



A Hybrid Deep Learning Approach for Forecasting Air Temperature

Gregory Gygax and Martin Schüle^(✉)

Institute for Applied Simulation, Zurich University of Applied Sciences ZHAW,
Wädenswil, Switzerland
{gregory.gygax,martin.schuele}@zhaw.ch

Abstract. Forecasting the weather is a great scientific challenge. Physics-based, numerical weather prediction (NWP) models have been developed for decades by large research teams and the accuracy of forecasts has been steadily increased. Yet, recently, more and more data-driven machine learning approaches to weather forecasting are being developed. In this contribution we aim to develop an approach that combines the advantages of both methodologies, that is, we develop a deep learning model to predict air temperature that is trained both on NWP models and local weather data. We evaluate the approach for 249 weather station sites in Switzerland and find that the model outperforms the NWP models on short time-scales and in some geographically distinct regions of Switzerland.

Keywords: Weather forecasting · Deep learning · Hybrid modeling

1 Introduction

To forecast the weather is a long-standing scientific challenge. Also, accurate weather forecasts have great economic impact and mitigate costs to lives and assets in the case of high-impact weather.

Our weather is produced by a physical atmospheric system with complex dynamics. The usual meteorological models used in numerical weather prediction (NWP) are based on modeling the atmospheric dynamics and atmosphere-land-sea couplings. The simulation of these models, initialized by a wealth of measured and inferred data, then allows to forecast various parameters such as air temperature or precipitation. In recent years and decades, the models and simulation techniques have been developed to the point where they allow a fairly accurate weather forecast for up to 10 days.

Despite these successes of the common meteorological models, more and more work is recently being undertaken that aims to produce weather forecasts using a machine learning (ML) approach. It is hoped that this further improves weather forecasting, especially in terms of fast and accurate spatio-temporal resolution. Such forecasting on a relatively short time-scale is also called nowcasting.

Instead of aiming to replace the elaborate physical models completely by ML approaches, we propose in this contribution a hybrid approach, combining the

NWP models with a deep learning (DL) approach. We believe this combines the advantages of NWP models such as an accurate representation of atmospheric physics and a global approach to weather forecasting with the advantages of data-driven ML approaches such as fast and comprehensive local data analysis.

Our approach develops a DL model which is trained on local measurement data of weather stations and the corresponding forecasts of the NWP models plus local past weather data. The main contributions of this work can be summarised as follows: 1. We analyze the performance of the main NWP models for Switzerland in regard to forecast accuracy of air temperature at 249 weather station sites, 2. We design a DL model that learns to make local air temperature forecasts based on the performance of the NWP models and additional local data. As we will discuss in more detail, the results indicate that our approach allows to generate improved forecasts on short time-scales and for some geographically distinct regions of Switzerland.

The rest of the work is structured as follows: Next, we discuss the background and related work. In Sect. 3 we describe the data situation and the NWP models used. In Sect. 4 we describe the technical details of our approach. In Sect. 5 we present the results of experiments with the model on Swiss weather data. In Sect. 6 we conclude with a discussion of our approach and results.

2 Background and Related Work

Weather prediction has a long and successful history. In numerical weather prediction (NWP) the methodology usually centers on modeling the physics of the atmosphere and taking samples from numerical simulations of these models to generate forecasts [1, 2]. The approach essentially consists in modeling the physics of the atmosphere and its couplings e.g. to sea and land, in model initialization schemes, and running large simulations on super-computing facilities. Ensemble modeling allows for an estimate of the uncertainty of forecasts. Advanced computational techniques are used in order to run simulations of models. The current state of the art in numerical weather prediction is reviewed by Bauer et al. [3]. The NWP models are developed at large research centers such as the European Centre for Medium-Range Weather Forecasts (ECMWF). We will also use ECMWF models in this work (see Table 1), but refrain from explaining these models in more detail as this does not constitute the focus of this contribution.

In recent years there have been more and more approaches to weather forecasting with ML models. These models sometimes try to predict a number of parameters [4, 5], as NWP models, but often the focus is on certain parameters, e.g. on precipitation forecasts [7, 8]. Also, some authors devise a hybrid approach, by combining different models or modeling strategies, where others rely on a straight ML or DL modeling approach. Our approach is in the spirit of Reichstein et al. [6] where the authors argue for hybrid modeling strategies for the earth sciences, combining physical models with ML approaches.

Some work on weather forecasting with DL methods we would like to mention specifically: Grover et al. [4] develop a hybrid approach where they combine

machine learning algorithms locally trained on key weather variables with a deep learning model that models the joint statistics of the variables and a statistical method for spatial interpolation. The model predicts wind, temperature, pressure and dew point for weather station locations in the US and in some cases outperforms the NOAA (National Oceanic and Atmospheric Administration) models. Weyn et al. develop a convolutional neural network (CNN) approach to model the atmospheric state [5]. Xingjian et al. develop a convolutional LSTM model for precipitation nowcasting and outperform an operational precipitation forecasting algorithm using radar map data [7]. Hernández et al. use an autoencoder and FNN (feedforward neural network) to forecast accumulated daily precipitation for a meteorological station in Colombia [8]. The cited work shows that the main DL models such as FNN, CNN, conv-LSTM, etc. are currently being explored for weather forecasting tasks.

3 Description of Data and Weather Prediction Models

In order to build and evaluate our approach we use weather data collected by weather stations and historic weather forecast data generated by some of the main NWP models for Europe.

In regard to the measurement data we selected a number of key weather parameters and collected these for 249 weather stations locations in Switzerland for the time period 1990–2020. In this contribution we however only consider the mean temperature 2 m above ground in 1 h frequency. The data was collected by MeteoSwiss (Swiss Federal Institute of Meteorology and Climatology) weather stations and provided by Meteomatics, a private weather data provider.¹

The forecast data was collected for the NWP models listed in Table 1 for the time period 2019-09-17 to 2020-03-24. These models constitute some of the main NWP models for Europe. Some regional models such as COSMO are however missing.

We analysed the performance of the NWP models by the mean squared error (MSE) of their forecasts per forecast horizon for the 249 weather station sites for the air temperature 2 m above ground. Figure 1 shows an example for the location Wädenswil, Switzerland. We can see that, as expected, the predictions become worse with growing forecast horizon.

We aim to beat these models or the best of these models in accuracy. However, what is the best model for a given time and location? At some point in time t it is not a priori clear which model will perform best for the next hours and days. We therefore constructed a benchmark model in the following way: given a location and some point in time we determine the model that has performed best in the past 60 h. The forecasts of that model will be picked as comparison to the forecasts made by our own model for the prediction made at time t . This procedure is repeated for every location and point in time. In Fig. 1 the thereby generated benchmark predictions, averaged over the entire time-period, are shown, yielding the lowest prediction errors compared to the original models.

¹ Meteomatics, <https://www.meteomatics.com>. Last accessed June 2020.

Table 1. NWP models used in the work.

Short name	Description
cmc-gem	Global Environmental Multiscale model operated by the Canadian Meteorological Center
ecmwf-ifs	The European Center for Medium-Range Weather Forecasts' (ECMWF) Integrated Forecasting System (IFS). Atmospheric global circulation model used for medium-range forecasts
ecmwf-ens	Ensemble Prediction System (EPS) by ECMWF
ecmwf-mms	Long-range seasonal forecast by ECMWF
ecmwf-vareps	Long-range ensemble forecast by ECMWF
knmi-hirlam	High Resolution Limited Area Model from the Royal Netherlands Meteorological Institute.
mf-arome	Regional model by Meteo France
mix	Mixture model combining different models designed by Meteomatics
mm-swiss1k	High-resolution model for Switzerland designed by Meteomatics
ncep-gfs	Global Forecasting System by the National Centers for Environmental Prediction (NCEP)
ncep-gfs-ens	Ensemble model of Global Forecasting System by NCEP
ukmo-euro4	European model by the UK MetOffice

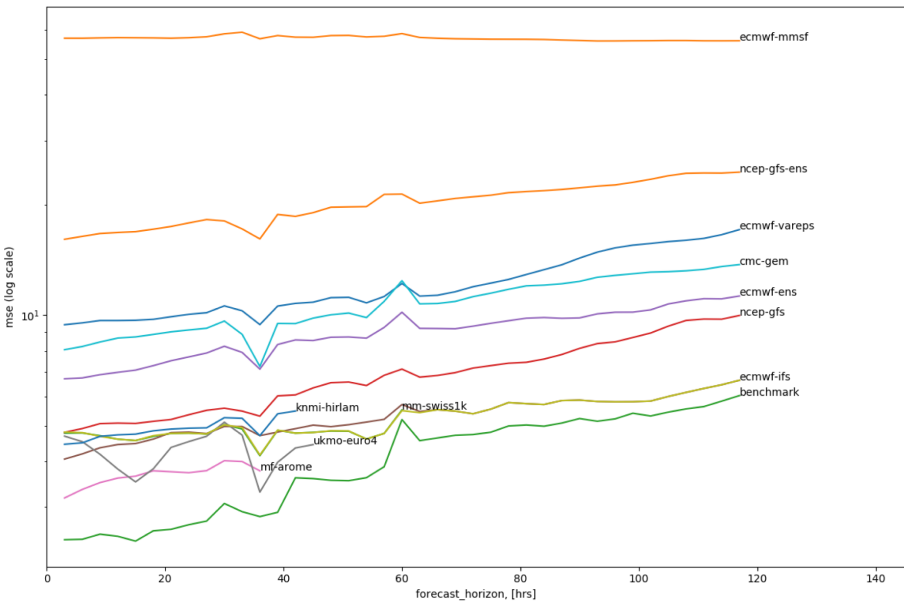


Fig. 1. MSE of the NWP model predictions of the air temperature 2 m above ground for the location Wädenswil, Switzerland for different forecast horizons. The lowest line indicates a benchmark model. Data considered for the time-period 2019-09-17 to 2020-03-24.

4 Method

The main idea is to train locally, i.e., at each selected weather station location, a DL model that takes as input time-series forecast data generated by the NWP models and the measurement data at that site. There are thus several time series, one for the measurements and several for the NWP models, which we collect in one feature time series vector. The target is the forecast h steps into the future which is then compared to the actual, measured values. That is, for each location where data is available (the site of a weather station), the model aims to close the gap between forecasted values and actually recorded values by training the model accordingly on the historic data.

Formally, let $M(t)$ denote the measured value at time t and let $F_h^{(i)}(t)$ denote the prediction made by model i at time t for time $t + h$. In other words, $F_h^{(i)}(t)$ is an estimate of $M(t + h)$. In this contribution, we only consider the air temperature 2 m above ground as value. Because M is available on an hourly basis, but forecasts are available on a 3 h basis only, we split M into 3 time series with a 1 h lag relative to each other and add these new time series to the feature vector. Given a forecast horizon h , we then construct for every time t the target $Y_h(t) = (M(t + h))$ and the input or feature vector

$$X(t) = \begin{pmatrix} M(t) \\ M(t - 1) \\ M(t - 2) \\ F_h^{(0)}(t) \\ \vdots \\ F_h^{(n)}(t) \end{pmatrix}$$

We further transform, as the time series is not stationary, $X(t)$ and $Y(t)$ by subtracting $M(t)$ from each value, yielding $Y_h(t) = (M(t + h) - M(t))$ and

$$X(t) = \begin{pmatrix} 0 \\ M(t - 1) - M(t) \\ M(t - 2) - M(t) \\ F_h^{(0)}(t) - M(t) \\ \vdots \\ F_h^{(n)}(t) - M(t) \end{pmatrix}$$

Finally, we are using $X(t), \dots, X(t - l)$, with look back period $l = 20$, for predicting $Y(t)$. The value for l seems reasonable to us, amounting to a 60 h look back window, but we did not investigate that parameter value further.

The described data preparation was then carried out for each location.

4.1 Model

With relatively small amount of data (ca. 1000 time steps), a small model with one GRU (Gated Recurrent Unit) layer with 2 nodes and a subsequent dense layer seemed appropriate. Larger models quickly overfitted. However we did not yet look systematically into this matter.

4.2 Benchmark

We defined the benchmark as the prediction of the NWP model which performed best during the past 20 time steps (the same time window with a look back period $l = 20$) corresponding to 60 h. Formally, we have

$$bench(t) = F_h^{(i)}(t), \text{ where } i = \arg \min_i \sum_{j=0}^{20} (F_h^{(i)}(t-j) - M(t-j))^2$$

4.3 Training

Data was split for cross validation using the last 10% of the data for testing and the rest for training.

A model was trained using each possible combination of the following parameters:

- Forecast Horizon: 3 h, 6 h, 12 h, 18 h, 24 h
- Station: 1 of 249
- Look back: 20
- Weather parameter: hourly mean temperature 2 m above ground

This results in $5 \cdot 249 = 1245$ models. Each model was trained for 1000 epochs.

4.4 Implementation

All computation was done using python3.8 on linux. Models were built, trained and evaluated on an NVIDIA RTX 2060 GPU using Tensorflow 2.1. We used Tensorflow’s standard implementation of the ‘Adam’ optimizer and the mean squared error (MSE) loss function.

4.5 Evaluation

For each time t , the predicted values were evaluated against the benchmark model predictions. The error (MSE) was computed for the testing and training sets for both the newly trained models and the benchmark model for every station $s \in S$, resulting in $mse_{train}(s)$, $mse_{test}(s)$, $bench_{train}(s)$, $bench_{test}(s)$.

For each station we can then build the differences $mse_{train}(s) - bench_{train}(s)$ and $mse_{test}(s) - bench_{test}(s)$. If the difference is smaller than 0, the new model is better than the benchmark for the given station.

5 Results

The performance of the new model was assessed by looking at the difference of the prediction error to the benchmark error. The error used was the mean squared error (MSE) either over the training period or the testing period. Here, we focus on the results for the testing set.

5.1 Nowcasting and Geographic Differences

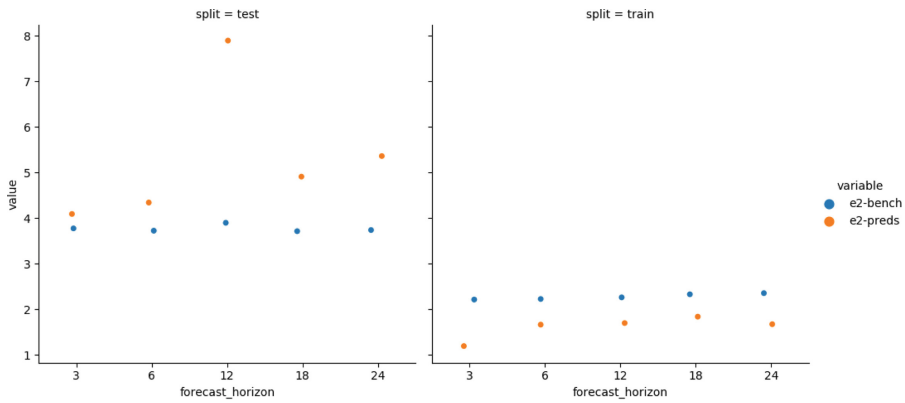
We evaluated the model for 249 weather station sites in Switzerland. We note two main points: 1. The model performs consistently better for a forecast horizon of 3 h, i.e., in the “nowcasting” range; 2. For larger forecast horizons the model performs better for some locations or some regions of Switzerland but not for whole country.

Figure 2 shows two examples for two distinct locations: in a) the model does not perform better while in b) the model performs better over all forecast horizons. As expected, the model performs better than the benchmark model on the training set in both cases.

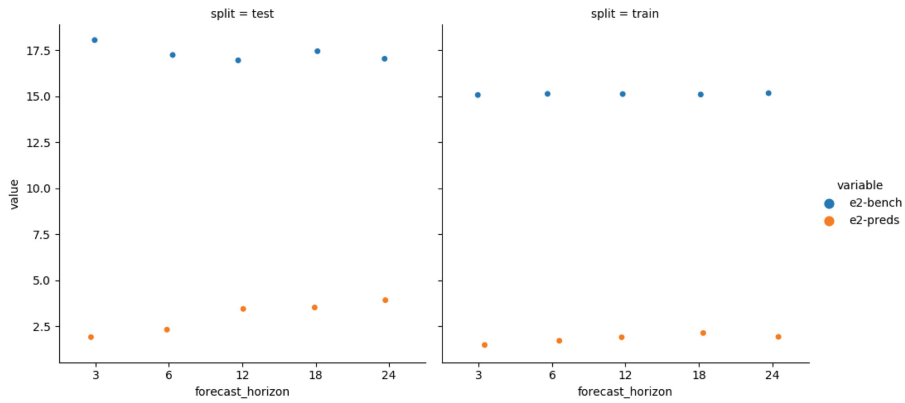
Figure 3 shows an overview of the errors per station for each forecast horizon on a national scale. The model performs well for the 3 h forecast for most stations. Forecasts quality seems to deteriorate for increasing forecast horizon and seem worst for 12 h forecast horizon. Interestingly, there seem to be clusters of locations, e.g. in the canton of Valais, where the model seems to be consistently better than the benchmark model. We therefore decided to look at the Valais example in more detail. Figure 4 shows a zoomed in view on the stations in the canton of Valais.

5.2 Evaluation Metrics and Error Distribution

We analyzed the differences between our model’s performance to the benchmark model performance by looking at the mean and median differences of the MSE of the predictions by our model and the benchmark model, further on referred to as MBP and MEBP, respectively. Also, we assessed the ratio of stations that performed better under the model than with the benchmark model. A station is assumed to perform better than the benchmark, if the difference of $MSE(\text{model}) - MSE(\text{benchmark}) < 0$. We performed this analysis on the level of the forecast horizons once for all available stations and once for the Valais stations. Results are summarized in Table 2. Both tables show that the majority of stations (81% for Switzerland and 89% for Valais) have better forecasts with the new model than with benchmark for the 3 h forecast. While the MPB seems to indicate



(a) Wädenswil, Canton of Zurich



(b) Turtmann, Canton of Valais

Fig. 2. Prediction errors of our model and benchmark model over forecast horizons for a) Wädenswil and b) Turtmann.

better performance for all forecast horizons except the 12 h, the ratio of improved models and the MEBP indicate that these values are probably caused by outliers. In the case of the stations in Valais, there seems to be improvement for all forecast horizons, although the improvement for 24 h forecast is very small.

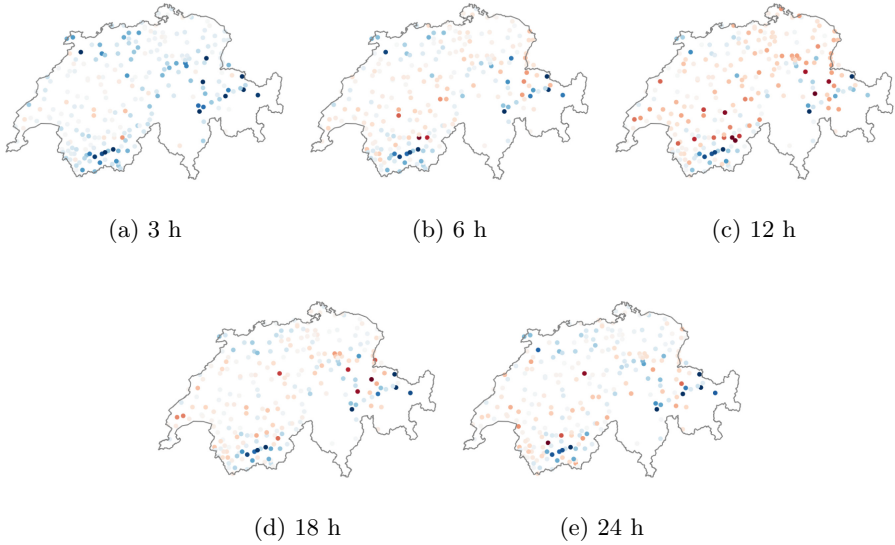


Fig. 3. Geographic error distribution: Each dot corresponds to one station. Red indicates station where model performed worse than the benchmark model while blue dots indicate stations where the model was better. Darker shades indicate larger absolute differences. (Color figure online)

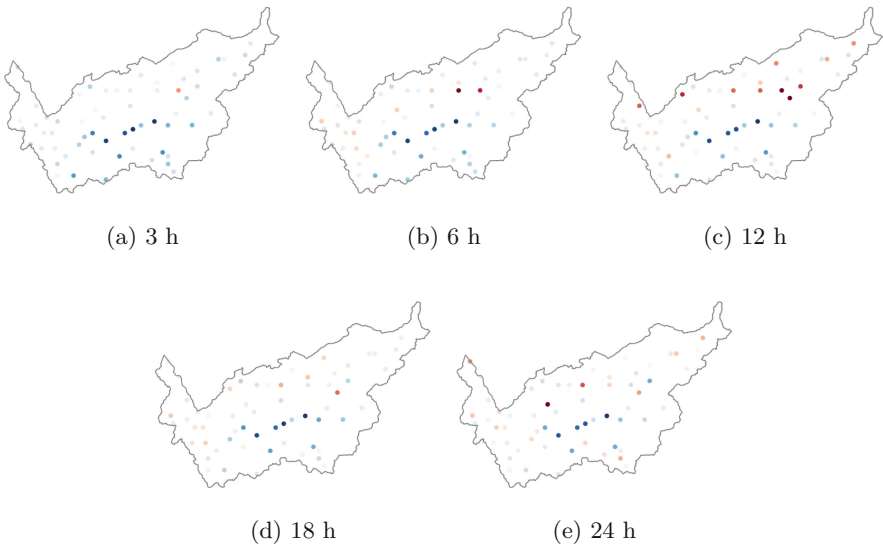


Fig. 4. Geographic error distribution in the canton of Valais. Coloring analogous to Fig. 3.

Table 2. FH: forecast horizon (h), R: Ratio of stations where the model is better than benchmark model, MPB: Mean difference of MSE of the predictions model vs. benchmark model, MEPB: median difference of MSE of the predictions model vs. benchmark model.

FH	MPB	R	MEPB	FH	MPB	R	MEPB
3	-2.01	0.81	-1.06	3	-2.37	0.89	-1.54
6	-0.54	0.45	0.20	6	-1.09	0.61	-0.64
12	0.73	0.36	0.70	12	0.46	0.56	-0.22
18	-0.30	0.47	0.10	18	-0.99	0.54	-0.34
24	-0.54	0.51	-0.02	24	-0.43	0.51	-0.02

(a) Switzerland

(b) Valais

Figure 5 shows the distribution of the difference $MSE(model) - MSE(benchmark)$ for all stations and for the stations in the Valais. Furthermore, the colors indicate the mean error of the station over all forecast horizons. We can easily recognize that stations either perform consistently bad or well over all forecast horizons on both geographic scales. That is, if a station benefits from the model forecasts for any forecast horizon, it is likely to benefit for the other forecast horizons as well. This seems to support the thesis that ML boosted models can improve forecast quality in difficult to model locations while other locations might not benefit from our approach.

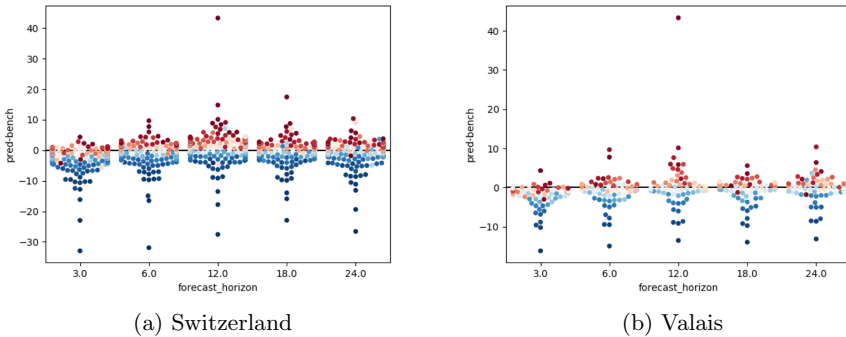


Fig. 5. MSE (model) and MSE (benchmark) distribution per forecast. Colors indicate the mean difference over all forecast horizons. Note that these colors do not correspond to the colorscale used on the map visualizations.

6 Discussion

We have developed in this contribution an approach that combines numerical weather prediction (NWP) models with a machine learning (ML) approach.

Specifically, we developed a deep learning (DL) model to predict air temperature 2 m above ground that is trained both on NWM models and local weather data and evaluated the approach for 249 weather station sites in Switzerland. Our preliminary results show that the approach has potential: in the nowcasting domain, i.e., for short time-scales, the model performs better almost everywhere, for longer forecast horizons it seems that the approach could bring improvements for some but not all regions. A new task may therefore be to identify the locations that could benefit from our approach, e.g., a classifier based on geographic features might come into play.

We currently interpret the results as shown on the map (Fig. 3) as follows: in mountainous regions such as the Valais, the potential for improvement seems to be highest, because there you might find yourself in a special micro weather situation, possibly created by the mountains that shield the region from the coarse-meshed macro weather situation simulated by the NWP models. However, this hypothesis should be examined more closely. Unfortunately, we have not collected forecast data for all mountain valley regions in Switzerland such as the Engadin.

We have not yet systematically analyzed the DL model in terms of architecture and parameter tuning. Therefore we think that with further experiments and analyses of the model substantial improvements are still possible.

In future work we will work further on the following approaches: 1. To forecast further parameters, for example to predict precipitation, 2. To use data from neighboring stations to forecast the weather at a particular station and 3. Collect more data, e.g. provided by small weather stations at local farmers, and develop the model further.

We believe that the blend of NWP models and ML models has great potential and will continue to find its way into the science of weather forecasting.

Acknowledgment. This work was supported by Innosuisse grant 26301.1 IP-ICT and Hydrolina Sarl.

References

1. Richardson, L.F.: *Weather Prediction by Numerical Process*. Cambridge University Press, New York (2007)
2. Kalnay, E.: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, New York (2003)
3. Bauer, P., Thorpe, A., Brunet, G.: The quiet revolution of numerical weather prediction. *Nature* **525**(7567), 47–55 (2015)
4. Grover, A., Kapoor, A., Horvitz, E.: A deep hybrid model for weather forecasting, In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 379–386 (2015)
5. Weyn, J.A., Durran, D.R., Caruana, R.: Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *J. Adv. Model. Earth Syst.* **11**(8), 2680–2693 (2019)

6. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science. *Nature* **566**(7743), 195–204 (2019)
7. Xingjian, S.H.I., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: *Advances in Neural Information Processing Systems*, pp. 802–810 (2015)
8. Hernández, E., Sanchez-Anguix, V., Julian, V., Palanca, J., Duque, N.: Rainfall prediction: a deep learning approach. In: Martínez-Álvarez, F., Troncoso, A., Quintián, H., Corchado, E. (eds.) *HAIS 2016. LNCS (LNAI)*, vol. 9648, pp. 151–162. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32034-2_13