The Effect of BERT Training on Atmospheric Data Interpolation

Belkis Asma Semcheddine[¥], Savvas Melidonis[¥], Michael Langguth[¥], Martin Schultz[¥,¤], Christian Lessig[¢] , Ilaria Luise[¢]

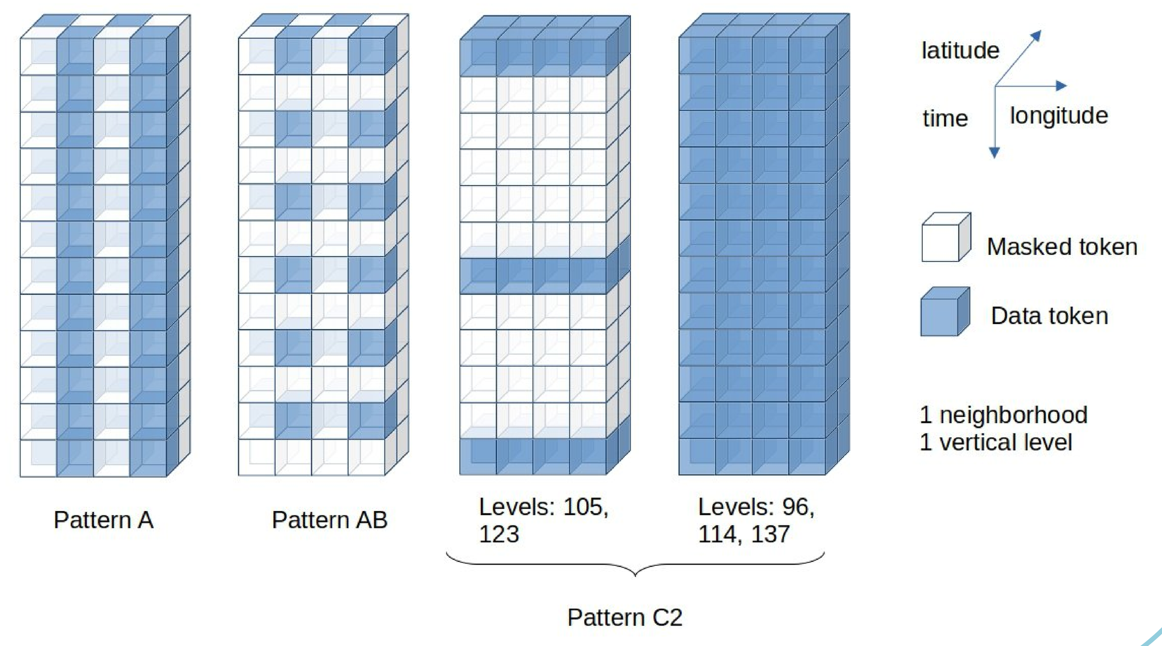
¥ Jülich Supercomputing Center, Forschungszentrum Jülich, 52428 Jülich, Germany. **Contact Person**: a.semcheddine@fz-juelich.de.
¤ University of Cologne, Department of Computer Science, Cologne, Germany.
¢ European Centre for Medium-Range Weather Forecasts, 53175 Bonn, Germany.

Motivation

- Transformer-based models are state-of-the-art in earth system modeling.
- AtmoRep, a transformer-based atmospheric model, uses BERT-style masking during training.
- Systematic masking strategies are tested to probe model learning.
- Systematic masking can uncover overlooked dependencies.
- Guide future pretrained model design choices.

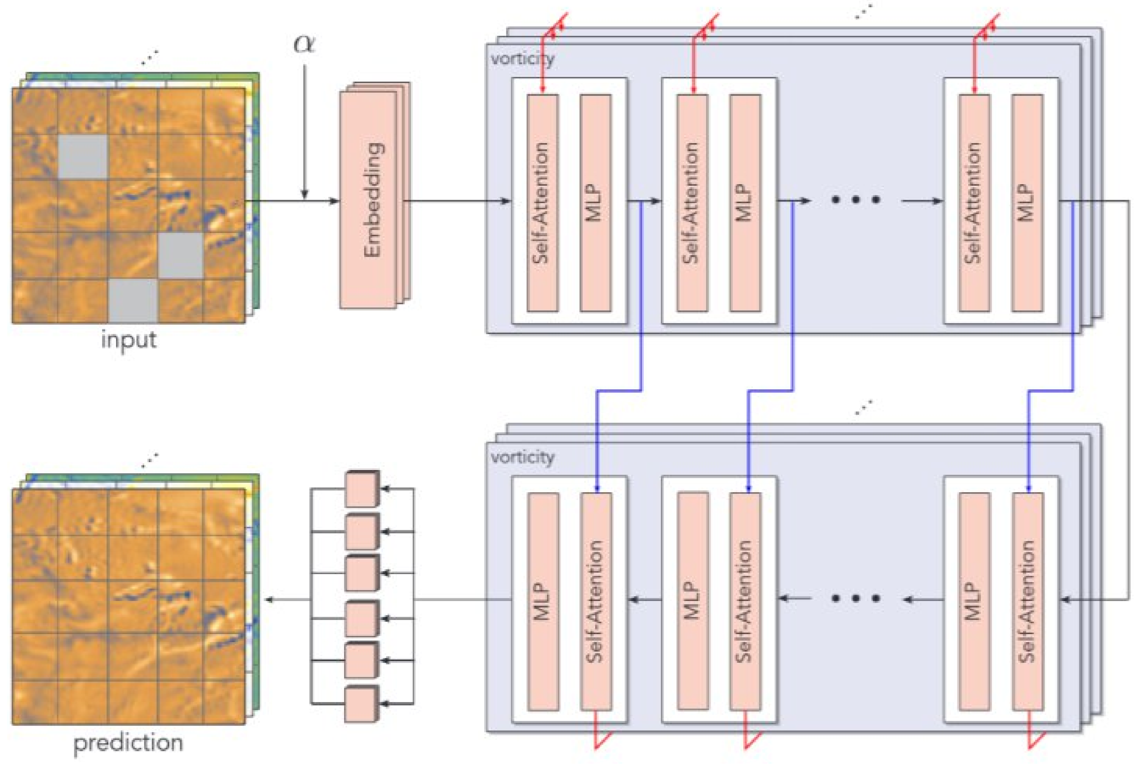
Method

- **Configuration A**: Checkerboard geographical masking at each model level and time step.
- **Configuration AB**: Combines geographical and temporal masking.
- **Configuration C2**: 9/12 Temporal masking on intermediate levels, excluding first, middle and last tokens.
- Zero-shot and fined tuned experiments on ERA5 temperature variable, year 2022.
- Study built on previous work[3]



AtmoRep[1]

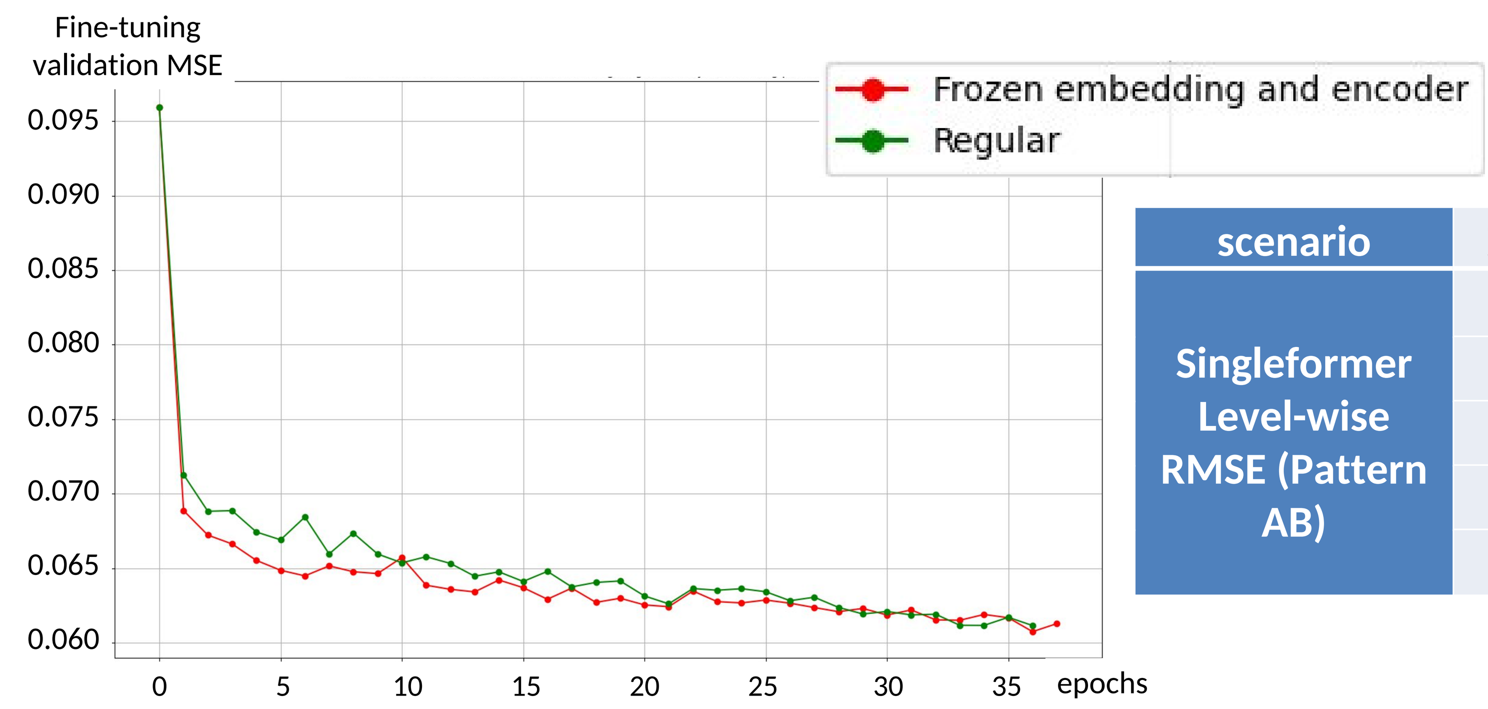
- Modelling of the atmosphere as a stochastical system: $p(y\,|x) \approx p_\theta(\hat{y}|\hat{x},\alpha)$
  $y, \hat{y}$: dependant (model) state.
  $x, \hat{x}$: initial (model) state.
  α: auxiliary information.
- Trained with ERA5[2] reanalysis data (1979 – 2017), in a self-supervised manner (BERT).
- Variables: u, v, w, vo, d, t, q on model levels [96, 105, 114, 123, 137] + tp as surface variable
- Modular configuration: Singleformer / Multiformer.



Results

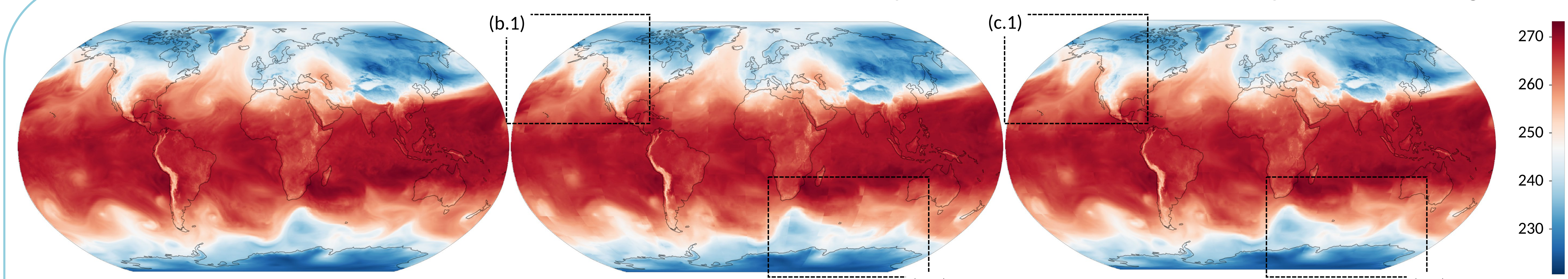
run	Zero-shot Level-wise RMSE singleformer					Zero-shot Level-wise RMSE multiformer				
A	0.4082	0.4541	0.6435	0.7033	0.8186	**0.3797**	**0.4252**	**0.6041**	**0.6787**	**0.7881**
AB	0.5806	0.6146	0.7835	0.8415	0.9427	0.8285	0.9010	1.0106	1.0858	1.0284
C2	-	0.4605	-	0.5187	-	-	0.4559	-	0.5114	-

- **Patterns A & C2**: Multiformer outperforms Singleformer.
- **Pattern AB**: Singleformer performs better.

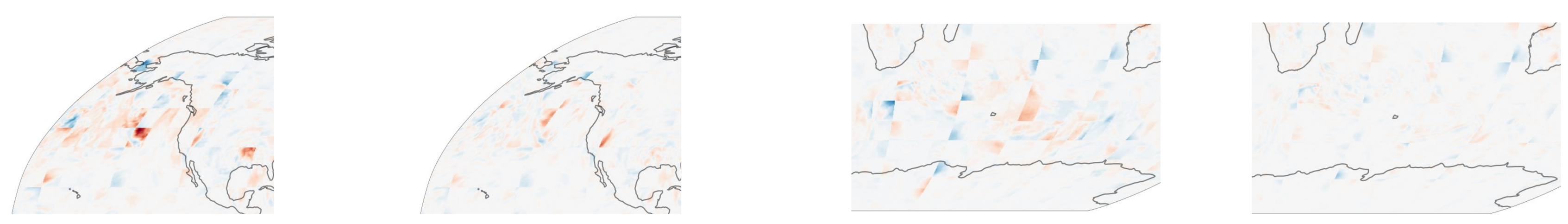


scenario	Zero-shot	Fine-tuned
Singleformer Level-wise RMSE (Pattern AB)	0.5806	**0.3471**
	0.6146	**0.3765**
	0.7835	**0.5243**
	0.8415	**0.5375**
	0.9427	**0.6143**

- Most RMSE improvement occurs within the first 5 epochs of **fine-tuning**.



(a) Ground truth (b) Zero-shot. RMSE = 0.4726 K. (b.2) (c) Fine-tuned. RMSE = 0.2964 K. (c.2)



(b.1) Zero-shot biais (c.1) Fine-tuned biais (b.2) Zero-shot biais (c.2) Fine-tuned biais

- **Fine-tuning** significantly reduces error — both visually and numerically (30% - 40% improvement).

Conclusion

- Temporal reconstruction is more effective than spatial/vertical.
- Fine-tuning is highly effective, 36% RMSE improvement on average, with rapid early gains.
- Results provide insight into how data is learned using BERT self-supervised learning.

References

1. Lessig et al., 2023. arXiv: 2308.13280
2. Hersbach, H. et al., 2020. Quarterly Journal of the Royal Meteorological Society 146, 1999–2049.
3. Patnala, A. et al. The NIC Symposium 2025, vol 52, pp 301-311.