Multi-Modal Observational Data For Weather Forecasting

Anonymous Author(s)
Affiliation
Address
email

Abstract

Existing AI weather forecasting systems have made strong progress in the last few years, but focus on predicting re-analysis targets (e.g., ERA5). AI forecasting systems share the same shortcomings of the re-analysis process including high computation cost, known non-physical artifacts, and oversampling in particular regions of the globe. In order to facilitate the development of end-to-end weather forecasting methods which bypass the need for re-analysis at operation time, we propose constructing a dataset of multi-modal weather observations containing weather stations measurements, microwave and infrared sounder results, and geospatial imagery. This dataset would help extend the results from other domains (e.g., computer vision, natural language) to weather forecasting by leveraging state-of-the-art multi-modal techniques, advancing both ML and weather forecasting. New methods enabled by this dataset will help reduce the train-serve discrepancy, and improve the operational usefulness of AI-driven weather forecasting.

1 Motivation

Modern state-of-the-art machine learning weather forecasting systems have demonstrated remarkable progress, outperforming traditional numerical weather prediction systems [1, 2, 3, 4, 5]. However, these systems are trained to predict and are evaluated on *gridded reanalysis* (e..g, ERA5 [6], MERRA-2 [7]) using data-assimilation, a technique to combine historical observations with physics-based numerical methods [8]. This couples the machine learning model to the assumptions and error characteristics of data-assimilation methods and induces a strong train-serve mismatch. In operational settings, individual weather stations and forecasters observe instrument-level readings and localized observations, not gridded re-analysis data. Traditional numerical weather prediction systems consolidate observations with data assimilation from multiple locations to produce a gridded state of weather dynamics, similar to re-analysis. This paradigm poses three unique challenges which motivates an end-to-end ML weather forecasting system which bypass traditional data assimilation.

Physical artifacts. Re-analysis and regridded weather data, specifically ERA5, is known to have artifacts and non-physical states, particularly for local phenomena and in the upper stratosphere and mesosphere [9, 10]. Near-surface variables and localized phenomena (e.g., precipitation, wind speeds) exhibit errors which can be pronounced for end-users. Near coastlines, ERA5 under-reports variance in temperature as a result of the land-surface model which blends the influence of the land and ocean. ERA5's topography is not representative of the true topography, and causing deviations from observational precipitation data, especially in the tropics and high topographic variability [11, 12].

Oversampling in resource rich locations. Because re-analysis generates a new global forecast state using observed quantities to correct for errors, locations with fewer observation instruments

Submitted to AI for Science workshop (NeurIPS 2025). Do not distribute.

have worse forecasts. Despite low overall forecast errors, existing ML weather forecast systems produce high errors in under-sampled locations, reflecting the data sampling discrepancy [2].

Online localized evaluation. Currently, generating localized predictions requires aggregating instruments across the globe, performing data assimilation, and disseminating forecasts. This workflow is computationally intensive, requiring large-scale CPU clusters, and is contingent upon a central authority (e.g., European Centre for Medium-Range Weather Forecasts or the National Oceanic and Atmospheric Administration). Enabling low-latency high fidelity forecasts necessitates developing novel ML weather forecasting systems which learn end-to-end from observational data.

2 Data Sources

Building an end-to-end ML weather forecasting system is blocked by the lack of a standardized ML-ready dataset of observational weather readings. Sensor and observation data represent a wealth of unstructured data, a regime deep learning models have often found success in [13, 14, 15, 16]. The heterogeneity which provides a rich source of data also poses a bottleneck due to differences in modalities with varying sampling geometries, cadence, units, quality control, and metadata. Surface stations (Integrated Surface Database ISD) report hourly variables at fixed sites with non-uniform coverage [17]. Radiosondes (Integrated Global Radiosonde Archive) profile and measure the atmosphere periodically [18]. Passive microwave and infrared sounders (Advanced Microwave Sounding Unit) provide brightness measurements along orbital paths [19, 20]. Geospatial imagery (GOES Advanced Baseline Imager) provide frequent multi-spectral imagery [21]. Individual observations are publicly available through government agencies, but there is no spatio-temporal standardization nor are they available in a format for high-throughput ML training workloads.

Existing ML-ready datasets. To date, there is one notable public dataset aiming to tackle this gap. Aardvark [22] introduces a dataset including station observations, satellite imagery, and sounder data. The released dataset is coarse in both time and space, down-sampling hourly observational readings into 1 day increments. Similarly, Aardvark down-samples in space to a 1 degree grid while many operational forecasts use a 0.25 degree grid [23]. Aardvarks' dataset, though limited in resolution, shows both the feasibility and necessity of a multi-modal observational dataset.

Technical Challenges. To demonstrate the challenge of curating this dataset, we examine the Integrated Global Radiosonde Archive (IGRA) provided by NOAA [18]. Each station's data is stored in plain text, and each station reports sounding data at different pressure levels and a different number of levels based on the available operational instruments. IGRA stations may also provide sounding data at heights without a pressure level reference, making it difficult to associate observed values across stations. The lack of a uniform format, even within a single modality, has been a barrier to the development of end-to-end ML forecast systems.

Compute and storage requirements. By its nature, observational data is readily available through government entities. However, the primary storage and compute costs will be processing the data, transforming into an ML-ready format, and distribution. The ISD in an uncompressed format, is over 600 GB [17] and growing daily. GOES-ABI is around 1GB per hour (unformatted > 1 TB). We anticipate that the primary cost will be the storage requirements of different modalities and aim for a final dataset size of < 3 TB. However, consolidating this final dataset will require more disk space during processing, quality control, and standardization.

3 Acceleration Potential

The proposed dataset of a multi-modal observation-first weather corpus would accelerate the development of end-to-end machine learning forecasting models by removing the standardization and processing required to convert observational data into an ML-ready format. This removes the need for researchers to perform error-prone data ingestion and standardization work. Multiple evaluation splits will also facilitate robust evaluations which reflect operational environments and probes the model's generalization capability. This dataset will help spur methods that work directly on observation streams, particularly relevant for near-surface variables and local hazardous weather conditions, improving operational AI forecasts.

References

[1] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, January 2025.

[2] Xiuyu Sun, Xiaohui Zhong, Xiaoze Xu, Yuanqing Huang, Hao Li, J. David Neelin, Deliang Chen, Jie Feng, Wei Han, Libo Wu, and Yuan Qi. FuXi Weather: A data-to-forecast machine learning system for global weather, 2024. _eprint: 2408.05472.

[3] Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, May 2025.

[4] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023.

[5] Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. AIFS – ECMWF's data-driven forecasting system, 2024. _eprint: 2406.01465.

[6] 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. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

[7] Ronald Gelaro, Will McCarty, Max J. Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A. Randles, Anton Darmenov, Michael G. Bosilovich, Rolf Reichle, Krzysztof Wargan, Lawrence Coy, Richard Cullather, Clara Draper, Santha Akella, Virginie Buchard, Austin Conaty, Arlindo M. da Silva, Wei Gu, Gi-Kong Kim, Randal Koster, Robert Lucchesi, Dagmar Merkova, Jon Eric Nielsen, Gary Partyka, Steven Pawson, William Putman, Michele Rienecker, Siegfried D. Schubert, Meta Sienkiewicz, and Bin Zhao. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14):5419 – 5454, 2017. Place: Boston MA, USA Publisher: American Meteorological Society.

[8] Marta Janisková and Philippe Lopez. Linearized Physics for Data Assimilation at ECMWF. In Seon Ki Park and Liang Xu, editors, *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*, pages 251–286. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[9] Hans Hersbach and National Center for Atmospheric Research Staff . The Climate Data Guide: ERA5 atmospheric reanalysis. September 2024.

[10] Bill Bell, Hans Hersbach, Adrian Simmons, Paul Berrisford, Per Dahlgren, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Raluca Radu, Dinand Schepers, Cornel Soci, Sebastien Villaume, Jean-Raymond Bidlot, Leo Haimberger, Jack Woollen, Carlo Buontempo, and Jean-Noël Thépaut. The ERA5 global reanalysis: Preliminary extension to 1950. *Quarterly Journal of the Royal Meteorological Society*, 147(741):4186–4227, 2021. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.4174.

[11] Robert W Carver and Alex Merose. ARCO-ERA5: An Analysis-Ready Cloud-Optimized Reanalysis Dataset, 2023.

[12] Lisa V Alexander, Margot Bador, Rémy Roca, Steefan Contractor, Markus G Donat, and Phuong Loan Nguyen. Intercomparison of annual precipitation indices and extremes over global land areas from in situ, space-based and reanalysis products. *Environmental Research Letters*, 15(5):055002, April 2020. Publisher: IOP Publishing.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs].

[14] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, October 2023.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, June 2022. ISSN: 2575-7075.

[16] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. arXiv:2508.10104 [cs].

[17] 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. Place: Boston MA, USA Publisher: American Meteorological Society.

[18] Imke Durre, Xungang Yin, Russell S. Vose, Scott Applequist, and Jeff Arnfield. Enhancing the Data Coverage in the Integrated Global Radiosonde Archive. *Journal of Atmospheric and Oceanic Technology*, 35(9):1753 – 1770, 2018. Place: Boston MA, USA Publisher: American Meteorological Society.

[19] Earth Observing System (EOS)/Advanced Microwave Sounding Unit-A (AMSU-A): Calibration management plan. Technical Report AEROJET-10356, September 1994. NTRS Document ID: 19950007291 NTRS Research Center: Legacy CDMS (CDMS).

[20] B. Lambrigtsen, E. Fetzer, E. Fishbein, S.-Y. Lee, and T. Pagano. AIRS - the Atmospheric Infrared Sounder. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages 2204–2207 vol.3, September 2004.

[21] Timothy J. Schmit, Paul Griffith, Mathew M. Gunshor, Jaime M. Daniels, Steven J. Goodman, and William J. Lebair. A Closer Look at the ABI on the GOES-R Series. *Bulletin of the American Meteorological Society*, 98(4):681 – 698, 2017. Place: Boston MA, USA Publisher: American Meteorological Society.

[22] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, May 2025.

[23] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models, 2023. _eprint: 2308.15560.