



Scale-dependent analysis of the accuracy–activity trade-off in AI weather forecasts

Britta Seegebrecht¹, Sabrina Wahl¹, Stefanie Hollborn¹, Erik Pavel², Wael Almikaeel², Michael Langguth², Martin Schultz², Christian Lessig³, Ilaria Luise³, Juergen Gall⁴, Anas Al-Iahham⁴, and Mohamad Hakam Shams Eddin⁴

¹Deutscher Wetterdienst (DWD), Offenbach, Germany (britta.seegebrecht@dwd.de)

²Jülich Supercomputing Centre (JSC), Jülich, Germany

³European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany

⁴Lamarr Institute for Machine Learning and Artificial Intelligence, Rheinische Friedrich-Wilhelms- Universität Bonn, Bonn, Germany

Data-driven weather prediction models based on artificial intelligence (AI) have rapidly advanced in recent years and are frequently reported to outperform traditional physics-based numerical weather prediction (NWP) models for selected verification scores. However, optimization with respect to a specific loss function can adversely affect other metrics, potentially leading to unrealistic forecast characteristics, such as overly smooth spatial structures when mean-squared or mean-absolute error–based loss functions are used.

A robust and meaningful comparison of AI-based and NWP models therefore requires a carefully chosen and diverse set of verification metrics that accounts for potential dependencies. The main focus is placed on the prominent forecast accuracy-activity tradeoff, associated with the double penalty problem of deterministic forecasts. Related questions include: How sensitive is the relationship between accuracy and activity metrics to the choice of verification measure? Are there systematic differences between AI-based and NWP models? What is the impact of the (in)dependence between the AI training loss function and the verification metrics on the assessment of forecast skill?

These questions are addressed using both scale-independent and scale-dependent verification metrics, allowing the quantification of forecast performance on individual spatial scales.

As a starting point, global deterministic forecasts are considered. The analysis is partly based on forecasts from the Weather Prediction Model Intercomparison Project (WP MIP), which provides a collection of NWP and AI-model forecasts from multiple national weather services and research institutions.

The work is conducted within the RAINA project, which aims to develop a foundation model for the atmosphere with a particular focus on reliable, high-resolution forecasts of extreme wind and precipitation events. Consequently, the relation between, e.g., forecast activity and the predictive capability for extreme weather are of special interest.

