GNNs for Global Weather Prediction

Cristi Blaga Robin Chan Gabriel Gavrilas Christopher Vogelsanger

{cblaga, chanr, ggabriel, cvogelsa}@student.ethz.ch

January 12, 2022

Abstract

Weather prediction is a domain mostly untouched by the advances in machine learning and data driven predictions. Based on a recently published benchmark, WeatherBench, several attempts have been made for fully data-driven global weather prediction. Despite the graph nature of global weather data, graph-based approaches, that model inherently non-uniform special dependencies using graph-based representations, have not yet been applied to weather forecasting. In this work, we present the feasibility of graph neural networks for global weather forecasting by predicting the temperature, which can easily be extended to any characteristic atmospheric variable. We compare multiple approaches for modelling the temporal dependencies and GNN architectures and assess them on baselines of the WeatherBench dataset. Even though simple GCN-architectures on simple graph representations struggle to capture weather dynamics fully, modern state-of-the-art GNN architectures can be adapted to achieve improved results over such baselines even with limited computational resources on a global scale.

1 Introduction

Despite the success of deep learning methods in problems spanning across various domains, rigorous studies of their application in weather prediction are scarce. [1, 2] Presently, purely physical, numerical models are used for state-of-the-art weather prediction. Since current numerical weather prediction (NWP) requires building large ensembles of models solving governing physical equations on a global grid, their use is limited not only due to the large computational cost, but occasionally also systematically, as parametrizations in physical models may not capture the complexity of specific climatological phenomena [3, 4]. Data-driven, deep learning methods might inherently address the aforementioned shortcomings due to their proven ability to capture high-dimensional dependencies in such complex, non-linear systems at reasonable computational cost [5].

In current weather prediction pipelines, deep learning methods are commonly used for post-processing the numerical simulation ensembles [6, 7]. In recent years however, several studies attempt to show their feasibility for full medium-range predictions, i.e. predicting specific climatological properties (i.e. individual atmospheric variables) globally with a lead time of a few hours based only on past reanalysis data gathered from satellite observations. Among the proposed prediction models were encoder-decoder CNNs [8, 9, 10], which were later refined by using a U-Net architecture [11] and ResNets pre-trained on climate simulations [12]. However, models were trained on different data of varying granularity, making results hard to compare. To improve comparability between such studies, the WeatherBench dataset was recently published, alongside baselines from purely data-driven as well as numerical approaches [13]. WeatherBench will act as the main point of reference to assess the feasibility of our proposed methods.

The data provided by WeatherBench is organized as distinct measurement points on a global grid, making the interdependence inherently non-uniform due to local topological irregularities and the non-linear projection from the latitude-longitude geographic coordinate system onto a plane. In this work, we study the use of graph representations to better encode the spatial dependencies between measurement points of weather data. We show the general feasibility of various GNN architectures for modelling the spatio-temporal dynamics and compare their performance with the WeatherBench baselines.

In sections 2.1 - 2.3 we state the problem, define our dataset and the necessary computational resources. In section 2.4 we present the four implemented graph-based models. In section 3 we present the results of our experiments of which a discussion can be found in section 4.

2 Models and Methods

2.1 Preliminary Problem Statement

Weather forecasting is a time-series regression problem, which can be understood as finding the most likely atmospheric property measurement $\hat{v}_{t+H}$ for a lead time of H, given the previous M observations:

$$\hat{v}_{t+H} = \arg\max_{v_{t+H}} \log P\left(v_{t+H} | v_{t-M+1}, ..., v_t\right)$$

Using graph data representations, a variable measurement v_t at time step t can then be represented as an undirected graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}, W)$, where vertices $\mathcal{V}_t$ each store values at a measurement location, edges $\mathcal{E}$ indicate some notion of connectivity between nodes and W represent the edge weights. More specifically, weather prediction then becomes a node prediction task for each vertex in the measurement grid graph.

2.2 Dataset

The WeatherBench dataset [13] includes several regridded versions of the ERA5 reanalysis dataset [14] to lower res-

olutions. WeatherBench contains 8 climatological variables, such as temperature, pressure and wind speeds, on 13 vertical pressure levels as well as 6 single-level variables measured hourly over the course of 40 years. Due to GPU memory constraints, I/O speed, and comparability with other studies, we used the coarsest dataset. It has a resolution of 5.625° in its latitude and longitude therefore yielding a 32×64 spatial resolution.

As the task of weather prediction is not defined precisely, we established a set of parameters that should guide our training procedure. Predictions were made for the `T850` variable (i.e. the temperature at the 850hPa level) with a lead time of 3 days based on its values in the previous 12 hours. The training dataset consisted of the measurements of the temperature variable at the 850hPa level between 1979-2016. The test years were set to 2017-2018 in accordance with precious studies. This configuration was decided upon to maximize comparability with previous studies as well as to limit computational cost as much as possible.

2.3 Computational Requirements

In order to train the models and perform hyperparameter tuning on the proposed architectures and pipelines of this study, we used multiple-CPU computation nodes with sufficient RAM to load the subset of the dataset used for training or a CPU-computation node with a GPU, again with enough RAM and VRAM. To store the subset of the dataset, physical storage with around 20GB of memory was required. Computations were done on the Euler cluster of ETH Zurich and on private sessions on the Google Colab platform.

2.4 Models

Four different models alongside with their own graph representations were tried out to try to capture the spatio-temporal dynamics of the `T850` variable.

2.4.1 Mixed Spatio-Temporal GAT Networks

As a feasibility assessment for the usage of graphs for weather prediction, a simple baseline was implemented, which processes temporal and spatial dependencies equally. For each prediction, a graph is created, where each measurement point is a node adjacent to its eight nearest neighbours spatially, together with itself in the previous and subsequent time step, as shown simplified in figure 1.

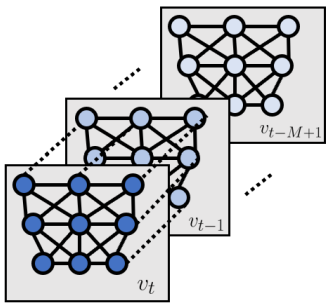

Figure 1: Mixed Spatio-Temporal Graph Representation

This graph is then passed through a set of residual graph-attention network blocks consisting of 2-headed GAT [15], BatchNorm and ELU activation, as shown in figure 2. Finally, the intermediary output is flattened and passed through a linear layer in order to make the prediction of the desired length.

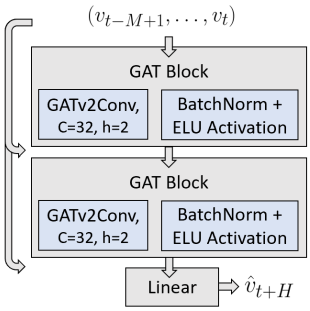

Figure 2: Mixed Spatio-Temporal GAT Model

2.4.2 GCN with Motion Attention

Based on the work by Mao et al. [16], where a GCN based approach for human motion prediction using motion attention was proposed, we have adapted the corresponding model in order to leverage the inherent graph structure of the data.

The main contribution of the original paper was the temporal encoding of the input sequence as a sequence of small windows transformed into spectral space to capture motion, upon which the attention values were computed. These attention values were then concatenated to a specially created input sequence and passed on to a GCN working on a fully connected graph over the input nodes in order to create the prediction sequence.

A graphical depiction of the model adapted to our input format is presented in Figure 3. Direct application of the proposed model to our data would induce a much larger graph, leading to performance reduction and long training times. To be able to efficiently run our model, we add linear encoding and decoding layers before and after the model, to reduce the number of nodes. This reduces training time significantly and also improves the performance of the model on our data.

2.4.3 3D WaveNet with Spatial Convolutions

Another approach we have built has been inspired by the WaveNet[17] approach of modelling and predicting time series data. In the standard WaveNet approach, only 1D time series are considered. However, this was not the case for us, since we had a time series for each of the $H \times W$ measurement points. Therefore, we needed to adapt the original approach to accommodate the increase in dimensionality in our use case.

Additionally, we allow message passing between adjacent measurement points based on the locality of the points. In the first and most straightforward approach, we do this using standard convolutions. Therefore, the model uses 3D convolutions, out of which two of the convoluted dimensions are spatial and the third one is temporal.

Finally, compared to the standard WaveNet, we have made the design choice of not computing intermediate time steps, but rather directly the wanted time step.

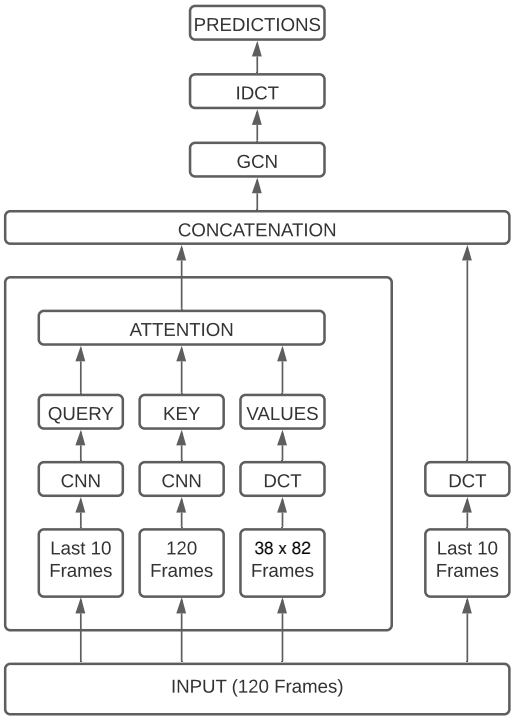

Figure 3: Architecture of the Motion Attention approach

2.4.4 3D WaveNet with Spatial Graph Convolutions

Building upon the previous model, we have also developed the "3D WaveNet with spatial graph convolutions" where we have replaced the message passing using standard convolutions with graph convolutions. By doing this, we take advantage of the inherent information present in the locality of the measurement points on the surface of Earth.

To this extent, the graph we use is a standard grid-mapping of the planet, where each measurement point is connected to all of its 8 neighbours. Moreover, even if we still use 3D convolutions for easiness of the implementation, the 2 dimensions representing spatiality have a kernel size of 1, whereas the actual focus is on the temporal convolution which is performed in this manner.

3 Results

If not stated differently, the models are trained for 10 epochs on the training dataset as described in section 2.2 for a lead-time of 3 days. The root-mean-squared test errors (RMSE) are shown for the listed given baselines from WeatherBench dataset alongside the results from the GNN approaches in Table 1. A summary of the WeatherBench baselines can be seen below.

WeatherBench Baselines

- Persistence: Use of the input image as output
- Climatology: The mean values over the entire training dataset
- Weekly Climatology: Mean computed separately for each calendar week of the year
- Linear regression: A simple linear regression model trained for direct and iterative prediction
- IFS: Integrated Forecast System of the European Center for Medium-range Weather Forecasting regridded to $5.625°$. IFS T42 and T65 are NWP models run at coarser resolutions.

Further, we also create plots of the predicted temperature, which are shown in Figure 4. We aim to use the shown temperature distributions to assess model shortcomings more concretely.

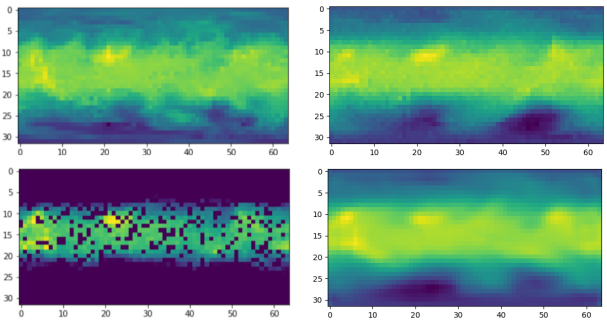

Figure 4: Output Prediction of GNN Models: Ground Truth Temperature Distribution (top left), GCN with Motion Attention (top right), Baseline STGAT model (bottom left), GraphWavenet (bottom right)

Model	T850 RMSE 3 days [K]
Climatology	5.51
Persistence	4.23
Weekly climatology	3.50
IFS T42	3.09
IFS T63	1.85
Rasp and Thuerey 2020 [12]	1.65
Operational IFS	1.36
Mixed STGATs	9.40
3D WaveNet + spatial graph convolutions	3.60
GCN + Motion Attention	3.27

Table 1: T850 RMSE in Kelvin for 3-day lead-time

4 Discussion

Our proposed models were able to beat simple baselines, even approaching the performance of a real physically modelled weather prediction (IFS T42) computed on a similar coarse sampled grid.

The baseline mixed STGAT model performed the worst and from figure 4 we see that it was able to capture the rough outline of the prediction, such as its general distribution and temperature hot spots, however the many learned irregularities in the grid show that the smoothing through information propagation through the edges was not sufficient. This could be addressed by including more GAT message passing layers. Or more generally, a separate, more sophisticated method is needed to capture the temporal relations between measurements, as was done in other approaches.

The Wavenet models are able to predict hot-spots correctly for temperature but fail to model more intricate details. We hypothesize that the WaveNet models primarily remember which areas on the earth are warmest at which point in the day and might derive and predict cloudiness. This would explain why the hot spots in the equator region are rather exact, but the predictions get worse the further away we move from the equator.

The motion attention approach was generally able to predict more details. Because the approach inherently models sequences of frames, we were able to investigate the prediction sequence up to three days. The model accurately captures the dominant heat distributions but also manages to model daily fluctuations and other temporal features of the data, yielding better overall results.

However, the proposed GNN models still lack in comparison to other neural network approaches, some of which have considerably less complexity. This fact stands in contrast to the perceived benefits of using graph-based approaches for global weather prediction.

To this extent, in the following, we discuss different reasons for the inferior performance of our models.

Limited Computational Resources

The largest inhibition to our models stem from the limited computational resources. The WeatherBench dataset offers a rich set of features to train on. Overall there are 40 years worth of data recorded at an 1 hour interval. For each interval there were 110 features recorded and an additional 5 constant features. Even at the roughest spatial sampling of 32×64 cells the size of the data still amounts to over 300 GB of data. Designing an efficient training procedure with this amount of data posed a serious problem to us, the main limiting factor being the VRAM of the GPU. Consequently, we had to apply significant reductions to the data to be able to train our models efficiently.

Therefore, regarding the data, we have resorted to the coarsest spatial sampling of the provided data. Moreover, we have only used a single feature (temperature at 850hPa) for both training and prediction.

On top of this, we have also significantly reduced the size of our models in order to be able to run them in a feasible time on our available GPUs. The most important hyperparameters that we have tuned down severely were the structural ones.

Moreover, other training hyperparameters, such as the batch size, have also needed to be reduced. It is important to note here that even if we could get around this type of problem by performing gradient accumulation, reducing the batch size also affects the valuable contribution of the Batch Normalization layer for example.

All in all, we strongly consider that these reductions have had a significantly negative effect on the performance of our model.

Furthermore, even after all of these reductions, our models still required several of hours of training to achieve good results using the configuration we described in earlier sections. This was also influenced mainly by the large amount of time steps available for training. However, using less time steps proved to reduce the performance of our models drastically, forcing us to favour number of time-steps over epochs.

Dataset Limitations

The richness of the dataset naturally calls for the exploration and study of the given features. It would be worthwhile to study the use of different features and their nature as well as the general nature of weather data such as yearly fluctuations, repetitions, and the interdependence of height levels and other inherent features of weather data. Moreover, the use of multiple features, rather than just one, in the training data could also prove to be extremely useful in order to leverage models to discover underlying patterns in the weather behaviour. Unfortunately, such a study would have exceeded the scope of this project and would have been unfeasible to be performed with respect to the computational resources we had at hand.

Finally, we see greater potential of graph-based representations if climate variable measurements are not interpolated and gridded, but measured at specific weather stations, distributed heterogeneously in different environments, such as in MeteoSwiss data. [18]

5 Summary

Despite limited computational resources, our proposed GNN architectures were able to outperform some rather simple WeatherBench baselines. This shows the general feasibility of applying graph-based representations to model. Separately modeling temporal and spatial dynamics can further improve their performance as seen in the Wavenet and motion attention approaches. We further suspect that using graph representations are most likely an over-representation of the current evenly gridded single-feature data, but could yield benefits if data is captured more heterogeneously with more features and relational information.Finally, if more computational resources are available, more features could be included in training and prediction to incorporate the correlation of other climatological variables with the predicted temperature.

References

[1] S. Scher. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, 45(22):12,616–12,622, 2018. doi: https://doi.org/10.1029/2018GL080704. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL080704.

[2] M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200097, 2021. doi: 10.1098/rsta.2020.0097. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0097.

[3] E. Kalnay. *Athmospheric modeling, data assimilation, and predictability*, volume 54. Cambridge University Press, 2003.

[4] Peter Vogel, Peter Knippertz, Andreas H. Fink, Andreas Schlueter, and Tilmann Gneiting. Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical africa. *Weather and Forecasting*, 33(2):369 – 388, 2018. doi: 10.1175/WAF-D-17-0127.1. URL https://journals.ametsoc.org/view/journals/wefo/33/2/waf-d-17-0127_1.xml.

[5] P. D. Dueben and P. Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009, 2018. doi: 10.5194/gmd-11-3999-2018. URL https://gmd.copernicus.org/articles/11/3999/2018/.

[6] Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, Oct 2018. ISSN 1520-0493. doi: 10.1175/mwr-d-18-0187.1. URL http://dx.doi.org/10.1175/MWR-D-18-0187.1.

[7] Peter Grönquist, Chengyuan Yao, Tal Ben-Nun, Nikoli Dryden, Peter Dueben, Shigang Li, and Torsten Hoefler. Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200092, Feb 2021. ISSN 1471-2962. doi: 10.1098/rsta.2020.0092. URL http://dx.doi.org/10.1098/rsta.2020.0092.

[8] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Can machines learn to predict weather? using deep learning to predict gridded 500-hpa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, 11(8):2680–2693, 2019. doi: https://doi.org/10.1029/2019MS001705. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001705.

[9] S. Scher. Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters*, 45(22):12,616–12,622, 2018. doi: https://doi.org/10.1029/2018GL080704. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL080704.

[10] Sebastian Scher and Gabriele Messori. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2830–2841, 2018. doi: https://doi.org/10.1002/qj.3410. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3410.

[11] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, 2020. doi: https://doi.org/10.1029/2020MS002109. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002109.

[12] Stephan Rasp and Nils Thuerey. Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13(2), Feb 2021. ISSN 1942-2466. doi: 10.1029/2020ms002405. URL http://dx.doi.org/10.1029/2020MS002405.

[13] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), Nov 2020. ISSN 1942-2466. doi: 10.1029/2020ms002203. URL http://dx.doi.org/10.1029/2020MS002203.

[14] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

[15] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks?, 2021.

[16] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. *CoRR*, abs/2007.11755, 2020. URL https://arxiv.org/abs/2007.11755.

[17] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.

[18] Meteoswiss: Datenportal für lehre und forschung. https://www.meteoschweiz.admin.ch/home/service-und-publikationen/beratung-und-service/datenportal-fuer-lehre-und-forschung.html. Accessed: 12.01.2021.