Applying AtmoRep for Diverse Weather Applications

Ankit Patnala[1], **Belkis Asma Semcheddine**[1],
Michael Langguth[1], **Martin G. Schultz**[1,2], **Christian Lessig**[3], **and Ilaria Luise**[4]

[1] Jülich Supercomputing Centre, Forschungszentrum Jülich, 52428 Jülich, Germany
E-mail: {a.patnala, a.semcheddine m.langguth m.schultz}@fz-juelich.de

[2] University of Cologne, Department of Computer Science, Cologne, Germany

[3] European Centre for Medium-Range Weather Forecasts, 53175 Bonn, Germany
E-mail: christian.lessig@ecmwf.int

[4] CERN, Geneva, Switzerland
E-mail: ilaria.luise@cern.ch

Machine learning has recently seen a rapid wide-spread adoption across various fields of science including atmospheric and weather research. The emergence of foundation models has marked a transformation in the science of machine learning. These foundation models are general-purpose models trained on huge amounts of data using self-supervised methods, eliminating the need for labeled data. Once trained, the parameters of these models can be utilized as a starting point for a range of domain-specific tasks. This approach is advantageous in terms of both cost and performance, as it minimizes the reliance on annotated data compared to models trained from scratch. Motivated by this, our study explores the foundational capabilities of AtmoRep, a stochastic atmospheric foundation model, for two distinct weather-related applications, data compression and statistical downscaling. The training of the 3.5 billion parameter AtmoRep model consumed about a few weeks of compute time on 32 JUWELS Booster nodes.

1 Introduction

Local weather is characterized by atmospheric variables such as temperature, specific humidity, and wind speed at a given location, time, and altitude. In meteorological studies, weather typically refers to time scales ranging from hours to several days[1]. Accurate weather predictions are crucial for mitigating severe weather impacts like high winds and flooding[2] and they are relevant for many planning purposes. Understanding weather patterns requires studying complex interactions among atmospheric variables. The physical laws describing these interactions are primarily derived from fluid dynamics and radiative transfer. They are governed by conservation laws of mass, momentum, and energy[3,4]. Numerical Weather Prediction (NWP) models forecast intricate weather patterns[5,6], utilizing preprocessed observational data to estimate the initial conditions[7]. The NWP models employ discretization in space and time with current operational models typically achieving resolutions of around 10 km in longitude and latitude for global forecasts. The output from NWP models is often post-processed with statistical tools, for example, to achieve bias correction and to further increase the spatial resolution with statistical downscaling[8]. Despite continuous improvements over decades and generally good predictive skills, NWP models suffer from inherent biases, limited spatial resolution, and structural errors[9], along with high computational costs.

Recently, advanced machine learning (ML) models have transformed weather forecasting. These AI-driven approaches have emerged as strong competitors to traditional NWP models, offering better predictions at a fraction of the computational cost[10,11]. Although purely data-driven and lacking explicit physics information, these models effectively capture complex interactions among atmospheric state variables and their spatio-temporal patterns[12]. ML models also offer enhanced flexibility and can be trained to directly predict the atmospheric state several hours into the future, unlike NWP models, which are constrained by the Courant-Friedrichs-Lewy (CFL) condition[14]. Additionally, advanced ML models can exploit the added value from multiple datasets with varying resolutions and they are able to provide efficient ensemble predictions, thus offering confidence intervals for understanding forecasting uncertainty[16,17].

The emergence of foundation models has enabled a new revolution in machine learning. These models are trained on vast datasets using unsupervised and self-supervised techniques, allowing adaptation for various tasks with minimal additional training. Foundation models are also making their way into the field of weather forecasting; one such model is AtmoRep[17]. The training on a large subset of data from the 5th European reanalysis (ERA5[24]) enables AtmoRep to learn comprehensive representations of atmospheric dynamics. The pretrained AtmoRep exhibits skilful capabilities for various tasks such as forecasting, temporal interpolation and counterfactuals. Through fine-tuning, the performance of AtmoRep can be further improved achieving state-of-the-art results (e.g. forecasting) or applied to other downstream tasks (e.g. statistical downscaling). In this paper, we explore the capabilities of AtmoRep for two downstream tasks: data compression and downscaling for 2m temperature.

In the following, we first provide an overview of the core AtmoRep model, focusing on the processing pipeline of the atmospheric variables and the employed training methodology. We then describe the two downstream tasks utilizing AtmoRep and discuss the results from initial sets of experiments. At the end, we conclude with a summary of our findings and future research directions.

2 The AtmoRep Model

AtmoRep[17] is a stochastic, generative neural network model for atmospheric dynamics, utilizing large-scale representation learning to identify patterns within the high-dimensional state space of atmospheric data. The inherently stochastic nature of the model is crucial to capture the inherent statistical nature of atmospheric dynamics. The model has been trained with ERA5 reanalysis data from 1980 to 2017 and evaluated on data of the year 2018, similar to other ML studies on weather forecasting. The architecture of AtmoRep is inspired by established transformer models[18] and Vision Transformer (ViT)[21], which have demonstrated remarkable success in natural language processing and computer vision, respectively. AtmoRep's training strategy has been adopted from BERT (Bidirectional Encoder Representations from Transformers,[20]). The model can be flexibly configured with respect to the variables and vertical levels.

The flexibility with respect to the variables is achieved through a two-step training process: In a first step, independent transformer models, termed *singleformers*, are trained separately for each atmospheric variable. In a second step, these per-variable transformers are combined with cross-attention heads added to the encoder to enable interaction between variables in the resulting multi-variable transformer model (termed *multiformer*). This approach proves efficient, since it significantly reduces the training time needed for a high-performing AtmoRep model compared to training a multi-variable model from scratch.

Various pre-trained configurations singleformers and multiformers are publicly available from `https://datapub.fz-juelich.de/atmorep/trained-models.html`. All the available models were trained on 5 model levels (96, 105, 114, 123, 137), ranging from the Earth's surface to about 5 km altitude. The downstream applications discussed in this work employ the `singleformer-t` configuration for temperature and the `multi3-uv` configuration trained on temperature and wind vector components.

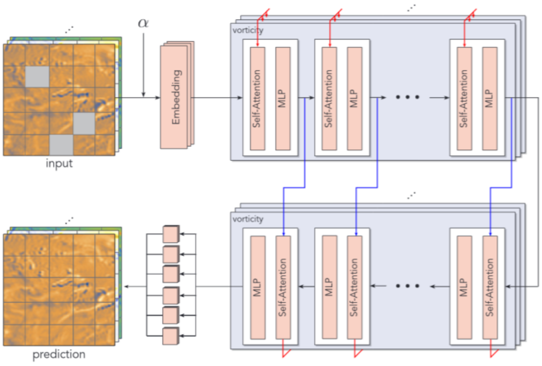

Figure 1. Schematic of AtmoRep's model architecture and training scheme[17]

Figure 1 illustrates the AtmoRep model architecture. Transformer models implement three main concepts: tokenization, embedding, and attention mapping. In AtmoRep, the tokenization process consists of dividing the randomly selected subset of gridded ERA5 data (36 × 54 × 108 in $time \times lat \times lon$ dimensions, respectively) into several 3-dimensional cubes known as patches or tokens. The standard configuration of these tokens are variable-dependent. In general, variables with higher modes of variability (e.g., vorticity and divergence) are cut into smaller tokens than variables with less high-frequency variations (e.g., temperature). For temperature, the standard token size is 3 × 27 × 27.

Subsequently, the tokens are embedded into high-dimensional vectors. Because the attention mechanism is position-independent, relative positional encoding is added to the tokens. Furthermore, latitude, longitude, model level, year, day-of-year, and time-of-day are added as auxiliary information to encode external forcings such as the seasonal cycle that is determined by the planetary motion of the Earth. The combined embeddings of positional encoding and tokens are subsequently processed by the attention blocks of the encoder network in the AtmoRep model. Self-attention is used to identify relations between patches of one variable, while cross-attention emphasizes correlations across variables. The output from the encoder encompasses an abstract, feature-rich representation of atmospheric dynamics. The purpose of the decoder is then to reconstruct physical fields based on this abstract representation. The final network layer consists of a tail network with multiple prediction heads that draw individual samples from the learned probability distribution of the atmospheric state.

To train the AtmoRep model, the principles of the BERT[20] protocol are adopted. In this framework, some tokens are randomly masked or modified during training. The model then learns to reconstruct the masked tokens based on contextual information provided by unmasked tokens. AtmoRep's training protocol is formulated as $p_\theta(y|x, \alpha)$ where x refers to the masked weather data, α refers to the auxiliary information, and y refers to the reconstructed tokens. The loss function employed to optimize the model's parameters combines Mean Squared Error (MSE) loss with a novel statistical loss that takes into account the first two statistical moments of the predicted ensemble. We refer to the original AtmoRep paper[17] for more details about the model architecture and the training process.

When the pre-trained model is applied to weather-related tasks without further fine-tuning, this is called zero-shot inference. In AtmoRep[17] zero-shot performance is evaluated for forecasting, bias correction, data interpolation, and counterfactual experiments. Here, we add results from the data (de)compression task and provide an update on 2m temperature downscaling. For the latter, the AtmoRep core model is extended with a downscaling tail network to account for the increased output dimension. In contrast to zero-shot applications, this extension requires fine-tuning of the task-specific AtmoRep model application.

3 Downstream Tasks

3.1 Data Compression and Reconstruction

The output of climate model simulations has been growing substantially due to increased model resolution and the increased demand for detailed and high-frequency output of a comprehensive set of variables[25,26]. The storage of climate model data is therefore becoming a fundamental bottleneck limiting the possible applications of climate simulations. Data compression is one way to potentially alleviate this issue. Here, we explore how we can use the rich representation of atmospheric dynamics learned by AtmoRep to reconstruct climate data from subsets of the original fields. In principle, AtmoRep should allow for the faithful reconstruction of variables even when large portions of the data are missing, since the model was trained with randomly masked data. In this section, we investigate how well AtmoRep can reconstruct data when certain systematic masking patterns are applied.

Figure 2 illustrates different masking patterns we employed to assess the reconstruction capabilities of AtmoRep. Our tests were constructed to assess the reconstruction quality along individual dimensions, whereas longitude and latitude were combined into a "geographic" masking pattern. The compression ratios varied from 1.42 to 4 (see Table 1), which means that up to 75% of the data is being omitted (i.e. masked). It should be emphasized that we tested the data reconstruction in a zero-shot setting, i.e. using the pre-trained `singleformer-t` AtmoRep configuration without any fine-tuning.

The space-time tokenization was set to 3 × 27 × 27, and the neighborhood was selected for each batch as 12 × 2 × 4. The masking patterns applied are summarized in Table 1. For every configuration, we randomly sampled 100 days from the test year 2021 (starting from December 2020). In the first configuration (A), a "checkerboard pattern" was applied at each model level and for all time steps: every second token in longitude and latitude dimension was masked resulting in a compression ratio of 2. The resulting reconstructions look physically plausible. The mean root mean square error (RMSE) ranges from

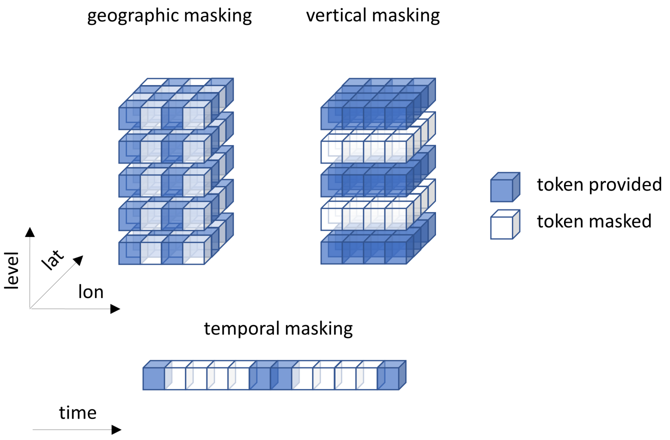

Figure 2. Illustration of masking patterns applied in the data reconstruction/decompression task

0.45 to 0.85 K across the vertical levels, with smaller errors near the surface (see Table 1). Configuration B explored temporal masking. In configuration B1, every second time step was masked, while configuration B2 explored a higher compression ratio of 3. Ablation studies on temporal masking indicated that the reconstruction results are best if data points at the beginning and end and at the center of the 36-hour time window are retained. This is the motivation for applying the pattern depicted at the bottom of Figure 2. Compared to geographic masking the reconstruction accuracy of temporal masking is slightly better. It is interesting to observe that a larger compression ratio has little influence on the reconstruction error near the surface and actually leads to smaller errors at higher model levels. The reasons for this behaviour are not fully understood. In configuration AB we combine geographic and temporal masking, thus achieving a compression ratio of 4. The results show a slight improvement in RMSE values.

run	compression ratio	level-wise RMSE				
A	2	0.4451	0.4878	0.6748	0.7374	0.8465
B1	2	0.3765	0.4117	0.5514	0.6568	1.2300
B2	4	0.4568	0.4726	0.6119	0.6551	0.9225
AB	4	0.4009	0.4318	0.5939	0.6379	0.7287
C1	2.5	NA	1.6845	NA	1.4229	NA
C2	1.42	NA	0.3996	NA	0.4506	NA

Table 1. Configuration and accuracy (RMSE) of the data reconstruction experiments. For explanations of the masking patterns, see Figure 2 and text. The compression ratio is defined as the ratio of available input tokens to the full number of tokens of the reconstructed field. Further information on the masking patterns is given in the text. RMSE values are given per model level with level indices (from left to right) 137, 123, 114, 105, and 96.

The third set of configurations explores vertical masking, i.e. leaving out data from specific vertical levels and asking the model to interpolate vertically. We found that masking entire levels of temperature data resulted in poor reconstruction accuracy (configuration C1 in Table 1). Therefore, we tested a second variant where we applied the temporal masking pattern on the intermediate levels, while leaving the other levels complete (configuration C2). As expected, the results with C2 are significantly better than with C1. However, the added value compared to the geographical and temporal masking patterns (configurations A to AB) is small, especially in light of the much smaller deployed compression ratio (more input information available).

As shown above, the data reconstruction is in principle possible, but further work is needed to achieve the desired level of accuracy (e.g., $RMSE < 0.1\,\mathrm{K}$) and computational performance (e.g., reconstruction time $< 1\mathrm{s}$). In any case, the experiments revealed interesting aspects of the model behaviour. Provided that it is possible to solve the issues described above, this novel data compression approach offers a lot of potential, because it would enable very high compression ratios (up to 100 or more) with relatively little dependence of the reconstruction quality on the masking ratio (since most of the information is stored in the model weights). We anticipate that proper fine-tuning and the use of multivariate information will further improve the results.

3.2 Statistical Downscaling

Localized and regional meteorological data is highly relevant for society, agriculture, and several industrial sectors, such as renewable energies. This particularly holds for regions with complex terrain which introduces significant spatial variability in key meteorological variables such as precipitation, wind speed, or the near surface temperature. The ERA5 reanalysis, which has been utilized to train AtmoRep, operates at a resolution of $\Delta x_{ERA5} \simeq 30\,\mathrm{km}$, which is clearly insufficient to reproduce orographic features. While ERA5 provides a comprehensive estimate of the atmospheric state[24], it has well-documented limitations in mountainous regions, such as the Alpine region in Central Europe. Even though there are ongoing efforts to generate meteorological data on the scale of $1 - 2\,\mathrm{km}$ with numerical models, these constitute a major computational challenge. Therefore, several weather centers developed statistical models to create higher-resolution information from coarser-resolution model output. ML models can be applied to this task with great efficiency and equal to better quality.

To demonstrate AtmoRep's adaptability for downstream applications, we applied it to perform statistical downscaling of 2m temperature (T2m) data to a resolution of approximately 6 km. For this purpose, we selected the COSMO REA6 reanalysis[27] as target dataset. COSMO REA6 provides much more accurate information than ERA5, especially over the Alpine region.[23] While the downscaling application has already been introduced in AtmoRep[17], we extend this analysis to further demonstrate the model's effectiveness for this task. This includes a more detailed analysis of spatial error patterns and of the spatial variability in the downscaled T2m field. To substantiate our findings, we compare AtmoRep's performance with an Wasserstein Generative Adversarial Network (WGAN[28]), offering a more advanced benchmark than previously used in AtmoRep[17].

The downscaling application utilizes the `multi3-uv` configuration of AtmoRep which has been pre-trained on temperature and the horizontal wind components. Note that

AtmoRep does not require input of high-resolution topography as many other downscaling models; it can extract the high-resolution features from the dynamic variables alone. For the downscaling task, we extended the core model with a tail network of 6 transformer blocks that is connected to the last transformer block of AtmoRep's decoder. Each block comprises a self-attention layer with 16 attention heads and and a multilayer perceptron with two layers. To achieve the desired resolution of the data, the output token size of the downscaling network is enhanced by a factor of 4. The increased token size necessitates an increased embedding dimension for the temperature data achieved with a linear layer at the beginning of the downscaling network. Accordingly, the local position encoding is updated. Again, an ensemble tail is deployed to provide a probablistic downscaling output. However, a small ensemble member size of 4 was chosen due to computational constraints. During fine-tuning, the network parameters of both the core model and the tail network were optimized, resulting in about 1.85B trainable parameters. For optimal hardware utilization, we employed both data and model parallelism. The downscaling network has been trained for three days on 8 nodes on JUWELS Booster.

Figure 3 showcases a sample from the test year 2018, demonstrating that the downscaling not only generates super resolution output, but also achieves a bias correction of the input data.

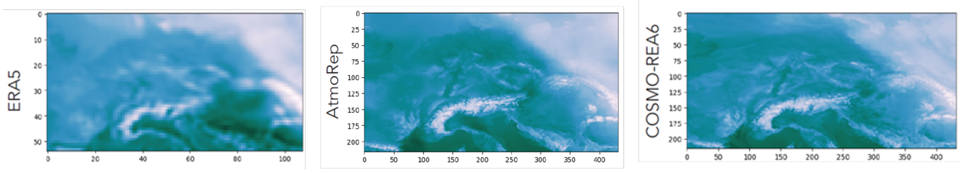

Figure 3. Downscaling sample from 2018 with an air mass boundary (AMB) in the north-eastern part of the domain. The AMB in ERA5 (left) is located further north-east compared to COSMO REA6 (right). The AtmoRep downscaling result (center) demonstrates that the location of the AMB is corrected towards the ground truth data.

To assess the potential benefit of using AtmoRep for downscaling, we compare our results with those from a WGAN. The WGAN utilizes a U-Net generator with 3.5 million trainable parameters that has been adopted from the 2 m temperature downscaling study of[29] and a convolutional critic network with 1.5 M trainable parameters. In analogy to AtmoRep, the generator is informed with temperature and wind information from several model levels. Additionally, it also inputs coarse- and high-resolved surface topography data to support the resolution mapping. The generator and critic components are trained adversarially for 40 epochs on a single A100 GPU requiring about 20 hours. No noise injection is performed in the generator, resulting in a determinstic WGAN downscaling model. To reduce the memory requirements during training, a smaller target region is chosen for the WGAN.

Figure 4a shows the diurnal cycle of the space-time averaged RMSE over the complete test year 2018 for both models. With an ensemble-averaged RMSE of $0.989\,\mathrm{K}$, the AtmoRep downscaling model clearly outperforms the WGAN ($\overline{RMSE} = 1.163\,\mathrm{K}$). The margin is largest for the afternoon and evening hours and can mainly be attributed to lower errors over the Alpine region. As depicted in Figure 4b, the spatial RMSE distribution is rather uniform with AtmoRep, whereas the WGAN exhibits RMSE values up to 3 K over the Alpine region. This clearly documents the superiority of AtmoRep for the downscaling task and its ability to fill in realistic orographic features in complex terrain even without explicit topographic information.

In contrast to the conclusion above, the WGAN model is slightly better in reproducing the spatial variability of the downscaled T2m field compared to AtmoRep (not shown). Power spectrum analysis, along with comparisons of the domain-averaged horizontal T2m gradient against the COSMO REA6 ground truth, indicates that AtmoRep underestimates small-scale spatial variability by approximately 10 % (not shown). This is not entirely surprising since we are evaluating the ensemble mean state of AtmoRep, which will decrease variability. When we look at individual ensemble members, the underestimation of variability is slightly reduced, but differences to COSMO REA6 remain. A possible reason for this could be the very small ensemble size of 4 members. An increased ensemble size would require a more efficient model configuration. Strategies for this include freezing portions of AtmoRep's encoder-decoder weights during fine-tuning or implementing a more light-weight tail network, for instance with Swin Transformers[30] or Perceiver IO-modules[31].

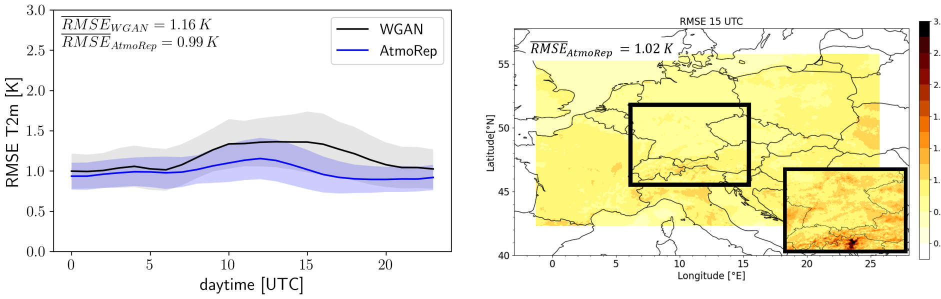

Figure 4. (a) Diurnal cycle of the domain-averaged RMSE for AtmoRep and the baseline WGAN downscaling model over the test data from 2018. The shaded area shows the standard deviation. (b) Spatial distribution of the RMSE with the AtmoRep downscaling model averaged over the test data at 15 UTC. The evaluation region is rendered in black. Additionally, the corresponding results of the WGAN model are displayed in the lower-right corner.

4 Summary and Outlook

AtmoRep is one of the first foundation models for weather and climate applications that fully exploits modern concepts of generative machine learning. In the 3 years since its conception, the model has demonstrated very good skills at a variety of meteorological tasks which had not been part of the original (pre-)training schedule. AtmoRep's capabilities for high-quality short-term forecasting, model correction, statistical downscaling, and counterfactual experiments have been demonstrated in[17]. Here, we extended the evaluation of AtmoRep by exploring its use as a data compression engine in a zero-shot setting and by further analysing the downscaling application including a comparison against a competitive Wasserstein GAN model.

The data (de)compression application explored a scenario where humongous amounts of climate data could be reduced by storing only every n-th grid box, k-th time step, or m-th model level. While this task has a lot of similarities with the pre-training task of random masking, our results nevertheless show that the systematic masking of specific patterns along the horizontal, vertical, or time dimension can introduce systematic biases in the reconstructed fields. RMSE values of reconstructed temperature fields range from about $0.4\,\mathrm{K}$ on the lowest model level to slightly higher values at the top level of $5\,\mathrm{km}$ altitude using compression ratios between 1.42 and 4. Although this is worse than the reconstruction quality of standard compression algorithms (e.g., JPEG), the advantage of AtmoRep is that it does allow for much larger compression ratios (combination of patterns) with relatively little degradation in performance. Furthermore, due to its probabilistic nature, AtmoRep can generate entire ensembles based on the compressed input of a single field. It can be expected that the reconstruction quality further improves when the model is fine-tuned and when we exploit cross-variable correlations.

Concerning the downscaling application, AtmoRep has demonstrated its superiority over a leading competitor model based on a WGAN. Although it failed to fully capture the enhanced variability of high-resolution temperature fields in complex terrain, it achieved very good scores in terms of absolute error and RMSE and generated credible high-resolution patterns following the complex orography over the Alps, even though no topographic information was provided to the model. Initial results suggest that the downscaling concept also works for other variables, in particular precipitation, which is most challenging but also highly relevant. In the current configuration, the ensemble size is very limited so that a robust assessment of the uncertainty of the downscaled field is not possible. Various approaches to overcome these limitations have been discussed above and are currently being explored. Already now, AtmoRep establishes a new state-of-the-art with respect to temperature downscaling and we are confident that this will also apply to other variables and regions.

The research on AtmoRep presented in this paper has been carried out with very little specific funding. Only recently, several projects that aim to further develop AtmoRep into a versatile model for weather and climate applications have been granted and the AtmoRep consortium continues to grow. While foundation models for weather and climate are still in their infancy, AtmoRep already allows some glimpses into what may become possible with such tools. It can be expected that foundation models for weather and climate will at some point replace classical numerical models in many different application areas as they are substantially faster and often better. However, there are still several fundamental questions to be solved and various technical challenges to be overcome. The evolution of supercomputing centers to provide more dedicated support for AI applications is one important cornerstone for building a bright future for weather and climate AI.

Acknowledgements

JSC, NIC and GCS for the provision of computing time and storage under the projects deeprain, deepacf, and atmorep. ECMWF for producing ERA5 and making it available. Several colleagues from JSC for technical support and advice. We also acknowledge the efforts of Nils Nobre Witter who conducted few data compression tests as a part of his bachelor thesis. We acknowledge the CERN Knowledge transfer fund for the support in

the development of the AtmoRep core model.

References

1. Freie Universität Berlin, *Definition of Weather and Climate*, URL: https://www.geo.fu-berlin.de/en/v/iwm-network/learning_content/-background/basics_climategeography/definitions/index.html (Accessed: 2024-10-11).
2. WMO, *WMO Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2019)*, WMO Publication No. 1267, 2021. URL: https://wmo.int/publication-series/wmo-atlas-of-mortality-and-economic-losses-from-weather-climate-and-water-extremes-1970-2019.
3. D. D. Holm, J. E. Marsden, T. S. Ratiu, *The Euler-Poincaré Equations in Geophysical Fluid Dynamics* Geophysical Fluid Dynamics **251–300**, Cambridge University Press, 2002 URL: https://authors.library.caltech.edu/19842/
4. D. R. Durran, *Numerical methods for fluid dynamics: with applications to geophysics*
5. P. Bauer, A. Thorpe, G. Brunet, *The quiet revolution of numerical weather prediction*, Nature **525**, 47–55, 2015 URL: http://www.nature.com/doifinder/10.1038/nature14956.
6. J. Rockström et al., *Safe and just earth system boundaries*, Nature **619**, 102–111, 2023 URL: https://doi.org/10.1038/s41586-023-06083-8.
7. T. N. Palmer, *Stochastic weather and climate models*, Nature Reviews Physics **1**, 463–471, 2019 URL: https://doi.org/10.1038/s42254-019-0062-2.
8. D. Maraun, M. Widmann, *Statistical Downscaling Concepts and Methods*, In: Statistical Downscaling and Bias Correction for Climate Research, Cambridge University Press, 2018, pp. 133-134.
9. T. Palmer, B. Stevens, *The scientific challenge of understanding and estimating climate change*, Proceedings of the National Academy of Sciences **116**, 24390–24395, 2019 URL: https://www.pnas.org/doi/abs/10.1073/pnas.1906691116.
10. S. Karthik Mukkavilli, D. Salles Civitarese, J. Schmude, J. Jakubik, A. Jones, N. Nguyen, C. Phillips, S. Roy, S. Singh, C. Watson, R. Ganti, H. Hamann, U. Nair, R. Ramachandran, K. Weldemariam, *AI Foundation Models for Weather and Climate: Applications, Design, and Implementation*, arXiv **2309.10808**, , URL: https://arxiv.org/abs/2309.10808.
11. Z. Ben-Bouallegue, M. C. A. Clare, L. Magnusson, E. Gascon, M. Maier-Gerber, M. Janousek, M. Rodwell, F. Pinault, J. S. Dramsch, S. T. K. Lang, B. Raoult, F. Rabier, M. Chevallier, I. Sandu, P. Dueben, M. Chantry, F. Pappenberger, *The rise of data-driven weather forecasting*, arXiv **2307.10128**, , URL: https://arxiv.org/abs/2307.10128
12. G. J. Hakim, S. Masanam, *Dynamical Tests of a Deep Learning Weather Prediction Model*, Artificial Intelligence for the Earth Systems **3**, e230090, 2024 Publisher: American Meteorological Society, Boston MA, USA. URL: https://journals.ametsoc.org/view/journals/aies/3/3/AIES-D-23-0090.1.xml
13. K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, *Pangu-Weather: A 3D High-Resolution Model for Fast and Accurate Global Weather Forecast*, arXiv preprint

arXiv:2211.02556 , , (2022). URL: https://arxiv.org/abs/2211.02556.

14. C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, J. Brandstetter, P. Garvan, M. Riechert, J. Weyn, H. Dong, A. Vaughan, J. K. Gupta, K. Tambiratnam, A. Archibald, E. Heider, M. Welling, R. E. Turner, P. Perdikaris, *Aurora: A Foundation Model of the Atmosphere*, arXiv preprint arXiv:2405.13063, 2024. URL: https://arxiv.org/abs/2405.13063.
15. T. Nguyen, R. Shah, H. Bansal, T. Arcomano, S. Madireddy, R. Maulik, V. Kotamarthi, I. Foster, A. Grover, *Scaling transformer neural networks for skillful and reliable medium-range weather forecasting*, arXiv preprint arXiv:2312.03876, 2023. URL: https://arxiv.org/abs/2312.03876.
16. I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, M. Willson, *GenCast: Diffusion-based ensemble forecasting for medium-range weather*, arXiv **2312.15796**, , URL: https://arxiv.org/abs/2312.15796
17. C. Lessig, I. Luise, B. Gong, M. Langguth, S. Stadtler, M. Schultz, *AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning*, arXiv preprint arXiv:2308.13280, 2023. URL: https://arxiv.org/abs/2308.13280.
18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Attention Is All You Need*, arXiv preprint arXiv:1706.03762, 2023. URL: https://arxiv.org/abs/1706.03762.
19. H. Hersbach, D. Dee, *ERA5: The fifth generation of ECMWF atmospheric reanalyses of the global climate*, ERA5 Documentation **2020**, , URL: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5.
20. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv preprint arXiv:1810.04805, 2018. URL: https://arxiv.org/abs/1810.04805.
21. A. Dosovitskiy, H. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhang, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv preprint arXiv:2010.11929, 2020. URL: https://arxiv.org/abs/2010.11929.
22. S. Rasp, S. Lerch, *Neural networks for postprocessing ensemble weather forecasts*, Monthly Weather Review **146**, 3885–3900, 2018. URL: https://journals.ametsoc.org/view/journals/mwre/146/11/mwr-d-18-0187.1.xml.
23. S. C. Scherrer, *Temperature monitoring in mountain regions using reanalyses: lessons from the Alps* Environmental Research Letters **15.4**, 044005, 2020 DOI: 10.1088/1748-9326/ab702d.
24. H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Munoz-Sabater, ... & J. N. Thépaut, *The ERA5 global reanalysis* Quarterly Journal of the Royal Meteorological Society **146.730**, 1999-2049, 2020 DOI: 10.1002/qj.3803.
25. P. Bauer, P. D. Dueben, T. Hoefler, et al., *The digital revolution of Earth-system science*, Nature Computational Science, vol. 1, pp. 104–113, 2021. URL: https://doi.org/10.1038/s43588-021-00023-0.
26. M. Govett, et al., *Exascale Computing and Data Handling: Challenges and Opportunities for Weather and Climate Prediction*, Bulletin of the American Meteorological Society, 2024, in press. URL:

https://doi.org/10.1175/BAMS-D-23-0220.1.

27. C. Bollmeyer, J. D. Keller, C. Ohlwein, S. Wahl, S. Crewell, P. Friederichs, A. Hense, J. Keune, S. Kneifel, I. Pscheidt, S. Redl, and S. Steinke, *Towards a high-resolution regional reanalysis for the European CORDEX domain*, Quarterly Journal of the Royal Meteorological Society, vol. 141, pp. 1–15, 2015. URL: https://doi.org/10.1002/qj.2486.
28. M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein GAN*, arXiv preprint arXiv:1701.07875, 2017. URL: https://arxiv.org/abs/1701.07875.
29. Y. Sha, D. J. Gagne II, G. West, and R. Stull, *Deep-Learning-Based Gridded Downscaling of Surface Meteorological Variables in Complex Terrain. Part I: Daily Maximum and Minimum 2-m Temperature*, Journal of Applied Meteorology and Climatology, vol. 59, pp. 2057–2073, 2020. URL: https://doi.org/10.1175/JAMC-D-20-0057.1.
30. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*, arXiv preprint arXiv:2103.14030, 2021. URL: https://arxiv.org/abs/2103.14030.
31. A. Jaegle, S. Borgeaud, J. B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Hénaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, *Perceiver IO: A General Architecture for Structured Inputs & Outputs*, arXiv preprint arXiv:2107.14795, 2022. URL: https://arxiv.org/abs/2107.14795.