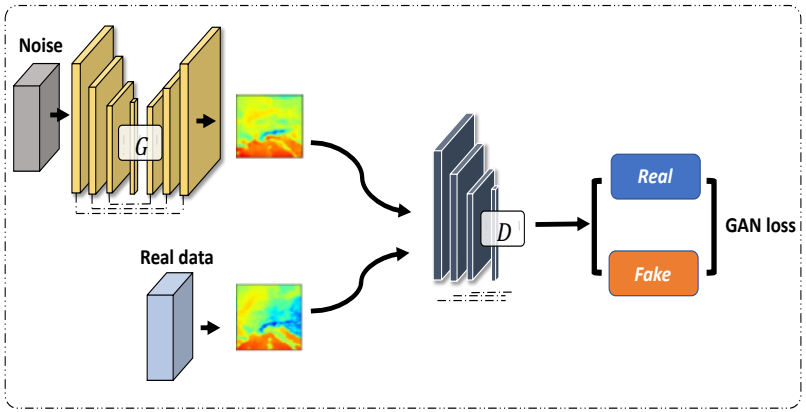
- Motivation
- **Deep learning architectures**
- Parallel deep learning workflow
- Experiment settings
- Results
- Conclusions and outlook

DEEP LEARNING ARCHITECTURES

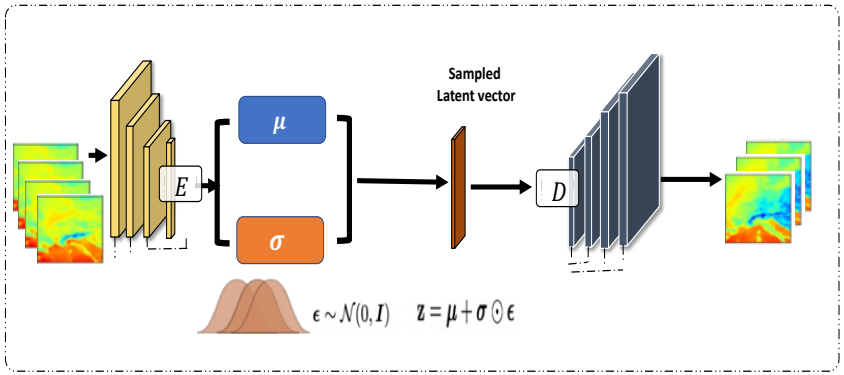
State-of-the-art architectures for video prediction



RNN architecture (click ▶ for more details)



GAN architecture (click ▶ for more details)



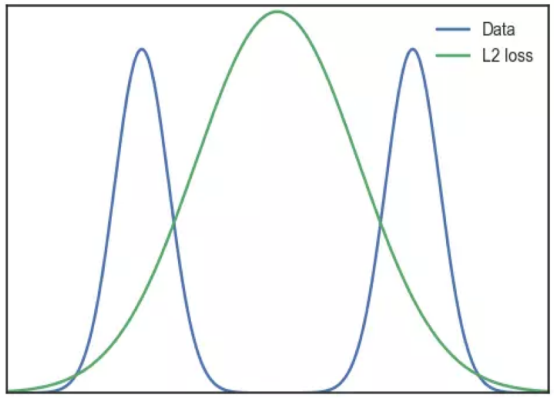
VAE architecture (click ▶ for more details)

- The state-of-the-art methodologies for video prediction can be roughly categorized into three groups, **Recurrent Neural Network** (RNN), **Variational Autoencoder** (VAE), and **Generative Adversarial Network** (GAN) based model.

DEEP LEARNING ARCHITECTURES

Downsides of state-of-the-art architectures for video prediction

Problem 1: Unrealistic images



$$L2 = \sum_{i=1}^{n} (y - \hat{y})^2 \qquad L1 = \sum_{i=1}^{n} |y - \hat{y}|$$

Adversarial loss

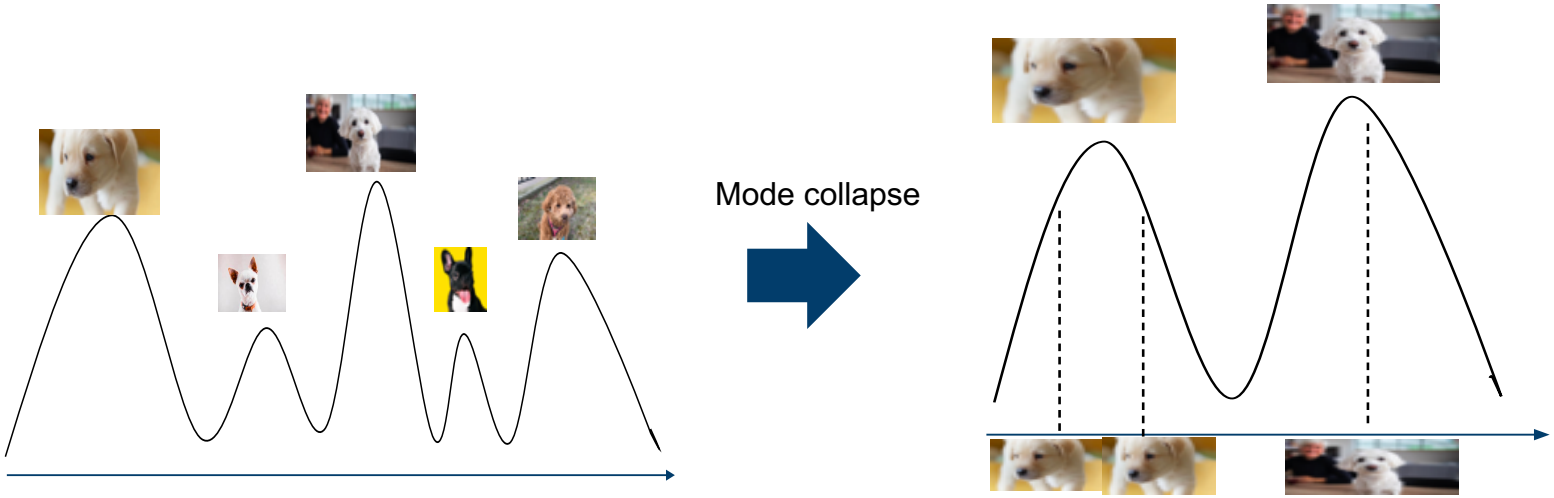
$$\mathcal{L}_{GAN}(G, D) = \mathrm{E}_{x_{1:T}}[logD(x_{1:T})] + \mathrm{E}_{x_{1:T, z\sim p(z_t)|_{t=0}^{T-1}}}\left[\log(1 - D\big(G(x_0, z_{0:T-1})\big))\right]$$

- For RNN and VAE, if each pixel follows multi-modal distribution, then using L2 Norm or L1 Norm will average the loss function. It means to produce the mean of image of all possible futures, as the global optimum. This will produce unrealistic prediction images.

DEEP LEARNING ARCHITECTURES

Downsides of state-of-the-art architectures for video prediction

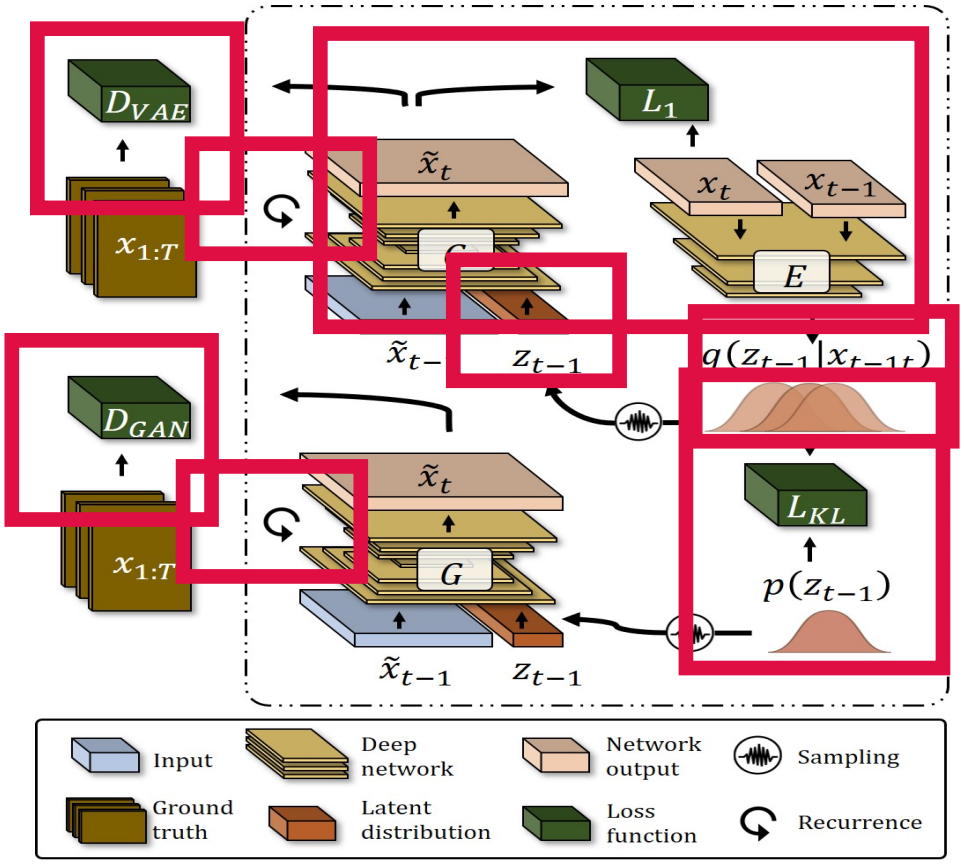
Problem 2: Mode collapse



- GAN based on adversarial learning can produce more realistic images than RNN and VAE with L2/L1 loss. On the downside, it potentially only produces a limited diversity of output (e.g. limited number of dog breeds), which is commonly called ***mode collapse***.

DEEP LEARNING ARCHITECTURES

Selected solution for weather forecasting



Lee et al. 2018

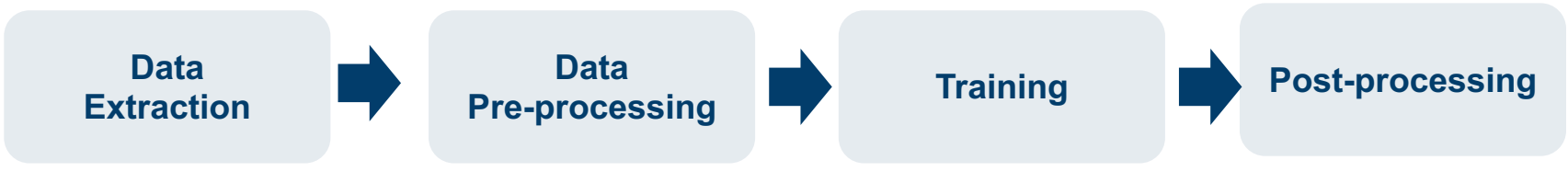
- Stochastic Adversarial Video Prediction (**SAVP**) takes advantages of VAE and GAN, and was selected in this study.
- SAVP compares with GAN and VAE architectures.
- VAE part: SAVP is conditioned on the latent variable $z(t-1)$, which is sampled from latent space q. q is approximated by the Kullback Leibler divergence from prior distribution p, which follows Gaussian distribution. This could be an effective way to generate diverse output.
- The images/scenes are then reconstructed by Generator G and the model is optimized by minimizing L1 norm function.
- GAN part: Discriminators (D_{GAN} and D_{VAE}) are used for distinguishing generated video from real ones. This could help to generate realistic videos/images.
- convLSTM as cell for generator and discriminator to preserve the spatio-temporal patterns.
- batchNorm and ReLu activation function are used for convolutional layers.

OUTLINE

- Motivation
- Deep learning architectures
- **Parallel deep learning workflow**
- Experiment settings
- Results
- Conclusions and outlook

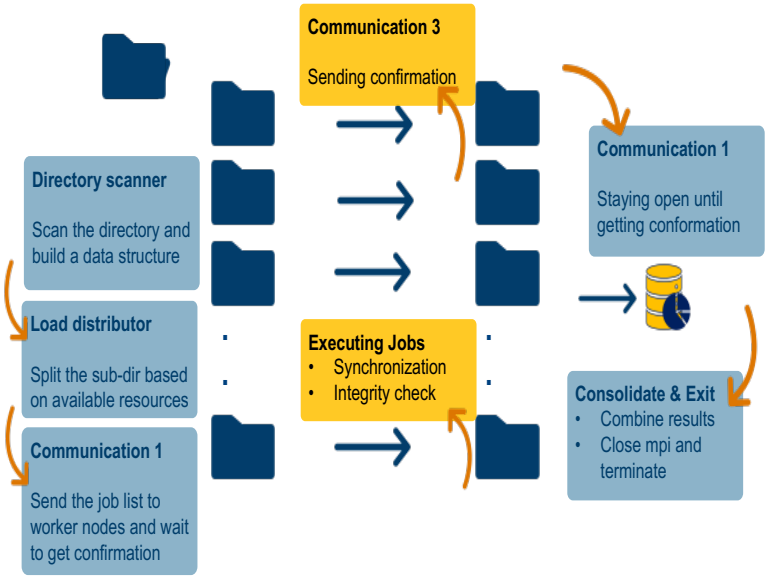
PARALLEL DEEP LEARNING WORKFLOW

Parallelization with Pystager and Horovod



Pystager:

http://bit.ly/PyStager1



HOROVOD

- We have Pystager for data extraction and pre-processing, and Horovod (Sergeev, Alexander et al., 2018) for the last two steps.
- Check scalability performance for training

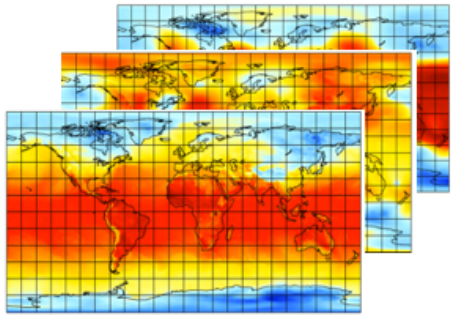
OUTLINE

- Motivation
- Deep learning architectures
- Parallel deep learning workflow
- **Experiment settings**
- Results
- Conclusions and outlook

EXPERIMENT SETTINGS

Temperature forecasting

ECMWF



ERA5 reanalysis data

- **Size**: ~ 500 TB,
- **Grid points/Pixel**: 601 * 1200
- **Time resolution:** Hourly

- We adopt a case study as proof-of-concept for our proposed parallel deep learning workflow. This study aims to **forecast 2m above sea level temperature.** We take the era5 reanalysis data from ECMWF as data source.

Data Extraction

Data Pre-processing

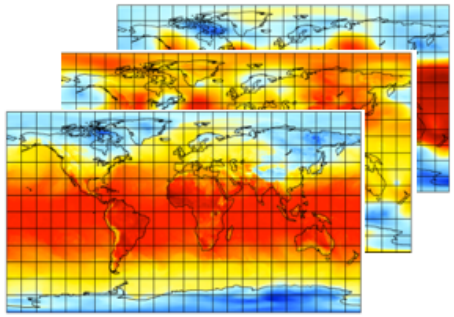
Training

Post-processing

EXPERIMENT SETTINGS

Temperature forecasting

ECMWF



ERA5 reanalysis data

- **Size**: ~ 500 TB,
- **Grid points/Pixel**: 601 * 1200
- **Time resolution:** Hourly

- **Size:** ~ 2.2 TB
- **Time:** 2016/2017
- **Region:** Europe
- **Var:**
  - 2m Temperature
  - 500hPa Geopotential
  - Sea level pressure

- We take 2016 and 2017 as our operational datasets along with different combinations of selected variables, based on which we conduct a series of experiments.

Data Extraction

Data Pre-processing

Training

Post-processing

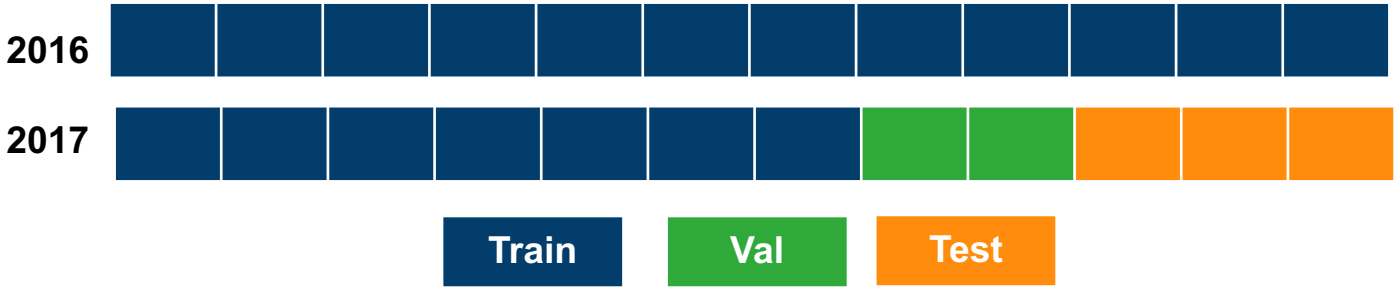
EXPERIMENT SETTINGS

Temperature forecasting

Forecasting the next 10 hours 2m temperature

- **Size**: ~ 3300 MB
- **Samples:** 64 * 64 * 3, 10 frames /sample as an input

Data splitting without shuffling



- Each block corresponds to a month.
- We partition the data into Training, validation and testing dataset, without shuffling.

Data Extraction

Data Pre-processing

Training

Post-processing

EXPERIMENT SETTINGS

Experiment settings

No.	Years	Arc.	Variables	Mode
1	2017	SAVP	3 * Temperature	End-to-end
2	2017	SAVP	Temperature, mean sea-level pressure, 500hPa geopotential	End-to-end
3	2016-2017	SAVP	3 * Temperature	End-to-end
4	2017	SAVP	3 * Temperature	KTH_pretrained
5	2017	GAN	3 * Temperature	End-to-end
6	2017	GAN	Temperature, mean sea-level pressure, 500hPa geopotential	End-to-end
7	2016-2017	GAN	3 * Temperature	End-to-end
8	2017	VAE	3 * Temperature	KTH_pretrained
9	2017	VAE	3 * Temperature	End-to-end
10	2017	VAE	3 * Temperature	KTH_pretrained

Note: More details about variables, mode selection 

- We conducted several experiments by varying dataset selected year, architectures, variables, and also training modes - pretrained or end-to-end learning.

Data Extraction

Data Pre-processing

Training

Post-processing

OUTLINE

- Motivation
- Deep learning architectures
- Parallel deep learning workflow
- Experiment settings
- **Results**
- Conclusions and outlook

RESULTS

Transfer learning VS end-to-end learning

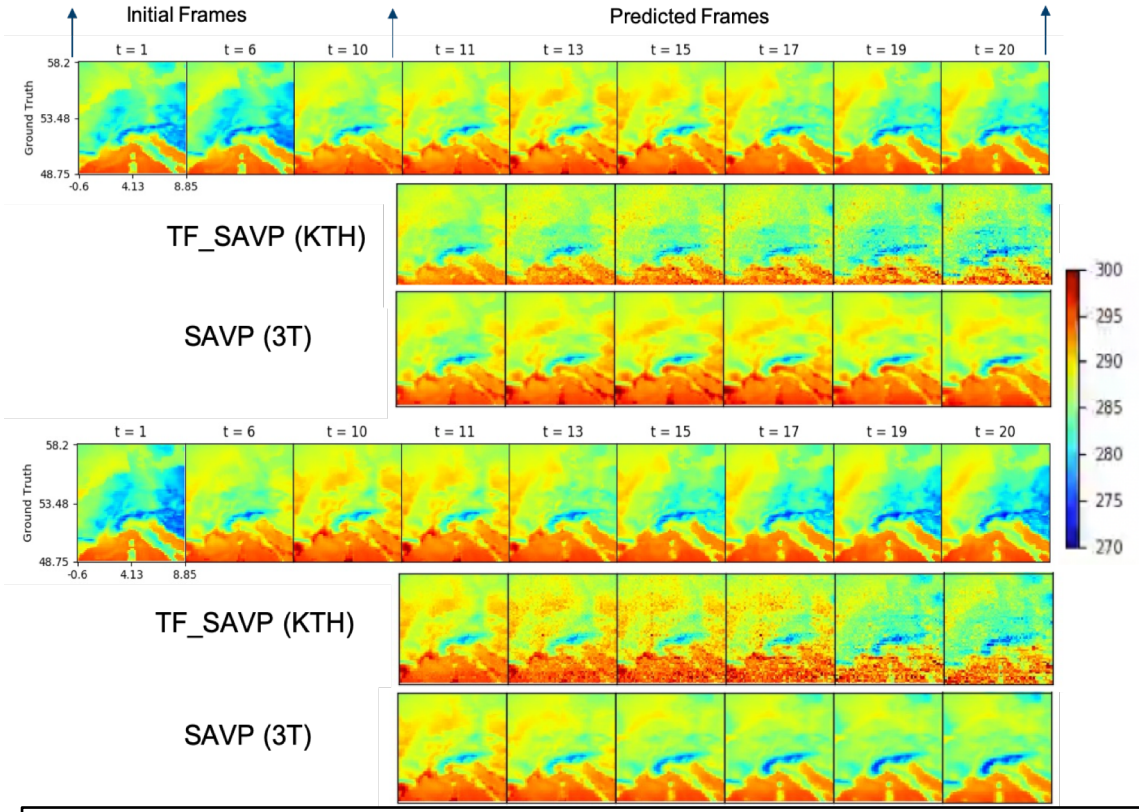


Figure. Visualization of two examples of ground truth (top row) and predicted frames from experiments 4 (second row) and 1 (third row);values temperature are in Kelvin

- The results show that the pre-trained model generates noisy frames where the temperature is homogeneous over the Mediterranean sea. By contrast, the end-to-end trained model (experiment 1) tends to underestimate cooling over Central Europe.

Data Extraction

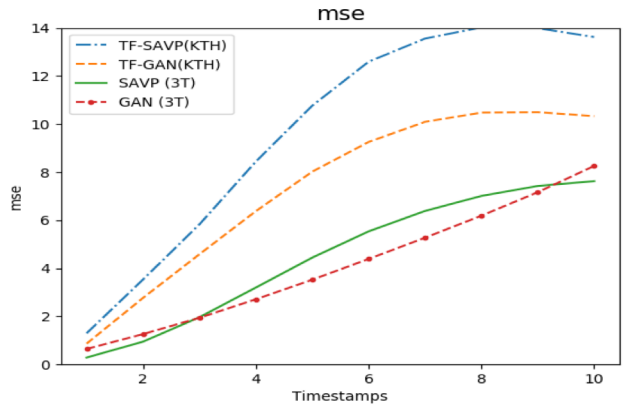
Data Pre-processing

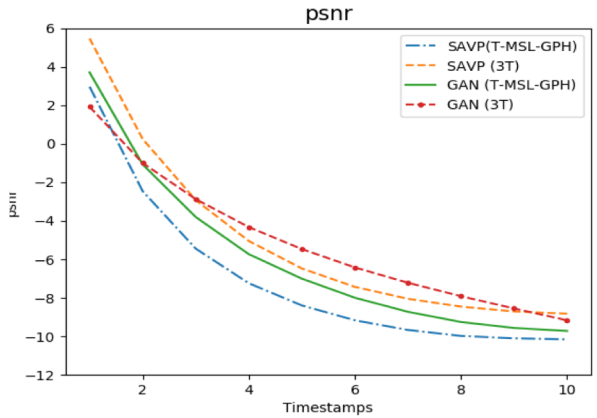
Training

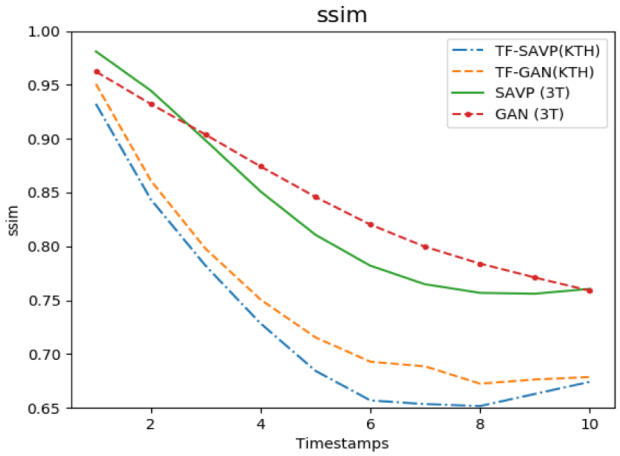
Post-processing

RESULTS

Transfer learning VS end-to-end learning







- Compares the forecast quality of the pre-trained experiments with the scores of the benchmark experiments with the three model architectures, showing In all cases, pre-training with the KTH dataset reduces the forecast skill of the models.

Evaluation metrics (SNR, SSIM, MSE and skill scores)

Data Extraction

Data Pre-processing

Training

Post-processing

RESULTS

GAN, VAE, and SAVP

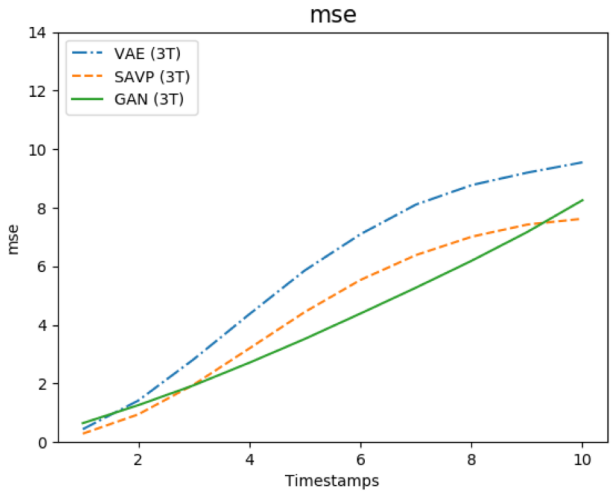


Figure. Temperature frame prediction performances of the end-to-end trained models SAVP, GAN, and VAE from experiments 1, 5, and 9

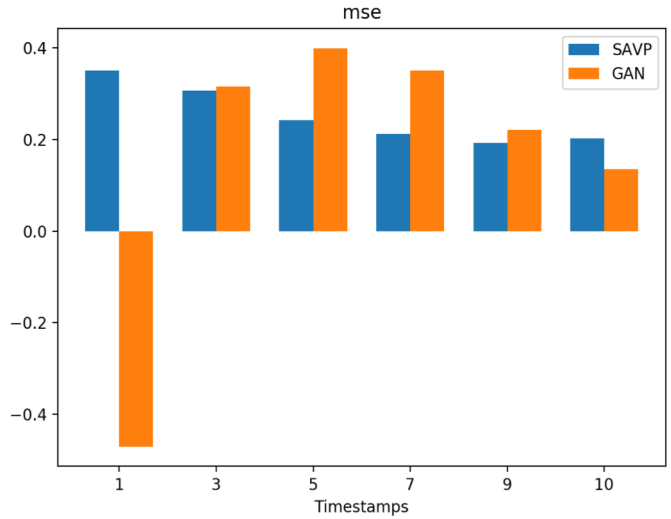


Figure. Skill scores of prediction performances from experiments 1, and 5 using experiment 9 as reference

- The VAE (experiment 9) is used as reference model.
- The results demonstrate that for all models, the accuracy degrades gradually with increasing lead time.
- In most cases, SAVP and GAN show positive skill scores, thus indicating that these models outperform VAE. Between forecast steps 3 and 7, the GAN network exhibits the best performance.

Data Extraction

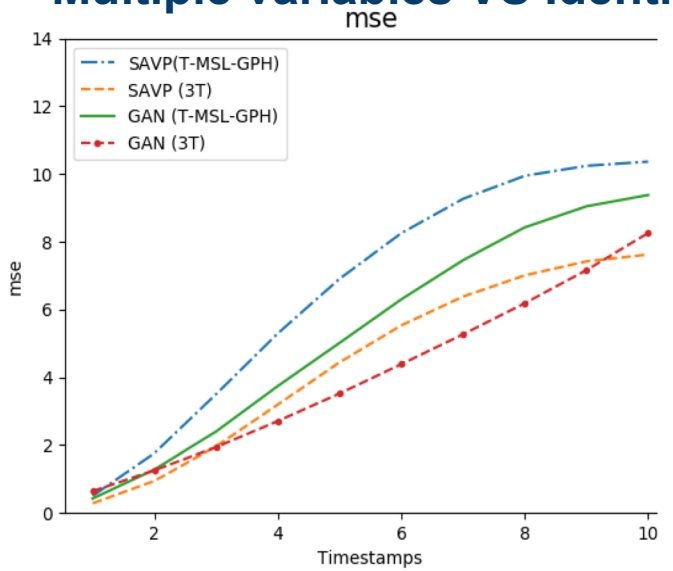
Data Pre-processing

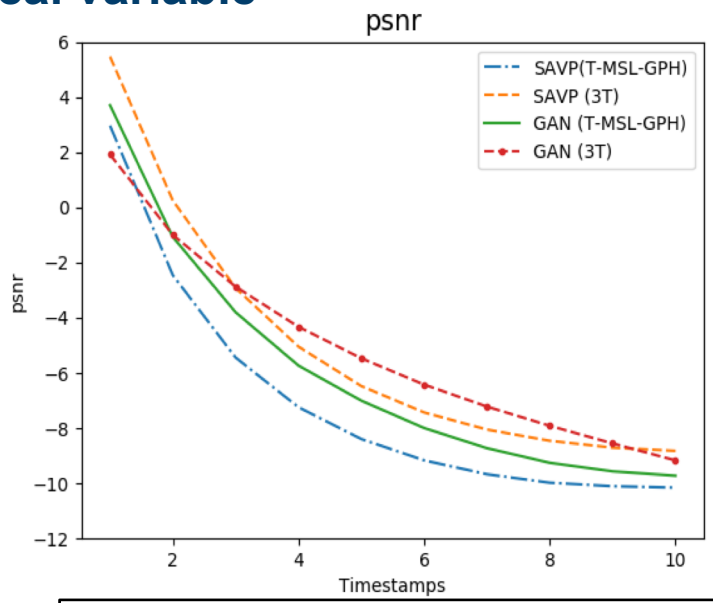
Training

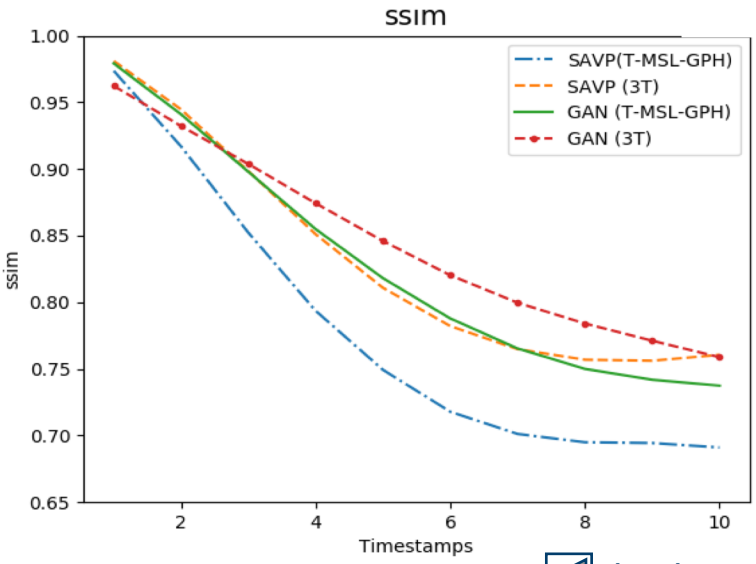
Post-processing

RESULTS

Multiple variables VS Identical variable







- Comparison of forecast scores from experiments with three duplicated temperature variables (experiments 1 and 5) versus those with three input variables (experiments 2 and 6).
- The two additional variables may provide valuable knowledge on general weather conditions.
- The networks do not seem to capture the physical relation between the near surface temperature and the large scale flow pattern on these time scales.

Data Extraction

Data Pre-processing

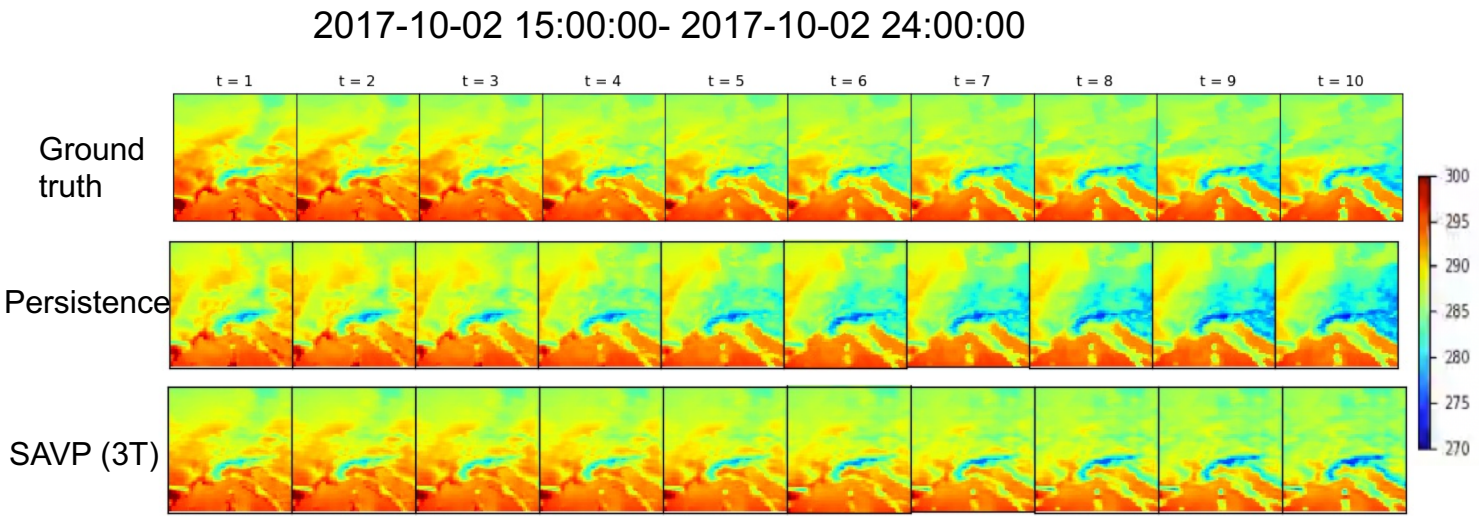
Training

Post-processing

JÜLICH Forschungszentrum

RESULTS

Persistence analysis



- We will answer: How much added value is the forecasting model to simply saying that tomorrow will be like today (Marion P. Mittermaier, 2007)
- Use the same hour of previous day as forecasting and compare with the SAVP (3T) forecasting.
- From the two samples (this slide and next one), the SAVP prediction outperforms persistent temperature.
- More cases:

Data Extraction

Data Pre-processing

Training

Post-processing

OUTLINE

- Motivation
- Deep learning architectures
- Parallel deep learning workflow
- Experiment settings
- Results
- **Conclusions and outlook**

CONCLUSIONS AND OUTLOOK

Application of the state-of-the-art video prediction methods for 2m temperature forecasting over Europe

Deep Learning

- ✓ VAE, GAN, and SAVP compared
- ✓ Transfer learning VS end-to-end training
- ✓ Multi-variables testing
- ✓ Persistence analysis

- X Lack of physical constraints
- X Multiscale in spatial/temporal space/ to catch global and long-term features
- X Latent space learning to get rid of the Gaussian distribution assumption

Parallel Training

- ✓ Horovod library can be efficiently employed for parallel training.

X Some issues remain with respect to convergence of results

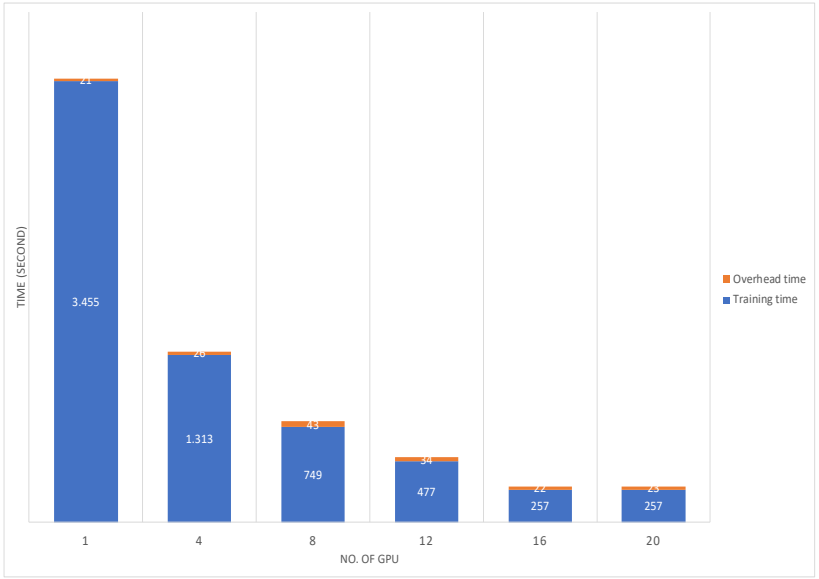
ACKNOWLEDGEMENT

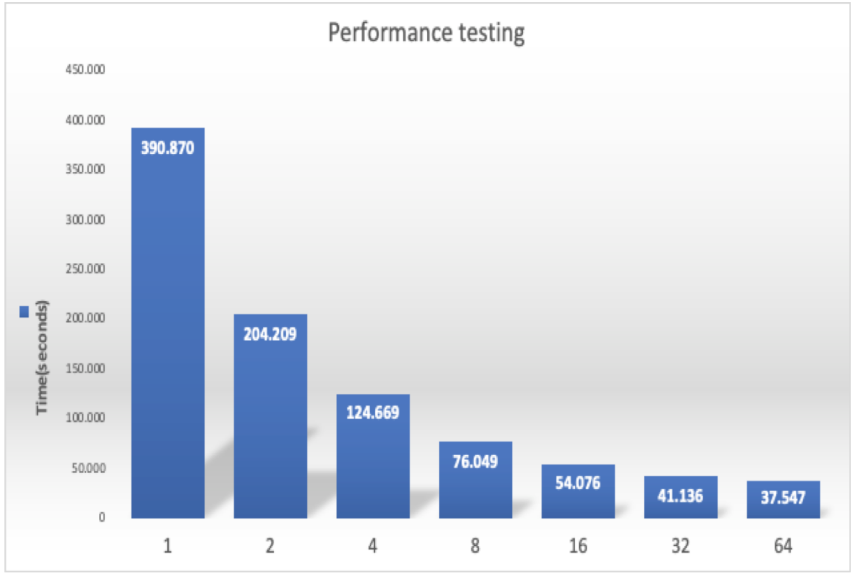
Funding is provided through ERC Advanced grant ERC-2017-ADG #787576 by Martin Schultz

DeepRain is funded by the Bundesministerium fuer Bildung und Forschung (BMBF),
under grant agreement 01 IS18047A

THANK YOU

SCALABILITY PERFORMANCE





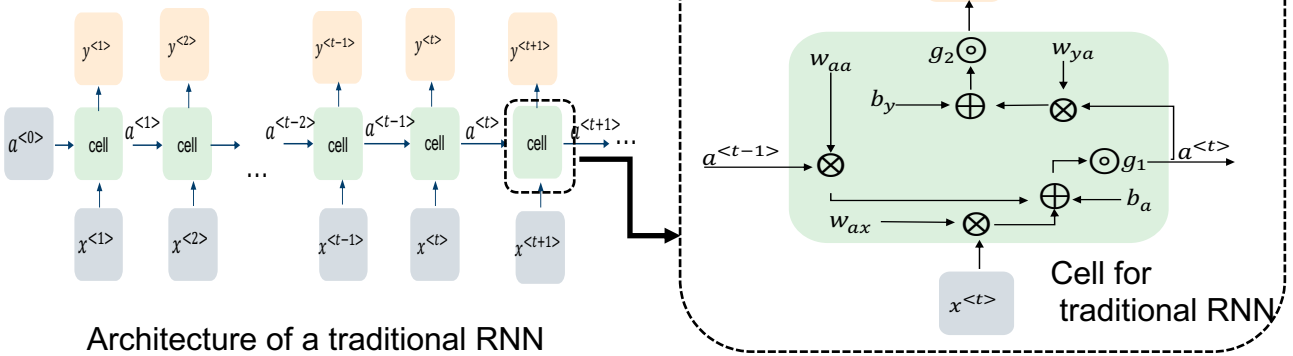
Parallel training for PredNet on JUWELS (left) from JSC and Piz Daint (right) from CSCS

- We carry out the parallel training in light of PredNet architecture for our first try.
- We scale the training process from 1 to 20 GPUs on JUWELS system in Jülich supercomputing Center (JSC). We also test this scalability with container technology on a different system – Piz Daint from Swiss National Supercomputing Center (CSCS).
- So far, some issues remain with respect to convergence of results, and the parallel training for SAVP is not implemented yet.

PredNet: https://coxlab.github.io/prednet/

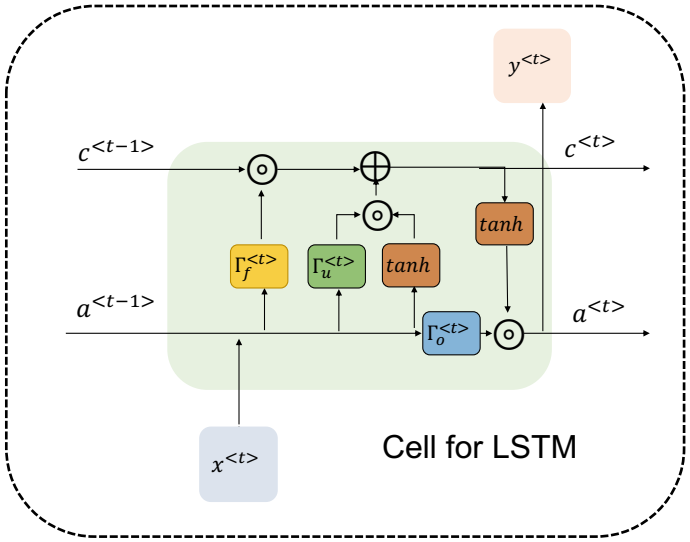
DEEP LEARNING ARCHITECTURES

RNN



Architecture of a traditional RNN

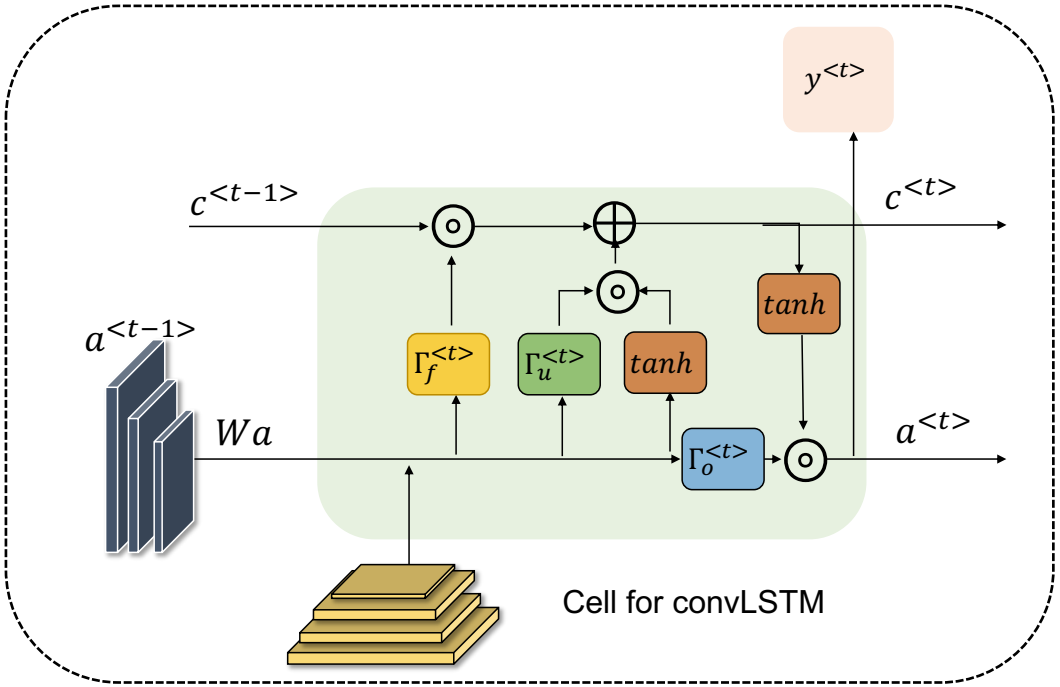
- **RNN (**Rumelhart et al. 1985**)** demonstrates the capability of capture the temporal information of data through connecting previous information to the present.
- **Traditional RNN** fails to reserve the long-term dependencies, leading to gradient vanish (Why? Bengio, 1994)



- **LSTM (**Hochreiter and Schmidhuber, 1997**)** consists of forget gate, update gate, and output gate.
- Forget gate Γ_f outputs between 0-1 made by sigmoid function to decide whether let the information $c^{<t-1>}$ from previous cell going through.
- Update gate Γ_u use sigmoid function to decide whether the new information should be updated. While the values of new information made by tanh will be multiply by updated gate to create an update to the state.
- Output gate Γ_o will decide which parts of the cell state that are going to output.

DEEP LEARNING ARCHITECTURES

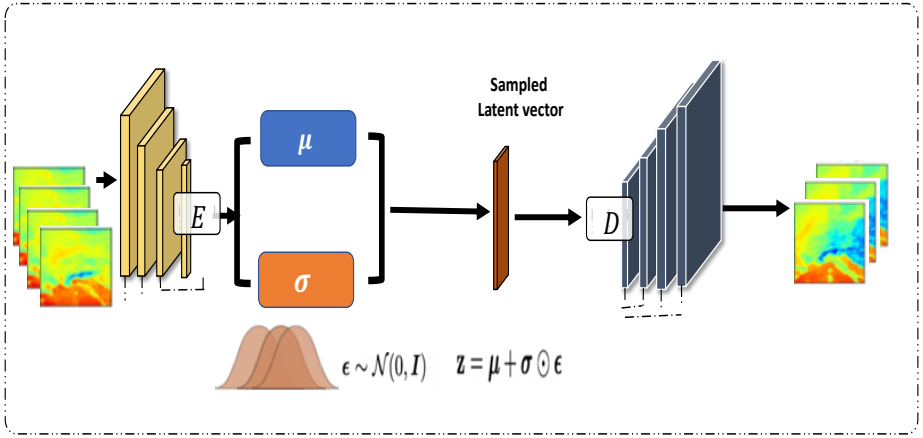
RNN



- **convLSTM** (Xingjian et al. 2015), performs convolution operator in each LSTM cell.
- Studies adopt RNN based architecture: **PredRNN** (Wang Y,. et al., 2017), **MIM** (Wang, Yunbo, et al. 2019)

DEEP LEARNING ARCHITECTURES

VAE

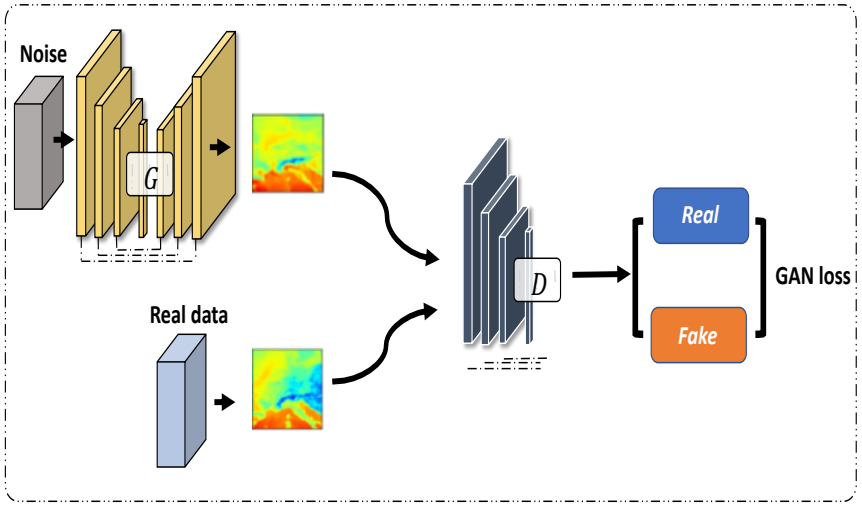


- **VAE** (Kingma & Welling 2013) is an approach to compress the input information from high dimension spaces to lower ones by convolutional layers, Then it reconstructs the original dimension in a generative process way by transposed convolutional layers with constraining the latent space's distribution resembling a predefined distribution (e.g. Gaussian distribution).

- Studies adopt VAE based architecture: **DDPAE** (Hsieh, Jun-Ting, et al., 2018)

DEEP LEARNING ARCHITECTURES

GAN



- The idea of this **GAN** (Goodfellow et al. 2014) is that, the generator is trained to generate images while the learned discriminator network tries to classify the generated image is real or not.
- Studies adopt GAN based architectures**: Adv+GDL (**Mathieu, et al. 2015**), VGAN** (Vondrick , et al. 2016), **MD-GAN** (Xiong , et al. 2018)

EXPERIMENT SETTINGS

Experiment settings

- Variables:
  - We use three variables to represent three channels of the frame as input.
  - 3 * Temperature: three duplicated temperature variables
- Mode:
  - **End_to_end:** perform the entire workflow and use our pre-processed datasets for training.
  - **KTH_pretrained:** use pre-trained models on KTH dataset from the study Lee et al. 2018 for the 2m temperature forecasting task.
  - KTH dataset consists of human six activities: walking, jogging, running, boxing, handwaving, and have clapping. It is a benchmark dataset, which is widely used in video prediction task (Schuldt et al. 2004).
  - The pre-trained models can be accessed from http://rail.eecs.berkeley.edu/models/savp/pretrained_models

Data Extraction

Data Pre-processing

Training

Post-processing

RESULTS

Evaluation metrics

Peak Signal to Noise Ratio (**PSNR)** measures the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its presentation

$$PSNR(\hat{Y}, Y) = 10\ log_{10}\ \frac{max_y^2}{MSE})$$

Structural Similarity Index (**SSIM**) is used for evaluating the similarity between two images in terms of luminance l, contrast c, and structure s

$$SSIM(\hat{Y}, Y) = [l(\hat{Y}, Y)]^{\alpha}.[a(\hat{Y}, Y)]^{\beta}.[s(\hat{Y}, Y)]^{\gamma}$$

Mean Square Error **(MSE) :** used for comparing the difference between two images.

$$MSE(\hat{Y}, Y) = \frac{1}{N}\ (\sum_{i=1}^{N}(\hat{Y}_i\ -\ Y_i)^2)$$

Skill scores: skill score compares the target model t and reference model ref based on metric m. Here, m_{pref} denotes the metric value of a perfect forecast.

$$skill_m = (m\ -\ m_{ref})/(m_{pref}\ -m_{ref})\ )$$

In terms of the PSNR and SSIM, better model performance is associated with higher values, while better MSE scores are smaller values.

Data Extraction

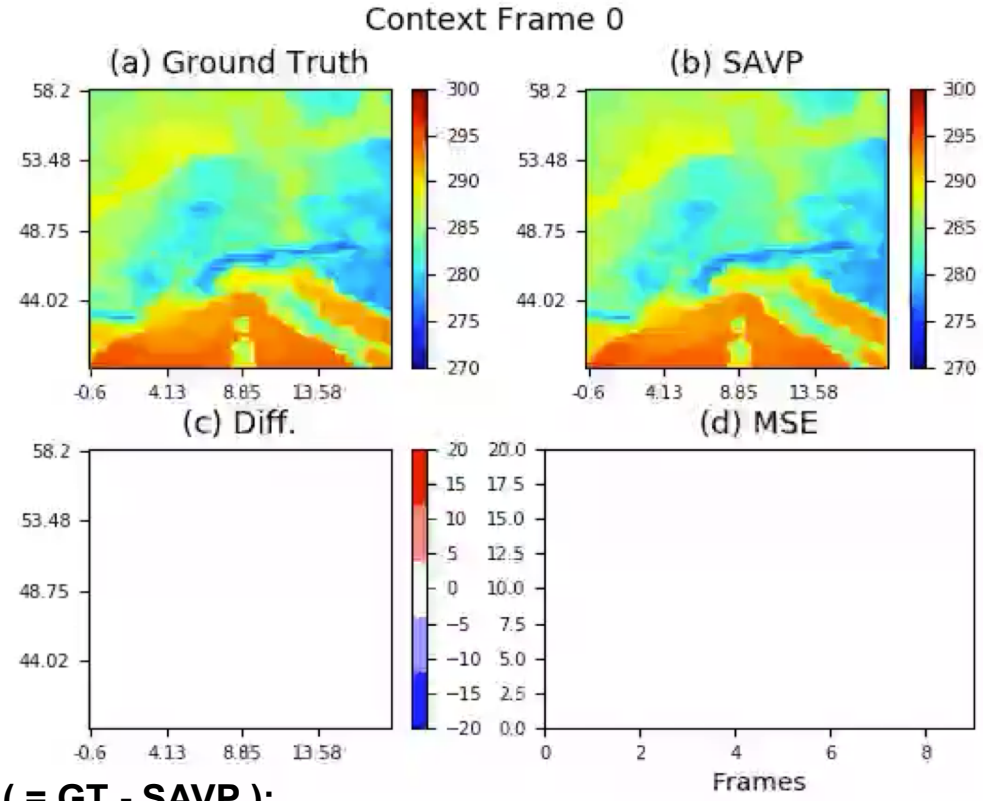
Data Pre-processing

Training

Post-processing

RESULTS

Visualization – SAVP



(C) Diff (= GT - SAVP):

- The value towards red color means the temperature were underestimated;
- The blue color corresponds to the overestimated values.

Data Extraction

Data Pre-processing

Training

Post-processing

RESULTS

Persistence analysis

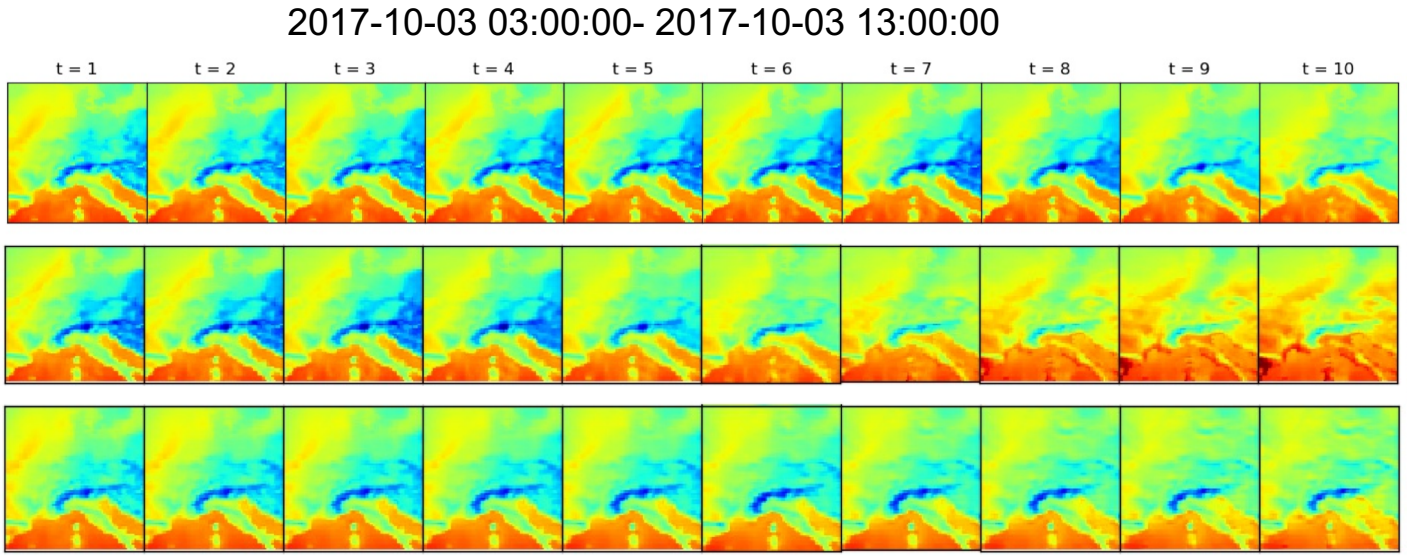


Figure. Row 1: Ground truth; Row 2: Persistent; Row 3: SAVP Prediction

Data Extraction

Data Pre-processing

Training

Post-processing

REFERENCES

Lee, Alex X., et al. "Stochastic adversarial video prediction." arXiv:1804.01523 (2018).

Lotter, William, Gabriel Kreiman, and David Cox. "Deep predictive coding networks for video prediction and unsupervised learning." arXiv:1605.08104 (2016)

Wang, Y., Long, M., Wang, J., Gao, Z., Philip, S.Y.: PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms. In: Advances in Neural Information Processing Systems. (2017)

Sergeev, Alexander, and Mike Del Balso. "Horovod: fast and easy distributed deep learning in TensorFlow." *arXiv preprint arXiv:1802.05799* (2018).

Lee, Alex X., et al. "Stochastic adversarial video prediction." arXiv preprint arXiv:1804.01523 (2018).

Xingjian, S. H. I., et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." Advances in neural information processing systems. 2015.

Wang Y, Long M, Wang J, Gao Z, Philip SY. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. InAdvances in Neural Information Processing Systems 2017 (pp. 879-888).

Wang Y, Zhang J, Zhu H, Long M, Wang J, Yu PS. Memory In Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity from Spatiotemporal Dynamics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019

Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. InProceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 2004 Aug 26 (Vol. 3, pp. 32-36). IEEE.

Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks. 1994 Mar;5(2):157-66.

Click the reference to go back to the slide

REFERENCES

Hsieh JT, Liu B, Huang DA, Fei-Fei LF, Niebles JC. Learning to decompose and disentangle representations for video prediction. InAdvances in Neural Information Processing Systems 2018 (pp. 517-526).

Yan Y, Xu J, Ni B, Zhang W, Yang X. Skeleton-aided articulated motion generation. InProceedings of the 25th ACM international conference on Multimedia 2017 Oct 19 (pp. 199-207).

Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013 Dec 20.

Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. InAdvances in neural information processing systems 2014 (pp. 2672-2680).

.

Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science; 1985 Sep

Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440. 2015 Nov 17.

Vondrick C, Pirsiavash H, Torralba A. Generating videos with scene dynamics. InAdvances in neural information processing systems 2016 (pp. 613-621).

Xiong W, Luo W, Ma L, Liu W, Luo J. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 2364-2373).