WeatherReal: A Benchmark Based on In-Situ Observations for Evaluating Weather Models

Weixin Jin[a], **Jonathan Weyn**[b], **Haiyu Dong**[b]

[a] *Microsoft Corporation* weixinjin@microsoft.com, jweyn@microsoft.com, haiyu.dong@microsoft.com

1. Abstract

In recent years, AI-based weather forecasting models have matched or even outperformed numerical weather prediction systems. However, most of these models have been trained and evaluated on reanalysis datasets like ERA5. These datasets, being products of numerical models, often diverge substantially from actual observations in some crucial variables like near-surface temperature, wind, precipitation and clouds - parameters that hold significant public interest. To address this divergence, we introduce WeatherReal, a novel benchmark dataset for weather forecasting, derived from global near-surface in-situ observations. WeatherReal also features a publicly accessible quality control and evaluation framework. This paper details the sources and processing methodologies underlying the dataset and further illustrates the advantage of in-situ observations in capturing hyper-local and extreme weather through comparative analyses and case studies. Using WeatherReal, we evaluated several data-driven models and compared them with leading numerical models. Our work aims to advance the AI-based weather forecasting research towards a more application-focused and operation-ready approach.

2. Datasets

WeatherReal comprises three near-surface observational datasets, two from diverse sources covering worldwide regions and one from weather reports collected from MSN weather users; a meticulously designed quality control system for raw observational data; and a benchmark for evaluating weather models, providing a unified standard to compare different models based on in-situ observations.

2.1 WeatherReal-ISD

WeatherReal-ISD (Fig. 1) is based on the Integrated Surface Database (ISD) [1], leveraging data from high-quality observation networks subjected to rigorous post-processing and quality control through independently developed algorithms.

For this dataset, initial data extraction and resampling at hourly levels is followed by station merging to consolidate data from stations with similar metadata or overlapping observation records. A comprehensive suite of quality control algorithms is then applied, which includes checks for value ranges, distributional gaps, spike detection, and cross-variable consistency. This rigorous quality control ensures the dataset remains accurate and reliable.

2.2 WeatherReal-Synoptic

Data of WeatherReal-Synoptic is obtained from Synoptic Data PBC, which brings together observation data from hundreds of public and private station networks worldwide, providing a comprehensive and accessible data service platform for critical environmental information. It encompasses a greater volume of data, a more extensive observation network, and a larger number of stations compared to ISD.

2.3 User Reports from MSN Weather

MSN Weather is a comprehensive global weather forecasting service integrated within the Microsoft Start ecosystem. To further improve the quality of weather forecasts, it continuously collects user reports through its weather service. Users can report the weather at their location, including the current approximate temperature, sky condition and whether there is precipitation along with other weather phenomena.

This dataset is under development and will be integrated into future versions of WeatherReal.

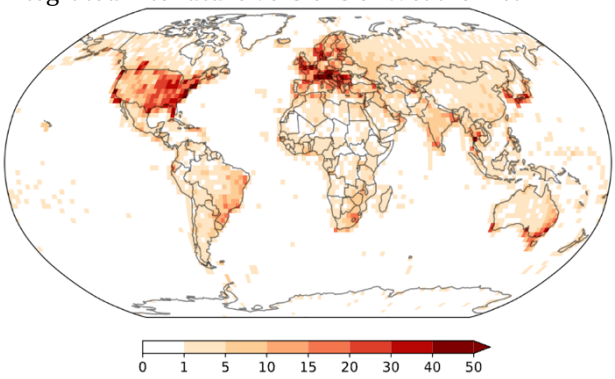

Fig. 1: Station density and 2m temperature reporting frequency for WeatherReal-ISD, shown as the number of stations within 2.5°×2.5° grids.

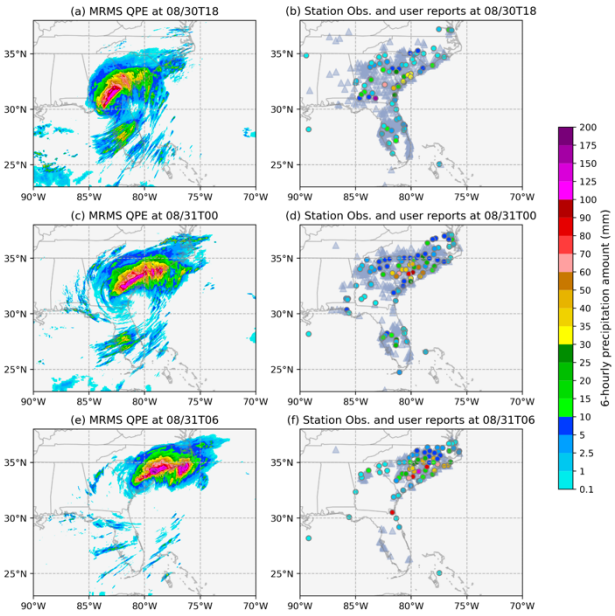

Fig. 2: The spatial distribution of 6-hour accumulated precipitation from August 30-31, 2023. The left column presents MRMS QPE Pass 2, while the right column includes station observations and user reports. In the right column, circles denote station observations, and triangles denote user reports. User reports are binary data, with only those indicating precipitation in the last 6 hours being marked.

Fig. 2 showcases the 6-hour cumulative

WeatherReal: A Benchmark Based on In-Situ Observations for Evaluating Weather Models

Weixin Jin[a], **Jonathan Weyn**[b], **Haiyu Dong**[b]

[a] *Microsoft Corporation* weixinjin@microsoft.com, jweyn@microsoft.com, haiyu.dong@microsoft.com

precipitation for three distinct time periods on August 30-31, 2023, as Hurricane Idalia made landfall in Florida, traversed Georgia, and moved into the Atlantic from the Carolinas. The figure contrasts Quantitative Precipitation Estimation (QPE) data from MRMS (Multi-Radar/Multi-Sensor) [2] with actual station observations and user reports. The right column of the figure displays only stations with complete observations, highlighting significant rainfall such as the peak 6-hour accumulation of 166.9 mm recorded in Georgia. User reports from MSN Weather, marked by triangles, effectively shows the changes in major precipitation locations over the 24-hour period. Within the area depicted on the map (20° by 20° range), over 6,000 user reports were collected, surpassing the number of 1-hour precipitation records from stations. Additionally, user reports are not confined to station locations, making their distribution more random and reflective of the spatial distribution of user density.

3. Tasks and Leaderboards

The WeatherReal benchmark introduces several evaluation tasks, each designed to assess the effectiveness of various weather forecasting models using in-situ observations, across varying time scales from short to medium range. The primary aim is to facilitate a comprehensive comparison of data-driven models against traditional numerical weather prediction models, thereby highlighting the strengths and potential areas for improvement in current AI-driven approaches.

All codes used for evaluation (along with quality control codes and part of the data) have been made publicly available on GitHub. Due to space limitations, please refer directly to our project at https://github.com/microsoft/WeatherReal-Benchmark.

Our complete work has been published on arXiv, please refer here for more details.

Acknowledgments

We sincerely thank NOAA and ECMWF for their efforts in constructing datasets, and for their open sharing of these data. Their work has been the primary sources for WeatherReal and this paper. We also extend our gratitude to Synoptic Data PBC for aggregating data from numerous sources. Their products and services empower public, private, government, and academic users with real-time and historical data, ensuring they can make informed decisions swiftly and confidently. Finally, we also thank Cristian Bodnar and the Aurora team for providing inference results from their model for evaluation.

References

[1] Adam Smith, Neal Lott, and Russ Vose. The integrated surface database: Recent developments and partnerships. Bulletin of the American Meteorological Society, 92(6):704–708, 2011. doi: 10.1175/2011BAMS3015.1.

[2] Jian Zhang, Kenneth Howard, Carrie Langston, Brian Kaney, Youcun Qi, Lin Tang, Heather Grams, Yadong Wang, Stephen Cocks, Steven Martinaitis, Ami Arthur, Karen Cooper, Jeff Brogden, and David Kitzmiller. Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. Bulletin of the American Meteorological Society, 97(4):621 – 638, 2016. doi: 10.1175/BAMS-D-14-00174.1.