Temperature Forecasting using Tower Networks

Siri S. Eide
sirise@met.no
Norwegian Meteorological Institute
Oslo, Norway

Michael A. Riegler
SimulaMet
Oslo, Norway

Hugo L. Hammer
Oslo Metropolitan University
Oslo, Norway

John Bjørnar Bremnes
Norwegian Meteorological Institute
Oslo, Norway

ABSTRACT

In this paper, we present the tower network, a novel, computationally lightweight deep neural network for multimodal data analytics and video prediction. The tower network is especially useful when it comes to combining different types of input data, a problem not greatly explored within deep learning.

The architecture is further applied to a real-world example, where information from historic meteorological observations and numerical weather predictions are combined to produce high-quality forecasts of temperature for 1 to 6 hours into the future.

The performance of the proposed model is assessed in terms of root mean squared error (RMSE), and the tower network outperforms even state-of-the-art forecasts from the Norwegian weather forecasting app `yr.no` from 3 hours into the future. On average, the RMSE of the tower network is approximately 6 % smaller than that of `yr.no`, and approximately 27 % smaller than that of the raw numerical weather predictions.

CCS CONCEPTS

• **Applied computing** → **Earth and atmospheric sciences**; • **Computing methodologies** → **Machine learning**; **Computer vision**.

KEYWORDS

tower network, temperature forecasting, video prediction, deep learning

ACM Reference Format:
Siri S. Eide, Michael A. Riegler, Hugo L. Hammer, and John Bjørnar Bremnes. 2021. Temperature Forecasting using Tower Networks. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3463944.3469099

1 INTRODUCTION

Multimedia is a term that encompasses infinite possibility, far beyond the combination of text and images. While it is often associated with entertainment, it is also very relevant in other fields, such as weather forecasting, where there are a number of different types of data that can be combined, e.g., radar, satellite images and weather station measurements.

Weather prediction is a complex and multifaceted problem. There are intricate relationships between physical properties of the atmosphere such as temperature, moisture, pressure, wind and so on, across time and in three spatial dimensions. The relationships occur on many scales, from micro to global, and can follow anything from sub-hourly to multi-annual cycles. A perfect model would need to understand all of these interactions, some of which are chaotic and unpredictable [12].

Operational weather forecasting is today done largely using what is called Numerical Weather Prediction (NWP) models. These are complex mathematical models based on the fundamental laws of standard physics, which aim to simulate the future state of the atmosphere on a three-dimensional grid. The NWP models make many approximations for processes that are smaller than the resolution of the model grid. However, with advances in technology, it is possible to run these models with increasingly high spatial resolution, providing more detailed forecasts.

Because of the enormous amount of data that is available, the possibility of using multimodal machine learning, and more specifically deep learning, in weather forecasting has been getting more attention in recent years. Nevertheless, mulitmodal machine learning is not very common and the focus lies mainly on single modalities, for example, [3] explored challenges and design choices for global weather and climate models based on machine learning.

Much of the work in deep learning weather forecasting has been focused on precipitation (condensation of atmospheric water vapor leading to for example drizzle or rain). For example, [15] proposed the convolutional LSTM for precipitation prediction for a very short time period, also called nowcasting, [14] used multi-task convolutional networks, and [16] developed MetNet, a highly complex deep neural network that outperformed the best operational NWP available. [2] performed high-resolution precipitation forecasting from radar images using a UNET convolutional network, and [1] developed the Temporal Recurrent U-Net.

Similar work has been done with regards to air temperature. [7] used historical observations in a deep neural network with Stacked Denoising Auto-Encoders. [8] used stacked LSTMs and [10] applied the convolutional LSTM networks introduced by [15] to a temperature forecasting problem, while [5] compared Stochastic Adversarial Video Prediction [11], Generative Adversarial Networks [6] and Variational Auto-Encoders [9].

ICDAR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8529-9/21/08.
https://doi.org/10.1145/3463944.3469099

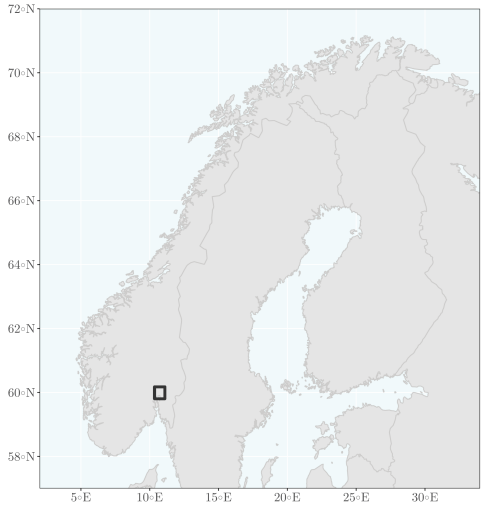

Figure 1: The geographical area used in this work.

There are many possible ways in which one can try to incorporate machine learning into weather forecasting or even into the NWP models themselves. The focus of this paper is to tackle the task as a multimodal data analysis problem through improving short-term temperature forecasts from NWP models by combining them with historical observations using a new, lightweight deep learning method. Multimodal data analysis in deep learning is not very deeply explored. Most approaches use early or late fusion for the analysis, which means either combining the features before the analysis or analysing them separately and combining them afterwards. Both come with their respective advantages and disadvantages. For this work we propose a different approach that combines several neural networks in one ensemble which allows feature extraction and training for all modalities at once. We call this type of network a tower network since each possible data source, or different view on the same data source, is represented as its own neural network. These networks are combined via a concatenation layer at the end which is then handed over to a dense network to either perform classification or regression. In this work we show the potential of these architectures for the weather forecasting task, but it is important to point out that this can be applied for any other multimodal data analysis problem.

2 METHOD

2.1 Data

The methods applied in this paper use two data sources. Firstly, there is a dataset containing observations of temperature at 2 m above the ground, which have been interpolated to a grid with a spatial resolution of 1×1 km [13]. The grid covers all of Norway, but for this work, a subset of 40 by 40 grid points, shown in Figure 1, has been selected. The subset is centered around Oslo, and includes both urban areas, forest and a small part of the Oslofjord inlet. The altitudes range from 0 to 647 above sea level. The observations cover a five year period from 2014 through 2018, and have a temporal resolution of 1 hour.

Secondly, there are NWP model data on a grid with a spatial resolution of 2.5×2.5 km [4], i.e. slightly lower resolution than the gridded observations, covering the same period with the same temporal resolution. These data have been regridded to have the same 1×1 km resolution as the observations. While the forecasts are hourly, fresh forecasts are only available 4 times daily, at 00, 06, 12 and 18 UTC.

For a period starting in February of 2018, data from yr.no are also available [13]. `yr.no` is a weather forecasting website and app hosted by the Norwegian government-owned national broadcasting corporation (NRK) and the Norwegian Meteorological Institute. The `yr.no` data are official forecasts based on the aforementioned NWP model data, that have gone through various forms of post-processing. These data are used for comparison, as an example of state-of-the-art weather forecasting.

2.2 Models

This paper deals with the problem of short term prediction of temperature, more specifically, forecasting temperature from 1 to 6 hours into the future. Intuitively, it can be surmised that the data have relationships in both space and time. It is unlikely to see large temperature differences within a small area at any given time, and extreme temperature changes from one hour to the next are rare. Therefore it makes sense to look at this as a video prediction problem, where given a set of frames (made up of the grid points in the data) the following frames will be predicted.

The primary method used in this paper is the tower network. Tower networks are a new type of neural network developed and conceptualized during this work. The networks are built up of so-called "towers" (see Figure 2), which consist of stacks of convolutional layers, batch normalization layers, max pooling and activation layers. Inputs are sent to the towers, which learn different views on the data. The output from the towers is then concatenated. A convolutional layer ensures the correct dimensions for the output predictions. By varying the kernel size and/or stride of the convolutional layers between the towers, the different towers are able to learn data patterns on different spatial scales. This method should theoretically be a good fit for weather forecasting, since the weather is influenced by both large and smaller scale systems.

Since the observations are historical data, while the NWP data are predictions for the future, there is no obvious way to combine the two directly. We therefore propose a modified, multimodal variant of the tower network. The historical observations are sent through three towers, while the NWP data go directly into the concatenation layer. After concatenation, layers equivalent of another tower are added. The new elements are shown in dashed lines in Figure 3.

The suggested models are further compared against a traditional statistical approach, namely the first order autoregressive model, AR(1), specified in each grid point. In AR(1), the observation X_t at time t is expressed in terms of the the previous time step in the following way:

$$X_t = C + \varphi X_{t-1} + \varepsilon_t$$

where C is a constant, φ is a regression parameter and ε_t is white noise.

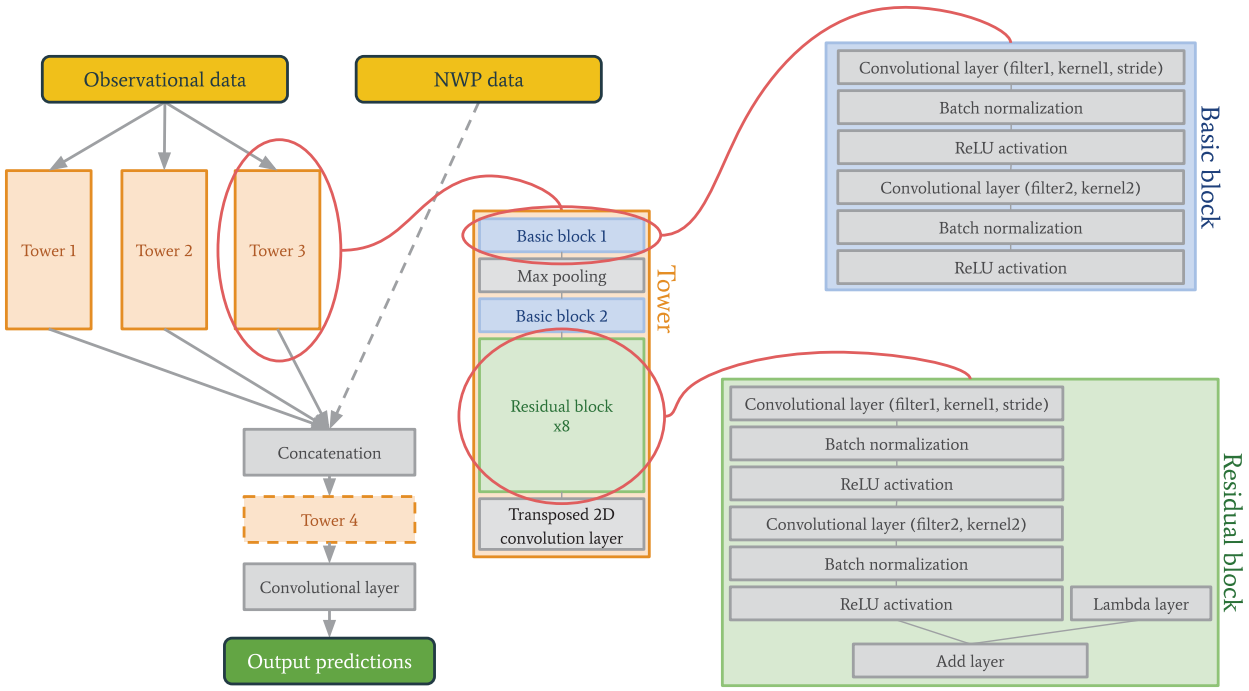

Figure 2: Network architecture for the normal and the modified tower network.

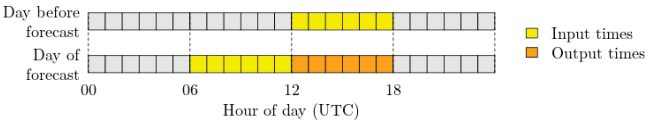

Figure 3: One sample of observation data.

3 RESULTS

The data were split into training, validation and testing sets, using 2014 to 2016 for training, 2017 for validation and 2018 for testing. Data from `yr.no` were only available starting on 19 February 2018, so in order to get a fair comparison, data from before this date are naturally not used when `yr.no` is included.

A sample of observation data consists of 12 hours used for input with 6 hours to be predicted. The input data are observations from 24 hours before the predicted times, as well as the last 6 hours before prediction, see Figure 3. By always predicting temperature for the hours from 13 UTC to 18 UTC, we simplify the prediction problem, as the model does not need to take into account the daily cycle present in temperature data. This also facilitates the use of and comparisons with the NWP data for which 12 UTC is one of the production times.

The tower networks are reasonably computationally light and used only 20 minutes to train a model on a NVIDIA V100 GPU. It was therefore possible to test a number of different architectures and to perform thorough parameter tuning. All parameters for the final architecture are presented in Table 1.

Table 1: All network parameters

		filters 1	filters 2	kernel size 1	kernel size 2	strides
Tower 1	Basic block 1	64	32	8	3	1, 1
	Basic block 2	64	32	3	3	1, 1
	Residual blocks	64	32	3	3	1, 1
	Transposed 2D	64	-	4	-	2, 2
Tower 2	Basic block 1	64	32	8	3	1, 2
	Basic block 2	64	32	3	3	1, 2
	Residual blocks	64	32	3	3	1, 1
	Transposed 2D	64	-	4	-	2, 8
Tower 3	Basic block 1	64	32	8	3	1, 4
	Basic block 2	64	32	3	3	1, 4
	Residual blocks	64	32	3	3	1, 1
	Transposed 2D	64	-	4	-	2, 20
Tower 4	Basic block 1	64	32	8	3	1, 1
	Basic block 2	64	32	3	3	1, 1
	Residual blocks	64	32	3	3	1, 1
	Transposed 2D	64	-	4	-	2, 2
Final convolutional layer		6	-	8	-	-

For all tower networks trained, the optimizer used is adaptive moment estimation (Adam), with mean squared error loss function, a batch size of 10, and up to 2000 epochs chosen with early stopping.

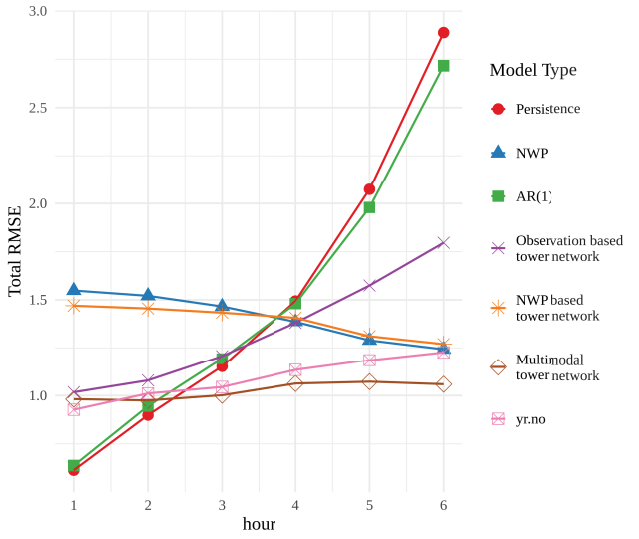

Figure 4: Root mean squared error averaged over the spatial grid.

The suggested model was compared against persistence which is a common baseline for short term forecasting in meteorology. Persistence tends to perform well in the first few hours, as the weather often does not change very quickly. It is also natural to compare any method to the performance of the raw NWP model, as well as the forecasts from `yr.no`.

Figure 4 shows the root mean squared error (RMSE) of the multimodal tower network compared to persistence, the NWP model, a first-order autoregressive model, a tower network trained only on observational data, one trained only on NWP data and finally temperature forecasts from `yr.no`, for predictions from one to six hours into the future. RMSE is computed taking the mean over both time and locations.

Persistence and the NWP model illustrate quite clearly the different qualities of observational and NWP model data. As expected, persistence performs very well in the first hour, but becomes worse every hour. The NWP model performs relatively evenly, but is best for hours 5 and 6.

The AR(1) model follows the same trend as persistence, performing well for the first two hours but losing performance over time.

The tower network based purely on observational data performs much better than the NWP model for the first hours, but can not compete with persistence. From hour 4 the NWP model beats the tower network, and persistence performs worse.

The tower network using only NWP data improves upon the NWP model for the first 3 hours, but is in fact slightly worse for hours 4 through 6.

The multimodal tower network combines historical observations and numerical weather prediction data and is overall the model with the best performance. The fact that both this and the observation based tower network perform so much better than the NWP model for the first hours indicates the importance of historical observations in predicting the near future. However, neither model achieves the same performance as the persistence, which is the clear winner in the two first hours. From hour 3 onward, the multimodal tower network is considerably better than both persistence and the NWP model, suggesting that combining these two data sources gives valuable insight beyond what they can provide separately.

The most interesting comparison to make is perhaps against what the common person sees every day when checking the weather on their phone. This is included in Figure 4 in the form of forecasts from `yr.no`, and we see that the multimodal tower network outperforms them from hour 2, with a growing margin for each hour. It is worth noting that the tower networks are beaten by both persistence and `yr.no` in the first hour, which suggests that there is still potential for improvement.

The different methods were also compared in terms of mean absolute error which resulted in similar conclusions.

Figure 5, is a visual presentation of one prediction from the test data. Each row is a model, with the top row displaying the ground truth, and each column is an hour. The observations (top row) show that there is a general reduction in temperature over the six hours, which the NWP model, the multimodal tower network and `yr.no` capture quite well. Persistence is by definition unchanging, and so remains warm for the six hour period. The AR(1) model exhibits more or less the same behaviour as persistence, while the observation based tower network both starts out and stays too warm, even though it also has a small temperature reduction.

It is easy to see that the NWP model, which has been interpolated from a grid with lower resolution, does not have the level of detail present in the other models. The NWP based tower network seems to have learned some smaller-scale features, but is distinctly less detailed than the other models. The remaining models appear to share many of the same shapes and features, and none look completely unrealistic.

The example presented in Figure 5 is the coldest day in the test set. This day was chosen because extremes are typically more difficult for the models to handle, and could give an indication of whether or not the models would still be able to produce realistic looking forecasts.

4 CONCLUSIONS AND FURTHER WORK

This paper shows how tower networks can be a useful tool in combining different data sources using deep neural networks, based on the example of using observational data along with output from numerical weather prediction models in order to create short-term weather forecasts of high quality. The simplicity and flexibility of the method means that it can easily be adapted to accommodate other types of input data if necessary.

So far, we have only looked at the model's performance in terms of summarized scores. Future work will therefore involve diving deeper into analysis of the results, e.g. by breaking them up according to different thresholds. Since mean squared error was used as the model's loss function, we need to investigate the degree to which the model is able to predict extremes.

It would also be interesting to try to extend the prediction horizon beyond +6 hours and see how far we can go while still maintaining good performance.

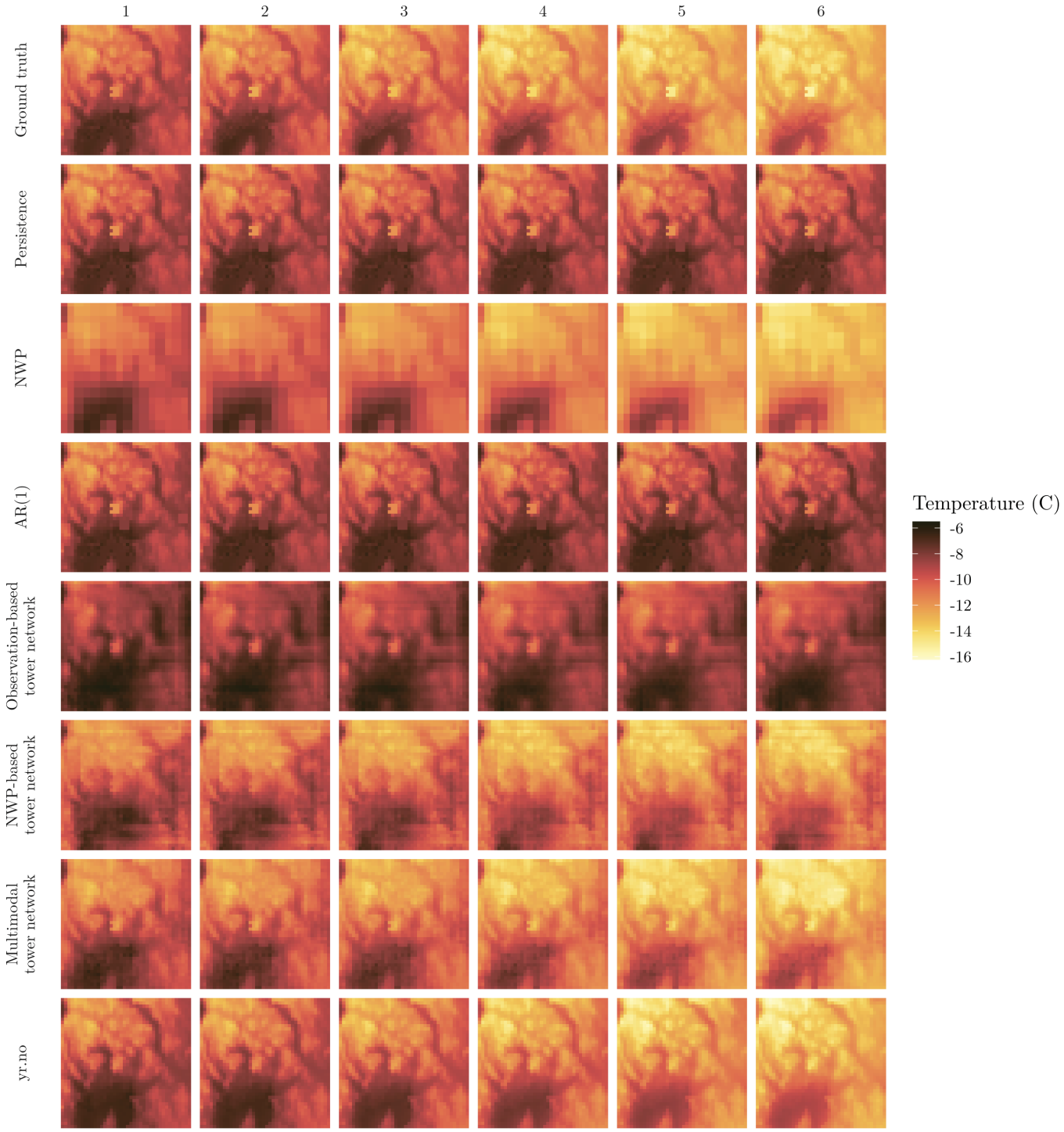

Figure 5: Example of forecasts with the ground truth for reference.

ACKNOWLEDGMENTS

A huge thank you to Steven Hicks for his contribution to the development of the code.

REFERENCES

[1] Rilwan Adewoyin, Peter Dueben, Peter Watson, Yulan He, and Ritabrata Dutta. 2021. TRU-NET: A Deep Learning Approach to High Resolution Prediction of

Rainfall. arXiv:2008.09090 [cs.CE]
[2] Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. 2019. Machine Learning for Precipitation Nowcasting from Radar Images. arXiv:1912.12132 [cs.CV]
[3] Peter D. Dueben and Peter Bauer. 2018. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development* 11, 10 (Oct. 2018), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018
[4] Inger-Lise Frogner, Ulf Andrae, Jelena Bojarova, Alfons Callado, Pau Escribà, Henrik Feddersen, Alan Hally, Janne Kauhanen, Roger Randriamampianina, Andrew Singleton, Geert Smet, Sibbo van der Veen, and Ole Vignes. 2019. HarmonEPS—The HARMONIE Ensemble Prediction System. *Weather and Forecasting* 34, 6 (Dec. 2019), 1909–1937. https://doi.org/10.1175/waf-d-19-0030.1
[5] Bing Gong, Severin Hußmann, Amirpasha Mozaffari, Jan Vogelsang, and Martin Schultz. 2020. Deep learning for short-term temperature forecasts with video prediction methods. (March 2020). https://doi.org/10.5194/egusphere-egu2020-17748
[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]
[7] Moinul Hossain, Banafsheh Rekabdar, Sushil J. Louis, and Sergiu Dascalu. 2015. Forecasting the weather of Nevada: A deep learning approach. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE. https://doi.org/10.1109/ijcnn.2015.7280812
[8] Zahra Karevan and Johan A. K. Suykens. 2018. Spatio-temporal Stacked LSTM for Temperature Prediction in Weather Forecasting. arXiv:1811.06341 [cs.LG]
[9] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]
[10] David Kreuzer, Michael Munz, and Stephan Schlüter. 2020. Short-term temperature forecasts using a convolutional neural network — An application to different weather stations in Germany. *Machine Learning with Applications* 2 (Dec. 2020), 100007. https://doi.org/10.1016/j.mlwa.2020.100007
[11] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. 2018. Stochastic Adversarial Video Prediction. arXiv:1804.01523 [cs.CV]
[12] EDWARD N. LORENZ. 1969. The predictability of a flow which possesses many scales of motion. *Tellus* 21, 3 (June 1969), 289–307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x
[13] Thomas N. Nipen, Ivar A. Seierstad, Cristian Lussana, Jørn Kristiansen, and Øystein Hov. 2020. Adopting Citizen Observations in Operational Weather Prediction. *Bulletin of the American Meteorological Society* 101, 1 (Jan. 2020), E43–E57. https://doi.org/10.1175/bams-d-18-0237.1
[14] Minghui Qiu, Peilin Zhao, Ke Zhang, Jun Huang, Xing Shi, Xiaoguang Wang, and Wei Chu. 2017. A Short-Term Rainfall Prediction Model Using Multi-task Convolutional Neural Networks. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE. https://doi.org/10.1109/icdm.2017.49
[15] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv:1506.04214 [cs.CV]
[16] Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. 2020. MetNet: A Neural Weather Model for Precipitation Forecasting. arXiv:2003.12140 [cs.LG]