WINDSET: Weather Insights and Novel Data for Systematic Evaluation and Testing

Rajat Shinde, Sujit Roy & Christopher E Phillips {rajat.shinde}@uah.edu
Earth System Science Center, The University of Alabama in Huntsville,
Huntsville, AL, USA

Aman Gupta & Aditi Sheshadri {ag4680}@stanford.edu
Doerr School of Sustainability, Stanford University, Stanford, CA, USA

Manil Maskey, Rahul Ramachandran {rahul.ramachandran}@nasa.gov
NASA Marshall Space Flight Center, Huntsville, AL, USA

Reviewed on OpenReview: https://openreview.net/forum?id=JZDnsKaVGE

Editor:

Abstract

The demand for accurate and timely weather predictions continues to rise due to the ubiquitous role of weather in our day-to-day activities. We present WINDSET - *Weather Insights and Novel Data for Systematic Evaluation and Testing* as a catalog of AI-ready datasets for validating the capabilities of weather foundation models across multiple downstream tasks. The WINDSET datasets are accessible on HuggingFace[1] and the code to create the datasets is available on GitHub[2].

Keywords: Weather Foundation Model, Downstream Studies, AI-ready datasets

1 Weather Insights and Novel Data for Systematic Evaluation and Testing

The development of foundation models (FM) has completely transformed data-driven research in recent years. These models are equipped with advanced natural language processing and machine learning (ML) capabilities. For instance, a Weather Foundation Model (Weather FM) not only provides general forecasting skill but also addresses specialized downstream tasks such as generating forecast discussions or predicting aircraft turbulence. AI-ready data plays a crucial role in fine-tuning FMs to these tasks. Our contribution, WINDSET (Weather Insights and Novel Data for Systematic Evaluation and Testing), consists of multiple multi-modal and multi-resolution AI-ready datasets that focus on different downstream tasks related to numerical weather prediction (NWP) and forecasting. Table 1 presents a summary of the WINDSET datasets and their uniqueness in terms of AI readiness.

Non-Local Gravity Wave Parameterization Atmospheric mesoscales (including gravity waves (GWs), clouds, and precipitation) are often *parameterized* in NWP and climate

1. https://huggingface.co/datasets/nasa-impact/WINDSET
2. https://github.com/NASA-IMPACT/WINDSET

Table 1: Description of the datasets and their contribution in terms of AI-readiness.

Dataset (Label modality)	Contribution
Non-Local Gravity Wave Parameterization (4-D Array)	Can be used to learn and represent the subgrid-scale GW fluxes and mesoscale processes in coarser climate models using ML
Weather Forecast Generation (Text)	First of its kind for generating natural language based weather forecast summaries
Aviation Turbulence Prediction (Binary)	Can be used for fine-tuning an AI model for predicting turbulence using PIREP reports
Searching Weather Analogs (Image)	Dataset for fast searching weather analogs from an unindexed database for a given weather parameter and time

models. These parameterizations are idealized representations of the observed process and typically assume strict vertical (or single-column) process evolution (Chen et al., 2018; Plougonven et al., 2020; Gupta et al., 2024), leading to large-scale momentum imbalances and cold biases in climate models (McLandress et al., 2012). GW observations, however, show significant horizontal propagation. While the single-column parameterization design is limited by the discretization of NWP and climate models, ML presents a fresh approach to learning the full 3D evolution of GW fluxes, and representing them in climate models.

Detection of Aviation Turbulence Turbulence in the lower and middle atmosphere presents risks for passenger and cargo airliners, especially when it is encountered unexpectedly (Ito et al., 2020; Golding, 2000). Turbulence, however, is a finescale feature that defies direct prediction by even high-resolution NWP models. ML techniques have been explored for turbulence prediction (Emara et al., 2021; Williams, 2014; Hon et al., 2020) using information from NWP or onboard sensors, but deep learning (DL) approaches remain largely uninvestigated. This dataset supports a DL approach to improve aviation turbulence forecasting using pilot reports (PIREPs). (See Figure 1, 2, A.1). Provided inputs are vertical profiles of wind, temperature, and humidity paired with turbulence observations. This approach enables near-real time DL predictions using either observations or NWP output.

Searching Weather Parameter Analogs in the Archive Weather analogs are historical events similar to the current atmospheric state. For forecasters, past events provide insight into current conditions. Weather analogs are also used to initialize large-ensemble NWP systems and can be leveraged to train deep-learning models. The proposed dataset enables training an image similarity search model capable of searching over an archive for similar events based on a weather parameter such as temperature or pressure (See Figure 3, A.2). AI methods for analog search are rapid due to performing the search in the AIs own encoded vector space, enabling the search of entire re-analysis archives in seconds.

Generating Natural Language Weather Forecast Discussions From the user's perspective, one of the prominent outputs of weather forecasting is to publish forecast reports in natural language. This dataset comprises a text-based summary of a corresponding atmospheric state (Refer Figure 4, A.3). To the best of our knowledge, this is a first-of-its-kind dataset that can be used to benchmark an AI model for generating automated natural-language-based weather forecast summaries.

References

Ting-Chen Chen, M. K. Yau, and Daniel J. Kirshbaum. Assessment of Conditional Symmetric Instability from Global Reanalysis Data. *Journal of the Atmospheric Sciences*, 75 (7):2425–2443, July 2018. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-17-0221.1.

Mariam Emara, Marcos dos Santos, Noah Chartier, Jamey Ackley, Tejas G. Puranik, Alexia Payan, Michelle Kirby, Olivia J. Pinon, and Dimitri N. Mavris. Machine Learning Enabled Turbulence Prediction using Flight Data for Safety Analysis. In *32nd Congress of the International Council of the Aeronautical Sciences*, Shanghai, China, 2021.

Wayne Golding. Turbulence and Its Impact on Commercial Aviation. *Journal of Aviation/Aerospace Education & Research*, 2000. ISSN 2329-258X. doi: 10.15394/jaaer.2002.1301.

Aman Gupta, Robert Reichert, Andreas Dörnbrack, Hella Garny, Roland Eichinger, Inna Polichtchouk, Bernd Kaifler, and Thomas Birner. Estimates of Southern Hemispheric Gravity Wave Momentum Fluxes Across Observations, Reanalyses, and Kilometer-scale Numerical Weather Prediction Model. *Journal of the Atmospheric Sciences*, -1(aop), January 2024. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-23-0095.1.

Kai Kwong Hon, Cho Wing Ng, and Pak Wai Chan. Machine learning based multi-index prediction of aviation turbulence over the Asia-Pacific. *Machine Learning with Applications*, 2:100008, 2020. ISSN 2666-8270. doi: 10.1016/j.mlwa.2020.100008.

J. Ito, H. Niino, and K. Yoshino. Large Eddy Simulation on Horizontal Convective Rolls that Caused an Aircraft Accident during its Landing at Narita Airport. *Geophysical Research Letters*, 47(6), 2020. ISSN 0094-8276. doi: 10.1029/2020gl086999.

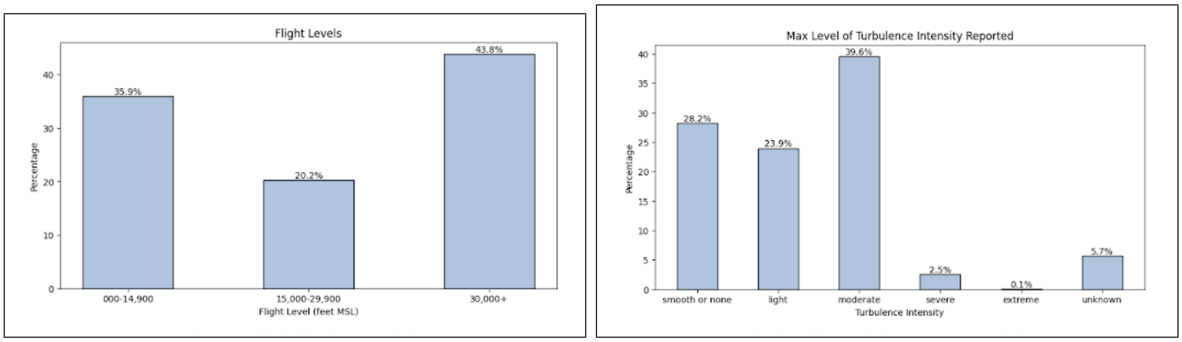
Charles McLandress, John F. Scinocca, Theodore G. Shepherd, M. Catherine Reader, and Gloria L. Manney. Dynamical Control of the Mesosphere by Orographic and Nonorographic Gravity Wave Drag during the Extended Northern Winters of 2006 and 2009. *J. Atmos. Sci.*, 70(7):2152–2169, December 2012. ISSN 0022-4928. doi: 10.1175/JAS-D-12-0297.1.

Riwal Plougonven, Alvaro de la Cámara, Albert Hertzog, and François Lott. How does knowledge of atmospheric gravity waves guide their parameterizations? *Quarterly Journal of the Royal Meteorological Society*, 146(728):1529–1543, 2020. ISSN 1477-870X. doi: 10.1002/qj.3732.

John K. Williams. Using random forests to diagnose aviation turbulence. *Machine Learning*, 95(1):51–70, 2014. ISSN 0885-6125. doi: 10.1007/s10994-013-5346-7.

Appendix A. Visualization of data samples from some of the proposed datasets

A.1 Detection of Aviation Turbulence



(a) (b)

Figure 1: Illustration of bar charts for (a) flight levels, (b) maximum turbulence intensity, associated with pilot aviation weather reports (PIREPs)

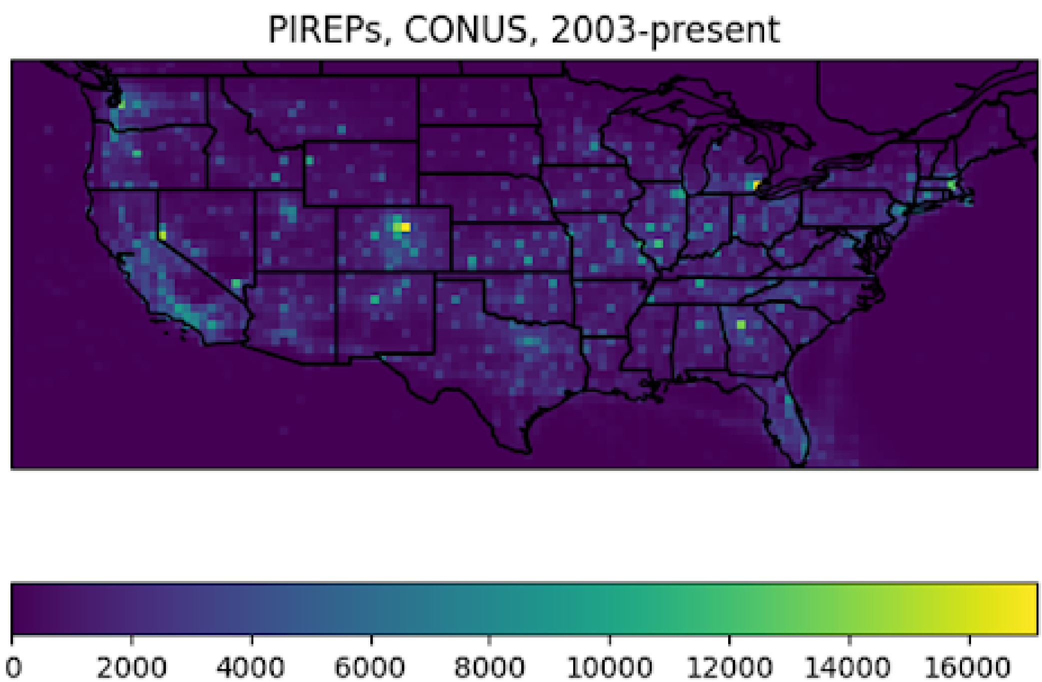
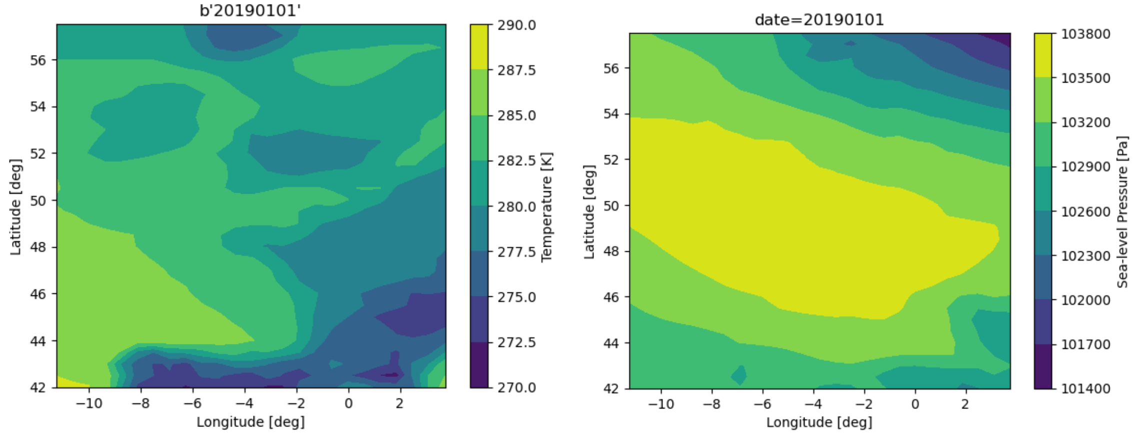


Figure 2: Spatial distribution of PIREPs binned to 0.5x0.625 over the contiguous U.S.A.

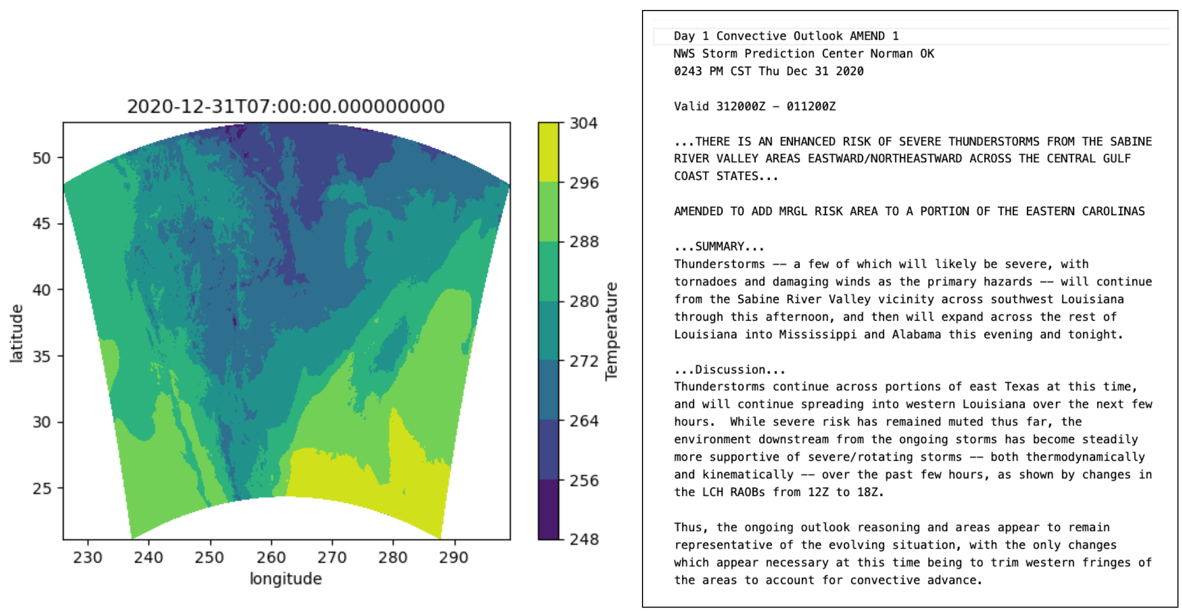
A.2 Searching Weather Parameter Analogs in the Archive



(a) (b)

Figure 3: Illustration of training images for (a) temperature (b) sea-level pressure (slp), for January 01, 2019. These training images are used for analog search of these parameters over a database of images.

A.3 Generating Natural Language Weather Forecast Discussions




```
Day 1 Convective Outlook AMEND 1
NWS Storm Prediction Center Norman OK
0243 PM CST Thu Dec 31 2020

Valid 312000Z - 011200Z

...THERE IS AN ENHANCED RISK OF SEVERE THUNDERSTORMS FROM THE SABINE
RIVER VALLEY AREAS EASTWARD/NORTHEASTWARD ACROSS THE CENTRAL GULF
COAST STATES...

AMENDED TO ADD MRGL RISK AREA TO A PORTION OF THE EASTERN CAROLINAS

...SUMMARY...
Thunderstorms -- a few of which will likely be severe, with
tornadoes and damaging winds as the primary hazards -- will continue
from the Sabine River Valley vicinity across southwest Louisiana
through this afternoon, and then will expand across the rest of
Louisiana into Mississippi and Alabama this evening and tonight.

...Discussion...
Thunderstorms continue across portions of east Texas at this time,
and will continue spreading into western Louisiana over the next few
hours.  While severe risk has remained muted thus far, the
environment downstream from the ongoing storms has become steadily
more supportive of severe/rotating storms -- both thermodynamically
and kinematically -- over the past few hours, as shown by changes in
the LCH RAOBs from 12Z to 18Z.

Thus, the ongoing outlook reasoning and areas appear to remain
representative of the evolving situation, with the only changes
which appear necessary at this time being to trim western fringes of
the areas to account for convective advance.
```

(a) (b)

Figure 4: Illustration of training data for (a) temperature (b) corresponding forecast report for December 31, 2020.