Atmos ML Journal Club

August Posch 5 Nov 2026

Google DeepMind

2026-9-3

WeatherNext 3: Increasing resolution and performance of global weather models with raw observations

Stephan Rasp[1*], Boris Babenko[2*], Dominic Masters[1*], Andrew El-Kadi[1*], Samier Merchant[2*], Guy Shalev[1*], Ilan Price[1*], Fred Zyda[2], Remi Lam[1], Sasha Shysheya[1], Matthew Willson[1], Stratis Markou[1], Shreya Agrawal[1], Suhani Vora[1], Mohammed Alewi Hassen[1], Sunny Mak[1], Tom R. Andersson[1], Megan Bela[3], Akib Uddin[2], Nofar Peled Levi[1], Ben Gaiarin[1], Ferran Alet[1*], Aaron Bell[2], Peter Battaglia[1] and Alvaro Sanchez-Gonzalez[1*]

[*]Equal contribution, [1]Google DeepMind, [2]Google Research, [3]Google

Overview

- State of the art medium-range weather forecasting model
- Previous MLWP used physics-based analysis state as initial conditions - this one does that, plus use geostationary satellite obs as inputs
- Hourly, 0.1 degree resolution, more variables (solar radiation, cloud cover, precipitation)
- Get forecasts at continuous values of location and time (not just a gridcell)

Overview

- Most MLWP initialized and trained on reanalysis like ERA5 - 6-hourly frequency, 6 hours latency operationally due to Data Assimilation
- Here, let's leverage obs directly
- "Station head" is how you do continuous location/time - query it

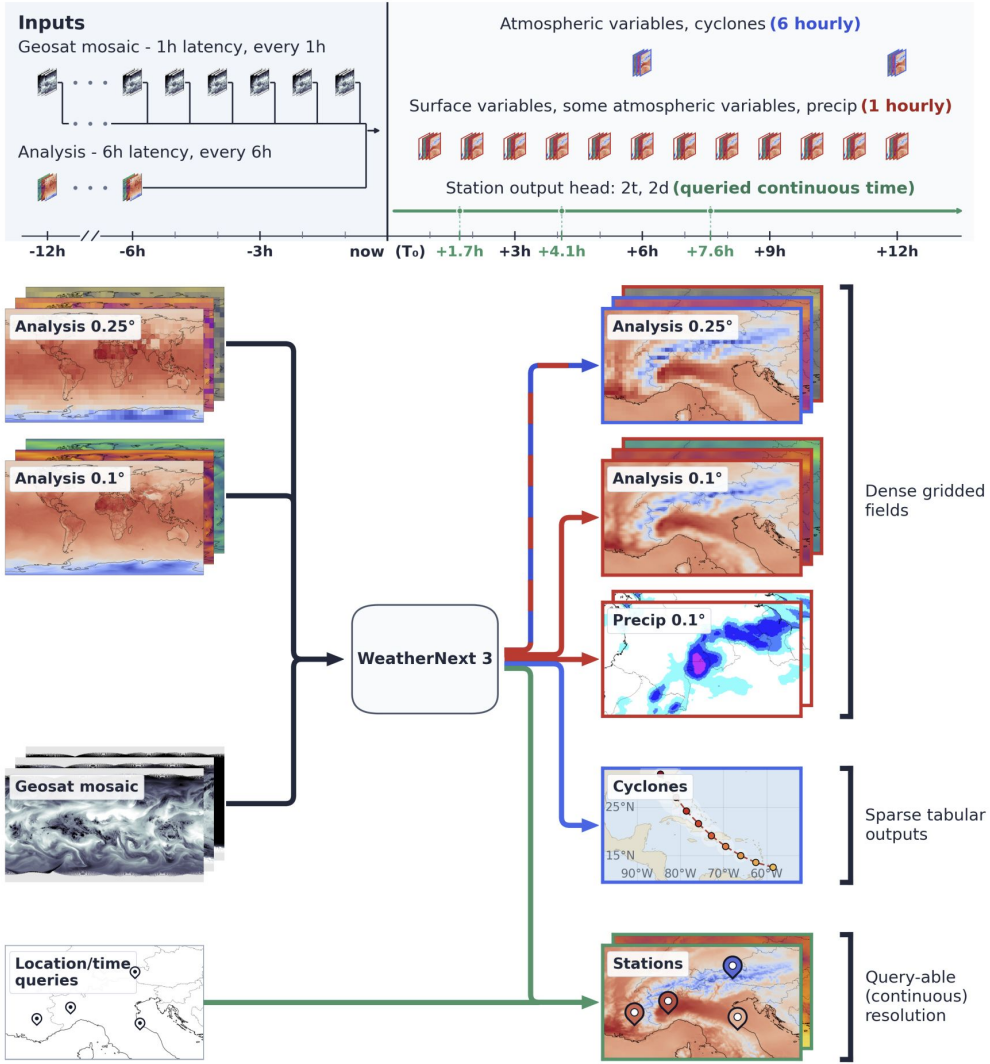
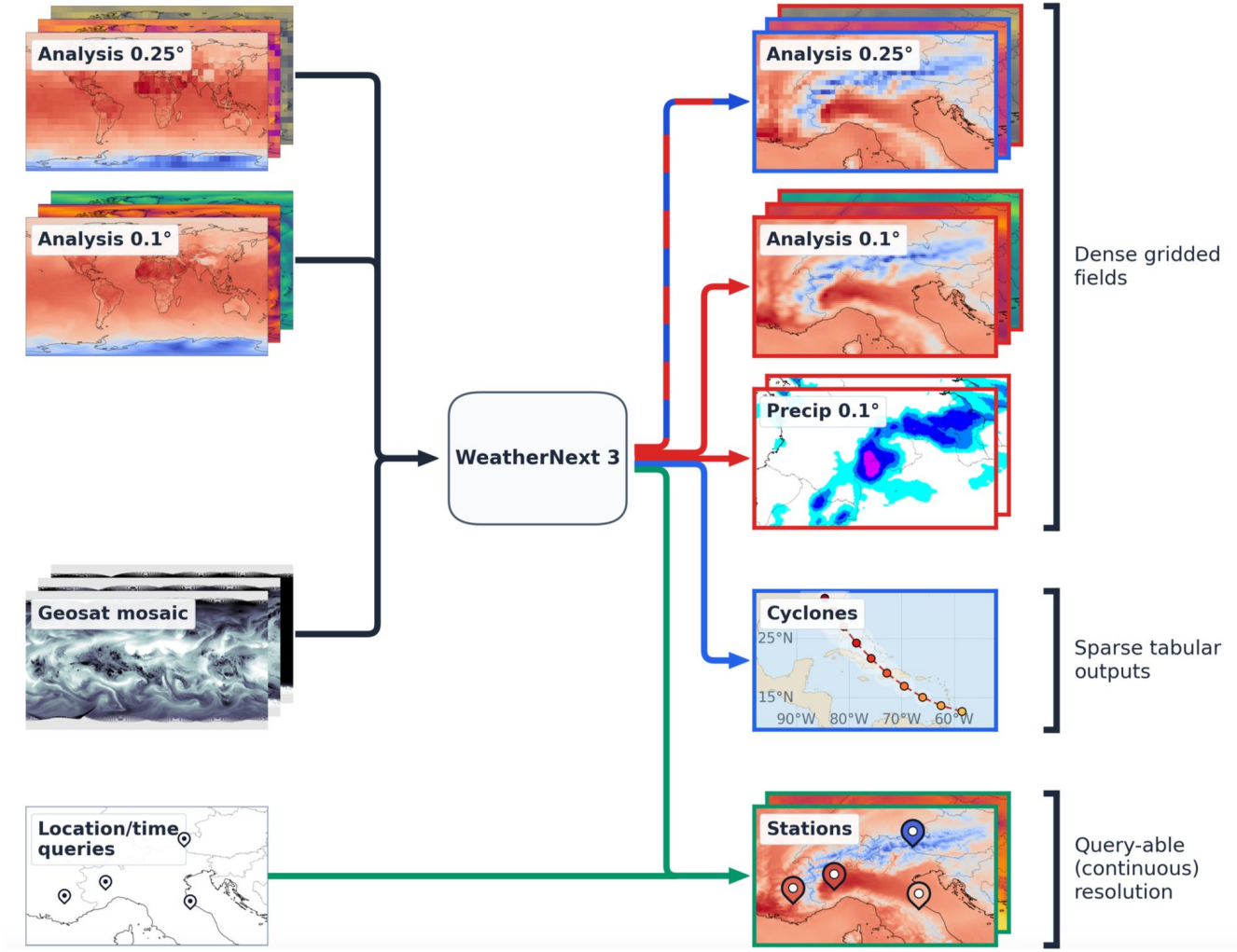


Figure 1 | **Model overview.** Schematic showing the different input and output modalities, and spatiotemporal resolutions of WeatherNext 3 (WN3). WN3 takes two sources of input: (a) two analysis frames (6 hours apart), with an operational latency of 5 hours, and (b) the 12 most recent geostationary satellite mosaic frames, which are available every hour and the most recent of which has an operational latency of just under 1 hour. Analysis inputs are 0.1° resolution for a subset of key atmospheric variables and surface variables, and 0.25° for other variables. WN3 predicts precipitation, surface variables and some atmospheric variables at hourly temporal resolution, with the remaining variables predicted in 6 hour timesteps. WN3's station output head predicts 2m temperature and 2m relative humidity at arbitrary spatial and temporal resolution (i.e. is continuously query-able in space and time). WN3 also produces off-grid sparse or tabular cyclone predictions at 6-hourly temporal resolution.


Analysis 0.25°
Analysis 0.1°
Geosat mosaic
Location/time queries
WeatherNext 3
Analysis 0.25°
Analysis 0.1°
Precip 0.1°
Cyclones
25°N
15°N
90°W
80°W
70°W
60°W
Stations
Dense gridded fields
Sparse tabular outputs
Query-able (continuous) resolution

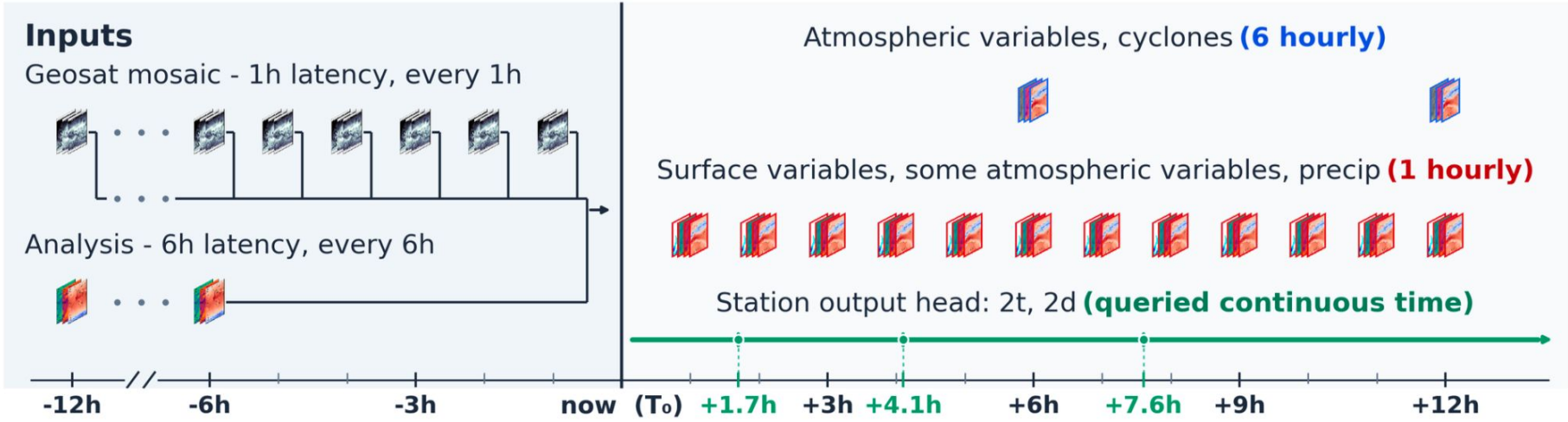


Figure 1 | **Model overview.** Schematic showing the different input and output modalities, and spatiotemporal resolutions of WeatherNext 3 (WN3). WN3 takes two sources of input: (a) two analysis frames (6 hours apart), with an operational latency of 5 hours, and (b) the 12 most recent geostationary satellite mosaic frames, which are available every hour and the most recent of which has an operational latency of just under 1 hour. Analysis inputs are 0.1° resolution for a subset of key atmospheric variables and surface variables, and 0.25° for other variables. WN3 predicts precipitation, surface variables and some atmospheric variables at hourly temporal resolution, with the remaining variables predicted in 6 hour timesteps. WN3's station output head predicts 2m temperature and 2m relative humidity at arbitrary spatial and temporal resolution (i.e. is continuously query-able in space and time). WN3 also produces off-grid sparse or tabular cyclone predictions at 6-hourly temporal resolution.

Key features

- Architecture: encoder-processor-decoder
- 15-day, 64-member ensemble
- Loss function: Continuous Ranked Probability Score (CRPS) - helps you get enough ensemble spread while having each ensemble member as accurate as possible

Key features

- Data modalities and resolutions - separate encoders/decoders for 0.1 or 0.25 degree
- Obs input and hourly initialization - geostationary satellite data directly (hourly) + HRES analysis (6-hourly)
- Latent interpolation - you can interpolate the latent state to any time/location, and decode it there
- Training logistics
- Epistemic dropout - create a good ensemble

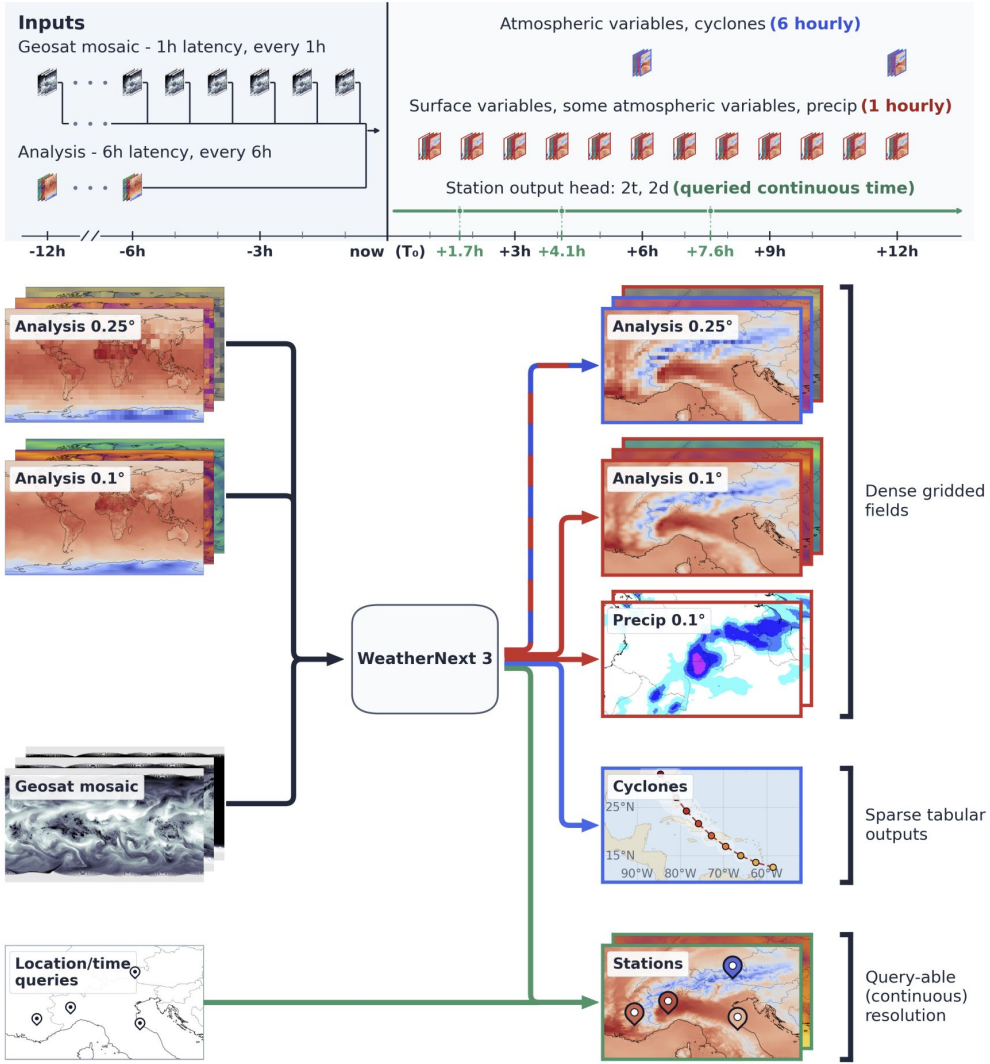


Figure 1 | **Model overview.** Schematic showing the different input and output modalities, and spatiotemporal resolutions of WeatherNext 3 (WN3). WN3 takes two sources of input: (a) two analysis frames (6 hours apart), with an operational latency of 5 hours, and (b) the 12 most recent geostationary satellite mosaic frames, which are available every hour and the most recent of which has an operational latency of just under 1 hour. Analysis inputs are 0.1° resolution for a subset of key atmospheric variables and surface variables, and 0.25° for other variables. WN3 predicts precipitation, surface variables and some atmospheric variables at hourly temporal resolution, with the remaining variables predicted in 6 hour timesteps. WN3's station output head predicts 2m temperature and 2m relative humidity at arbitrary spatial and temporal resolution (i.e. is continuously query-able in space and time). WN3 also produces off-grid sparse or tabular cyclone predictions at 6-hourly temporal resolution.

Inputs and targets

- Atmospheric analyses - ERA5 and HRES for training (1959-2015 and 2016-2023)
- Geostationary satellite mosaic - 11-channel 0.1 degree hourly mosaic from 5 satellites, infrared and visible. Helps with water vapor, temperature, and clouds.
- Precipitation - IMERG Final, PARDIG ("nevertheless, we include IMERG as well")
- In-situ surface obs - METAR airports, Mesonet stations, ICOADS ship and buoy
- Tropical cyclones - IBTrACS - positions, wind speeds, radii

Dataset	Variables	Time period	Spatial resolution	Temporal resolution
Inputs				
ERA5	13 pressure levels: *z*, *t*, *q*, *u*, *v*, *w* Single-level: *2t*, *2d*, *10u*, *10v*, *100u*, *100v*, *msl*, *sst*, ***tcc***, ***hcc***, ***mcc***, ***lcc***	**1959**-2015	0.25°	6h
HRES-fc0-5	Same as ERA5	2016-2023/**26***	0.25° + 0.1° **(Single-level variables)**	6h
Geostationary satellite mosaic	11 channels	2016-2023/**26***	0.1°	1h
Sparse decoder metadata	Elevation, land/sea mask	static	sparse training / 0.05° inference	static
Forcings	Elapsed year progress, local time of day, **tisr**	deterministic	global + 0.25° + 0.1°	6h + **1h** (tisr)
	Land/sea mask, geopotential at surface	static	0.25° + 0.1°	static
Targets				
ERA5	13 pressure levels: *z*, *t*, *q*, *u*, *v*, *w* Single-level: *2t*, *2d*, *10u*, *10v*, *100u*, *100v*, *msl*, *sst*, ***tcc***, ***hcc***, ***mcc***, ***lcc***, ***cdir***†, ***fdir***, ***ssrd***, *tp*	**1959**-2015	0.25°	6h + **1h** (single-level variables; 300/500 hPa *t*/*z*; 1000hPa *u*/*v*)
HRES-fc0-5	Same as ERA5	2016-2023/**26***	0.25° + 0.1° **(single-level variables)**	6h + **1h** (single-level variables; 300/500 hPa *t*/*z*; 1000hPa *u*/*v*)
Geostationary satellite mosaic	11 channels	2016-2023/**26***	0.1°	1h
PARDIG	1h precipitation accumulation	1999-2023/**26***	0.1°	1h
IMERG	1h precipitation accumulation	2000-2023/**25***	0.1°	1h
Surface observations	*2t*, *2d*	2016-2023/**26***	sparse training / 0.05° inference	1h
Cyclone data	Existence, minimum pressure, maximum wind, radius of max wind, wind quadrant radii at three wind speeds	1979-2023/**26***	0.25° → table	6h

Output maps from appendix! (A11-A13)

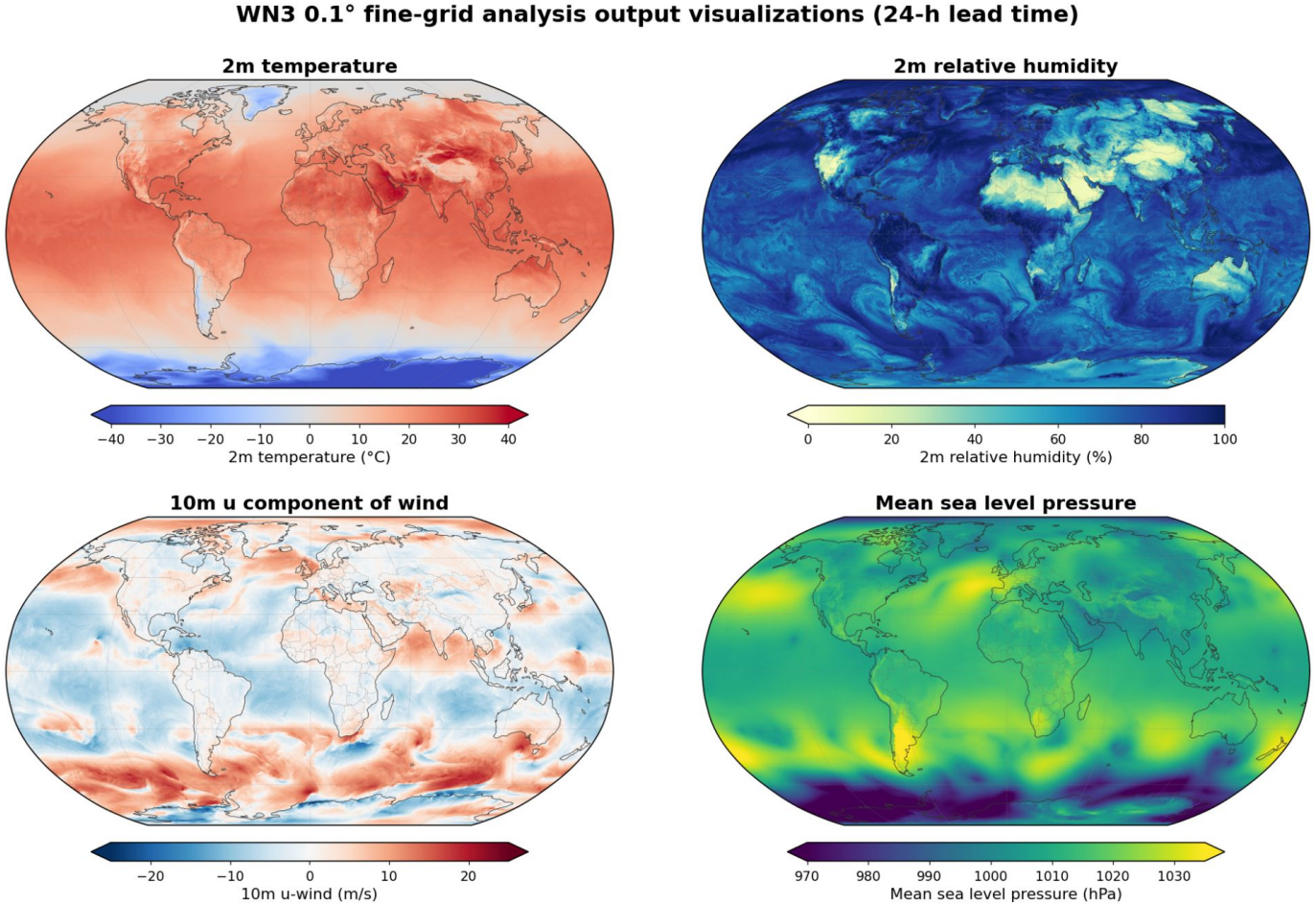


Figure A.11 | Visualizations of the 0.1° analysis output for a single sample for several single-level variables. 24h forecast initialized at July 1 2026 00 UTC.

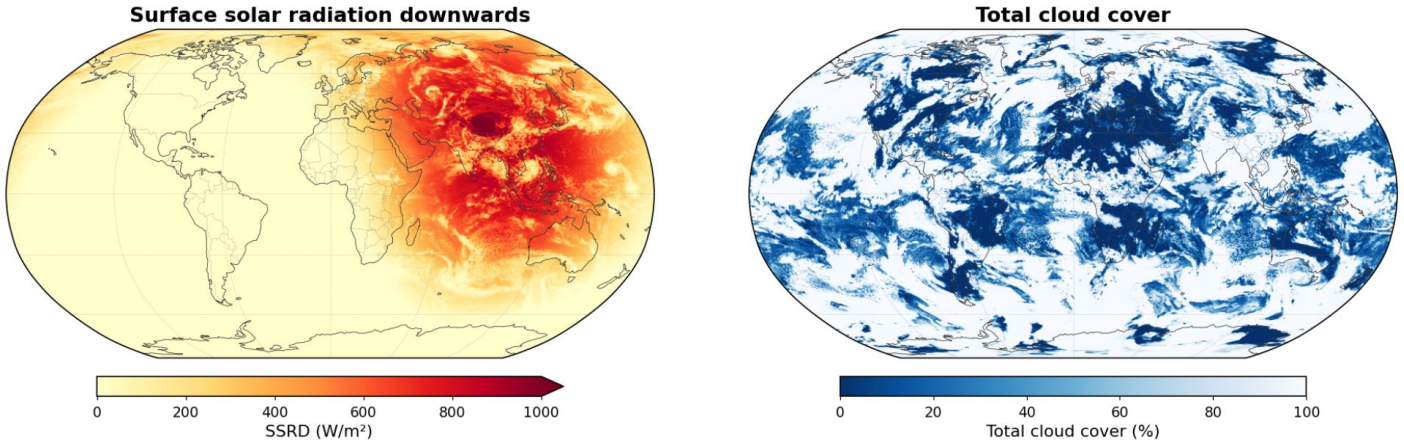


Figure A.11 | Visualizations of the 0.1° analysis output for a single sample for several single-level variables. 24h forecast initialized at July 1 2026 00 UTC.

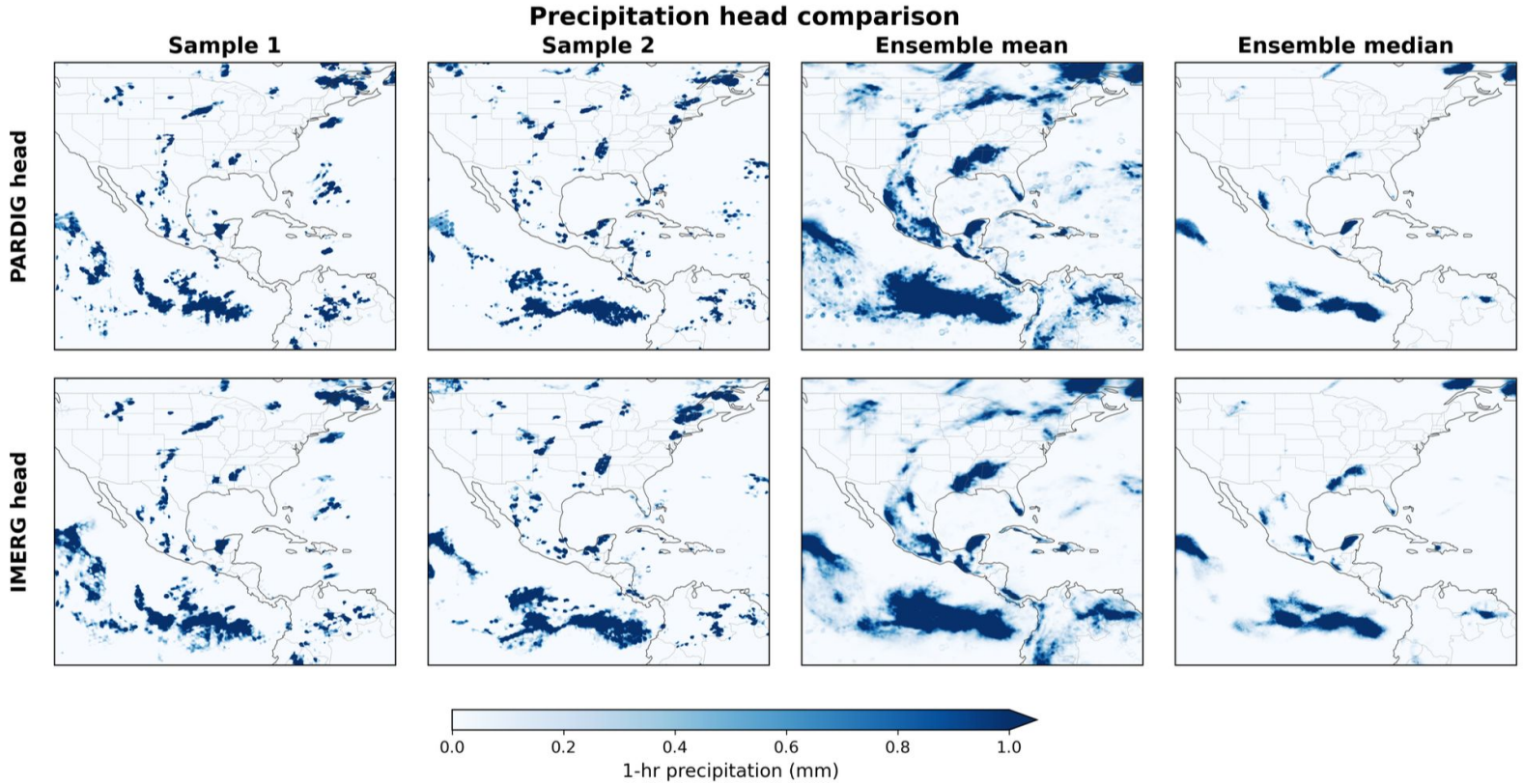


Figure A.12 | Visualizations of the PARDIG (top) and IMERG (bottom) output heads, showing two samples, the ensemble mean and the ensemble median. 24h forecast of 1h precipitation accumulation initialized at July 1 2026 00 UTC.

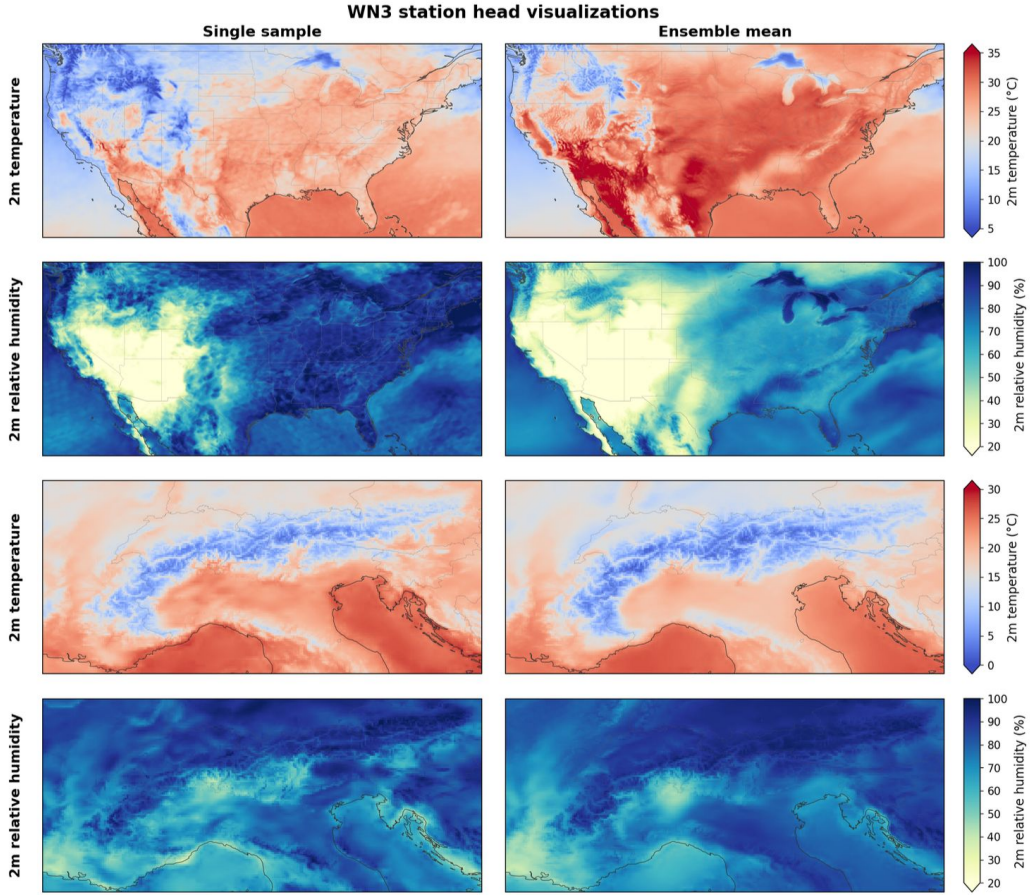
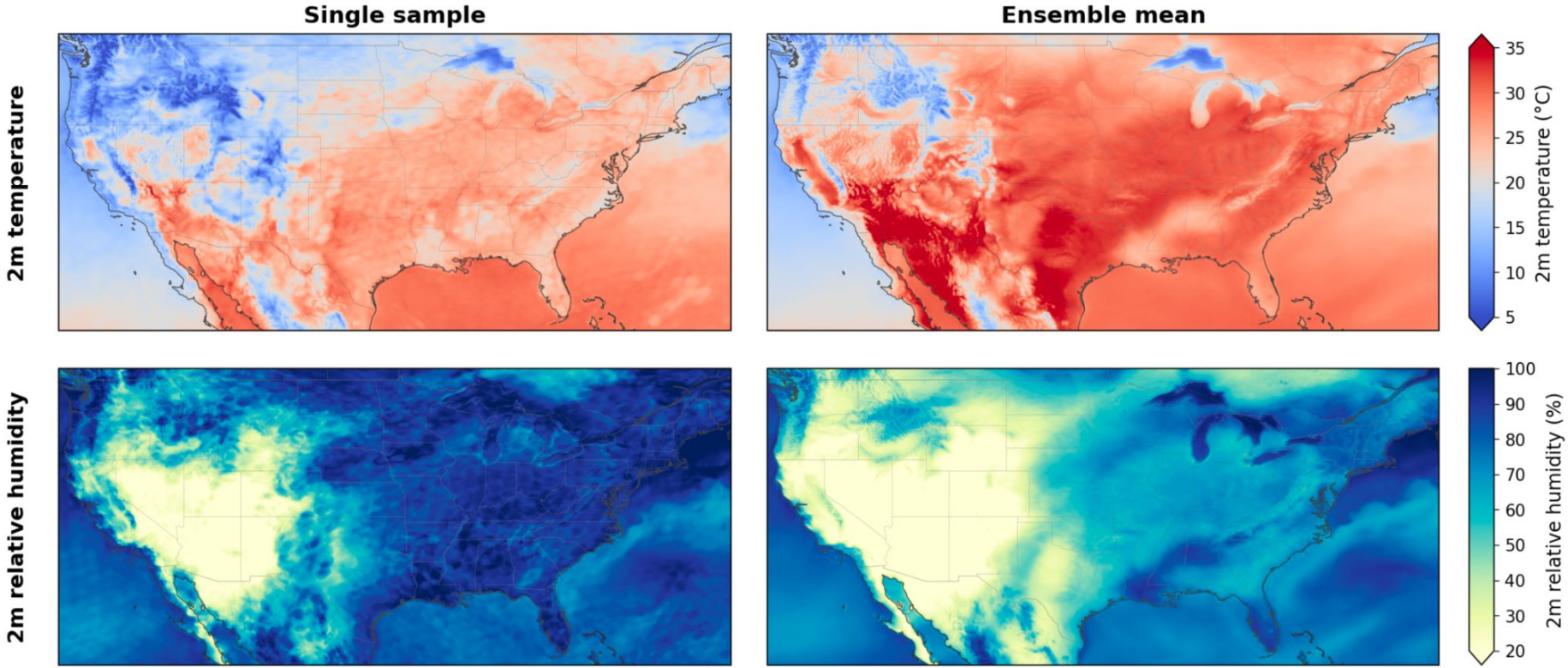
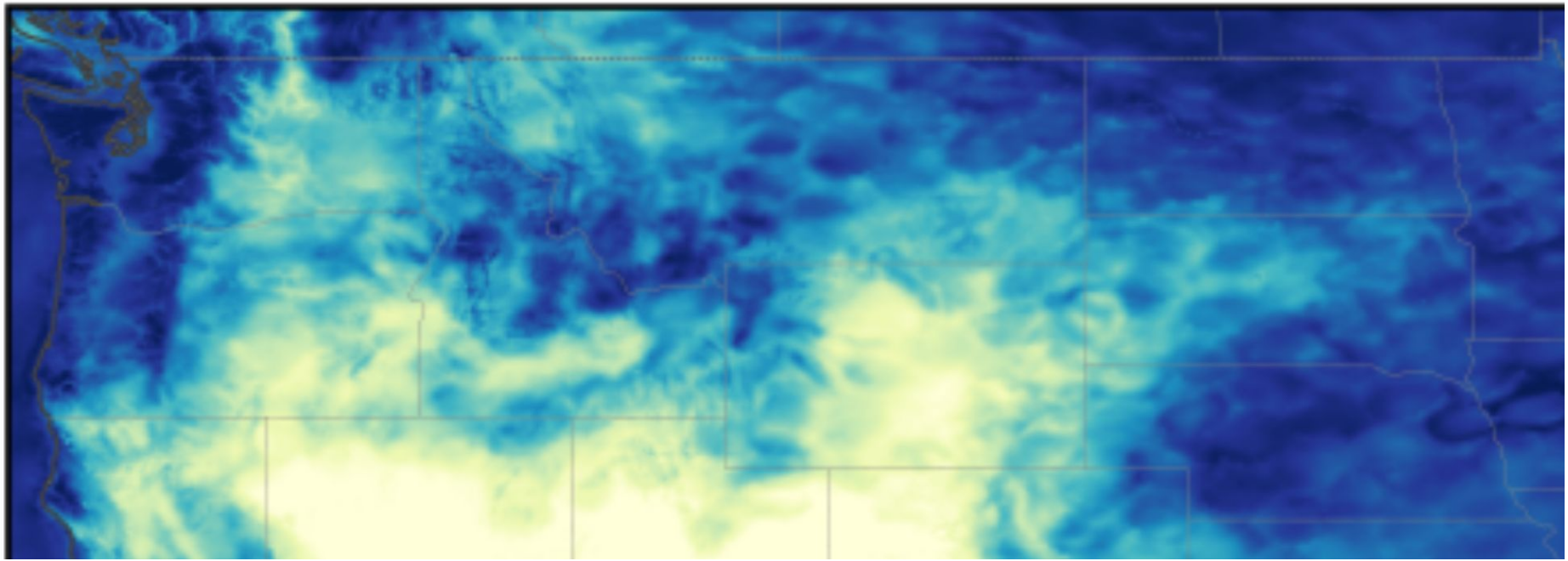


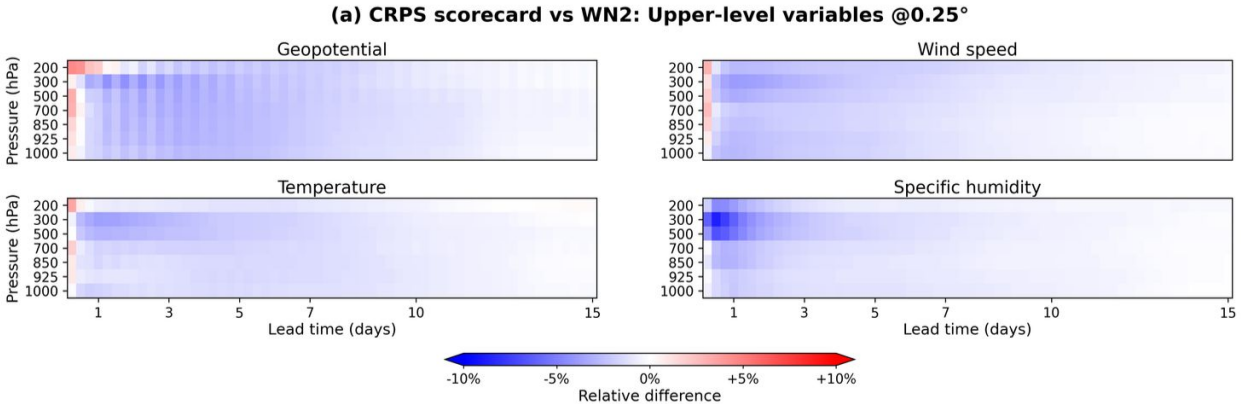
Figure A.13 | Visualizations of the 0.05° station head predictions for the United States (top two rows) and the Alpine region (bottom two rows). Left column shows an individual sample, right column shows the ensemble mean. 24h forecast initialized at July 1 2026 00 UTC.

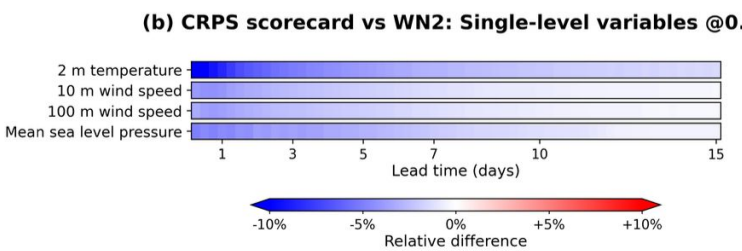

WN3 station head visualizations
Single sample
Ensemble mean
2m temperature
2m relative humidity
35
30
25
20
15
10
5
2m temperature (°C)
100
90
80
70
60
50
40
30
20
2m relative humidity (%)



Evaluation

- "Ground truth" reference products - HRES; METAR/Mesonet unseen stations; IMERG/MRMS/gauges for precipitation
- Baseline WP models - WeatherNext2, ENS, AIFS ENS v2
- Metrics - CRPS, Brier score
- Latency for operational evaluation - model's nominal k-hour lead time is an operational (k-7)-hour lead time
- Metrics - CRPS, Brier score
- Tropical cyclones - eval against IBTrACS





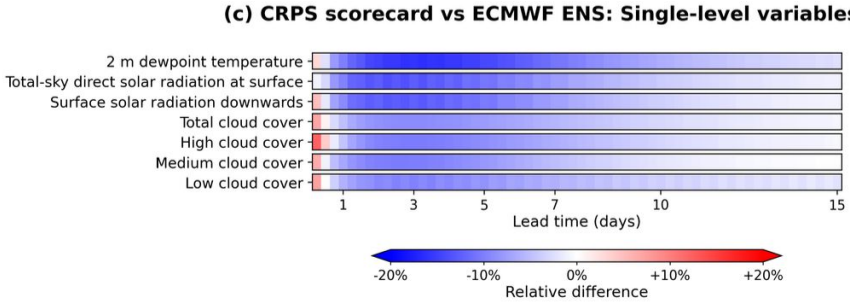


Figure 2 | **Analysis results.** (a) Upper-level CRPS scorecard comparing WN3 to the previous state-of-the-art model WN2 at 0.25°. Blue squares indicate improvement. (b) Single-level variable CRPS scorecard of WN2 vs WN3 evaluated at 0.1°. WN2 forecasts were interpolated bilinearly. (c) Scorecard of variables not in WN2 against ECMWF ENS at 0.1°. The ground truth for (a), (b) and (c) is HRES-fc0. All scores computed for 00/12 UTC initializations in 2024.

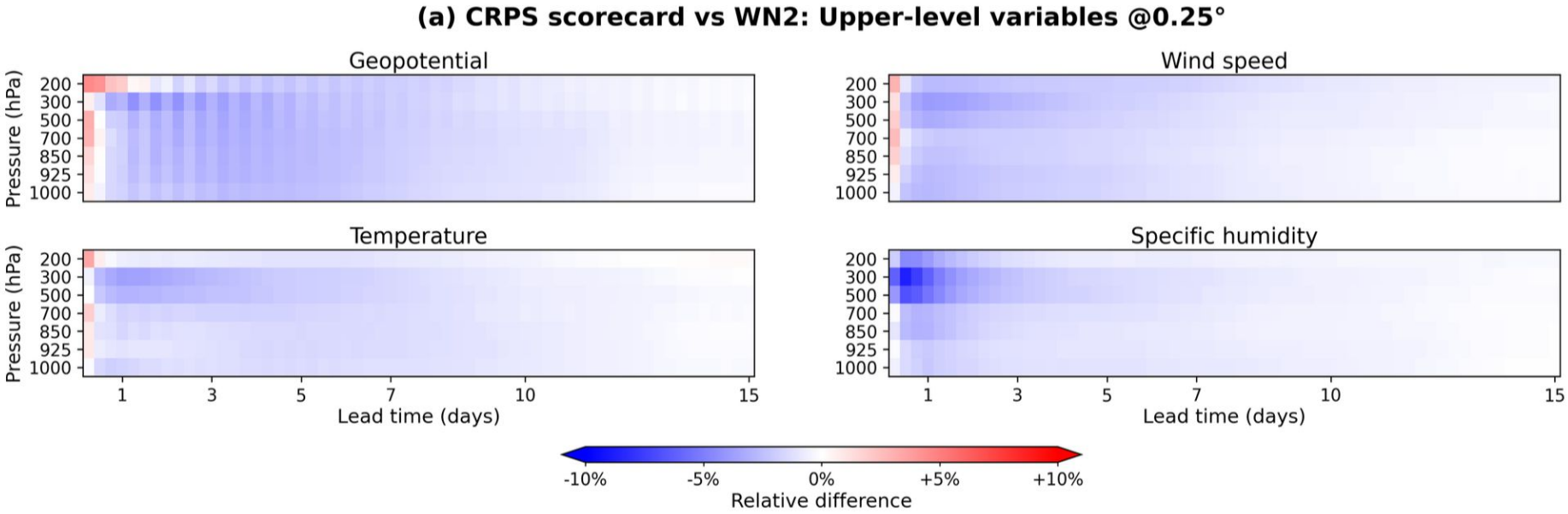
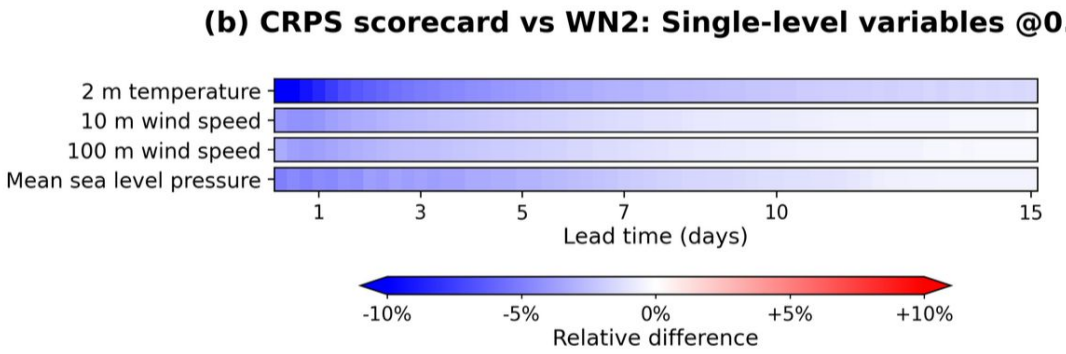


Figure 2 | **Analysis results.** (a) Upper-level CRPS scorecard comparing WN3 to the previous state-of-the-art model WN2 at 0.25°. Blue squares indicate improvement. (b) Single-level variable CRPS scorecard of WN2 vs WN3 evaluated at 0.1°. WN2 forecasts were interpolated bilinearly. (c) Scorecard of variables not in WN2 against ECMWF ENS at 0.1°. The ground truth for (a), (b) and (c) is HRES-fc0. All scores computed for 00/12 UTC initializations in 2024.



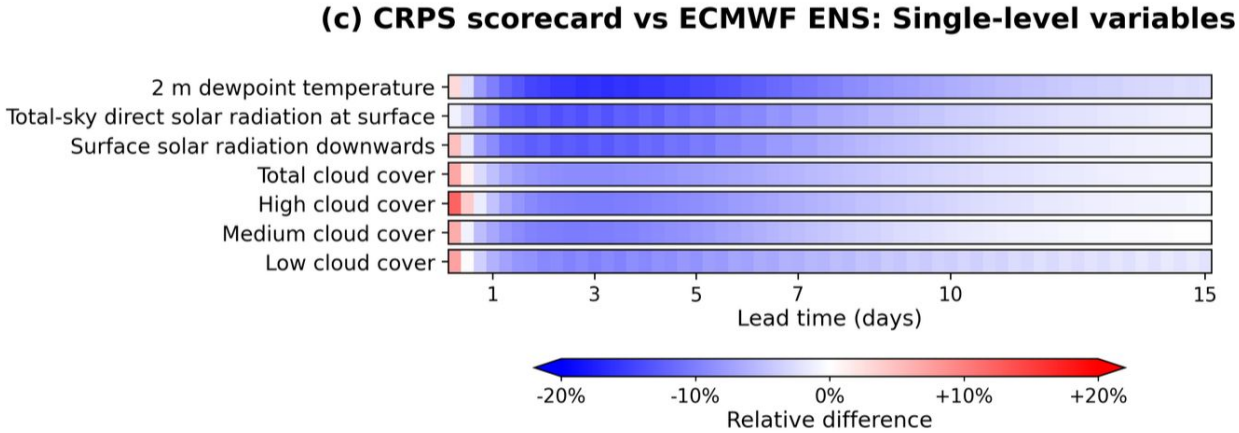


Figure 2 | **Analysis results.** (a) Upper-level CRPS scorecard comparing WN3 to the previous state-of-the-art model WN2 at 0.25°. Blue squares indicate improvement. (b) Single-level variable CRPS scorecard of WN2 vs WN3 evaluated at 0.1°. WN2 forecasts were interpolated bilinearly. (c) Scorecard of variables not in WN2 against ECMWF ENS at 0.1°. The ground truth for (a), (b) and (c) is HRES-fc0. All scores computed for 00/12 UTC initializations in 2024.

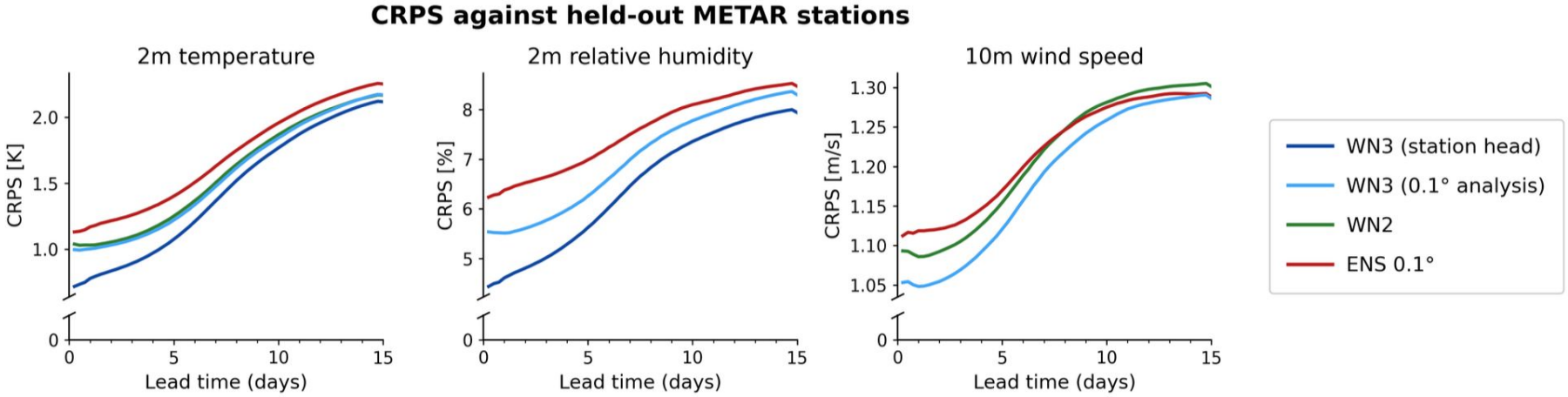


Figure 3 | **Station evaluation results.** CRPS of 2m temperature, 2m relative humidity and 10m wind speed computed against METAR weather stations that have not been used during training. WN3 station head forecasts were queried at 0.05°. All forecasts were bilinearly interpolated to the station locations. 0.25° and 0.1° 2m temperature forecasts are lapse-rate adjusted (see Section 3.1). All scores computed for 00/12 UTC initializations in 2024. For visual clarity, lines were smoothed with a 24 hour running average.

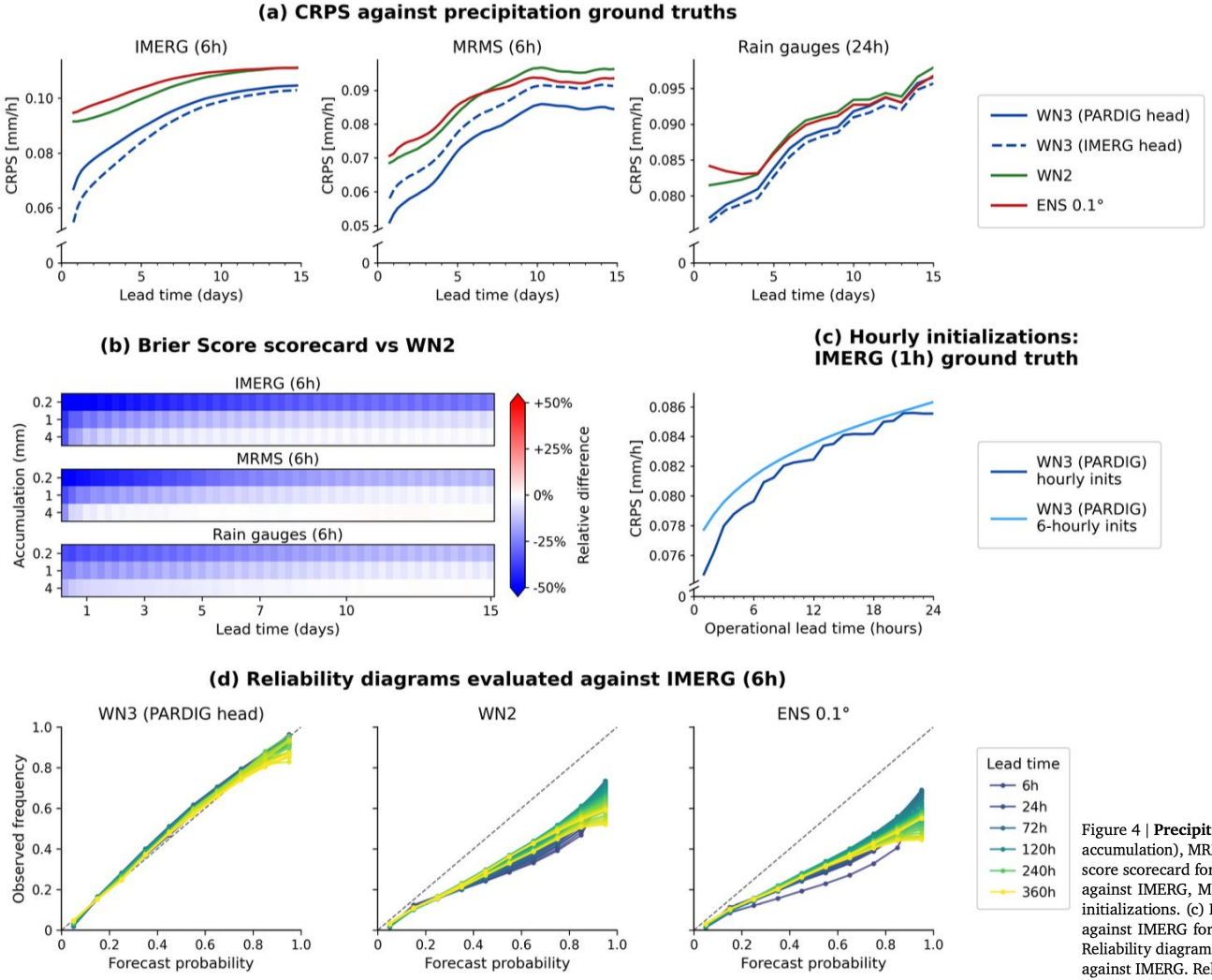


Figure 4 | **Precipitation results.** (a) CRPS of precipitation forecasts evaluated against IMERG (6h accumulation), MRMS (6h) and rain gauges (24h). Computed for 00/12 UTC initializations. (b) Brier score scorecard for the WN3 PARDIG head with respect to WN2 for different thresholds evaluated against IMERG, MRMS and rain gauges for 6h accumulation. Computed for 00/06/12/18 UTC initializations. (c) Latency-adjusted CRPS comparing hourly with 6-hourly initializations evaluated against IMERG for 1h accumulations. x-axis shows operational lead time (see Section 3.3). (d) Reliability diagrams of the WN3 PARDIG head, WN2 and ENS for 1mm / 6h accumulation evaluated against IMERG. Reliability diagrams computed for 00/12 UTC initializations.

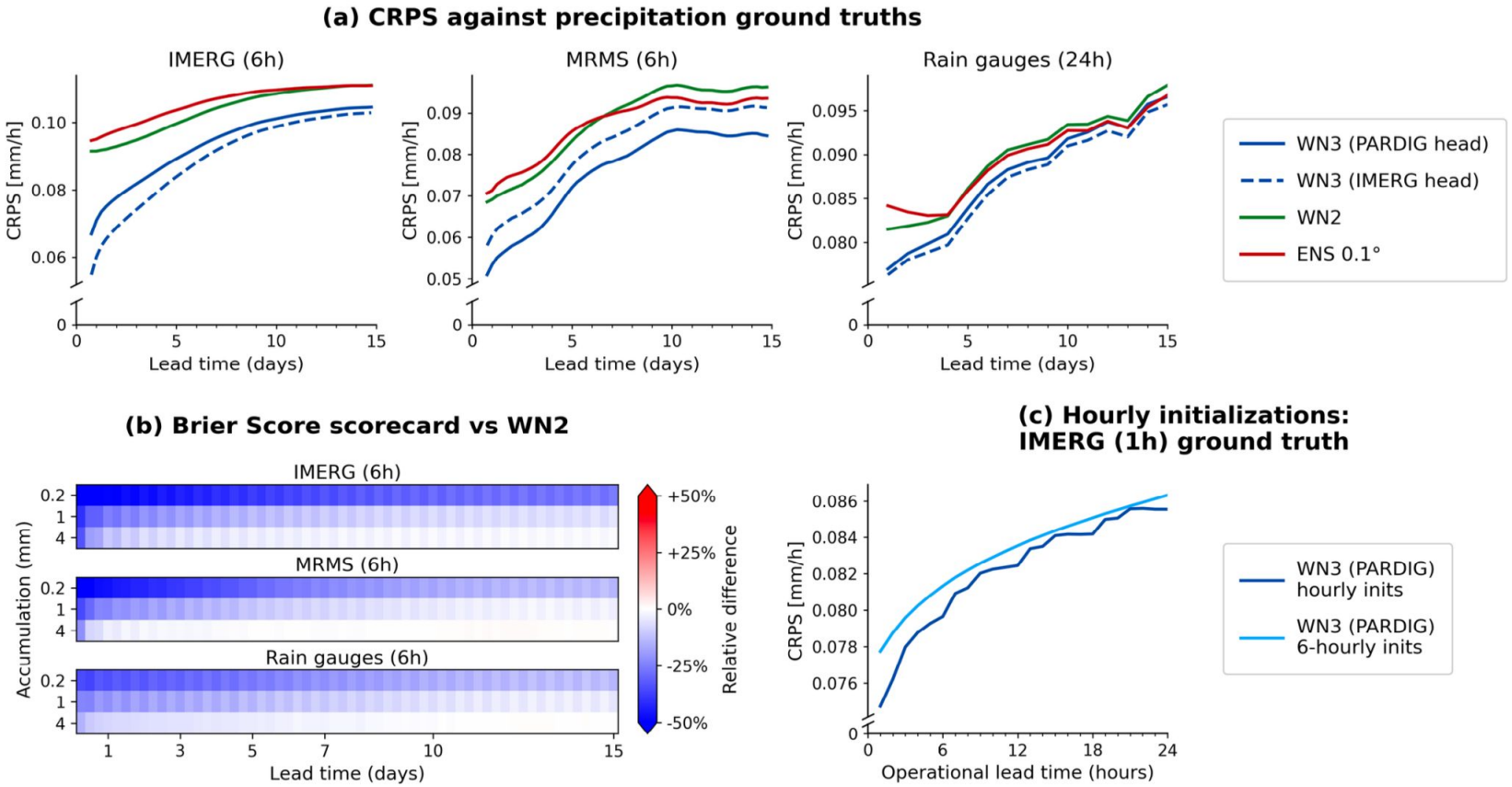


Figure 4 | **Precipitation results.** (a) CRPS of precipitation forecasts evaluated against IMERG (6h accumulation), MRMS (6h) and rain gauges (24h). Computed for 00/12 UTC initializations. (b) Brier score scorecard for the WN3 PARDIG head with respect to WN2 for different thresholds evaluated against IMERG, MRMS and rain gauges for 6h accumulation. Computed for 00/06/12/18 UTC initializations. (c) Latency-adjusted CRPS comparing hourly with 6-hourly initializations evaluated against IMERG for 1h accumulations. x-axis shows operational lead time (see Section 3.3). (d) Reliability diagrams of the WN3 PARDIG head, WN2 and ENS for 1mm / 6h accumulation evaluated against IMERG. Reliability diagrams computed for 00/12 UTC initializations.

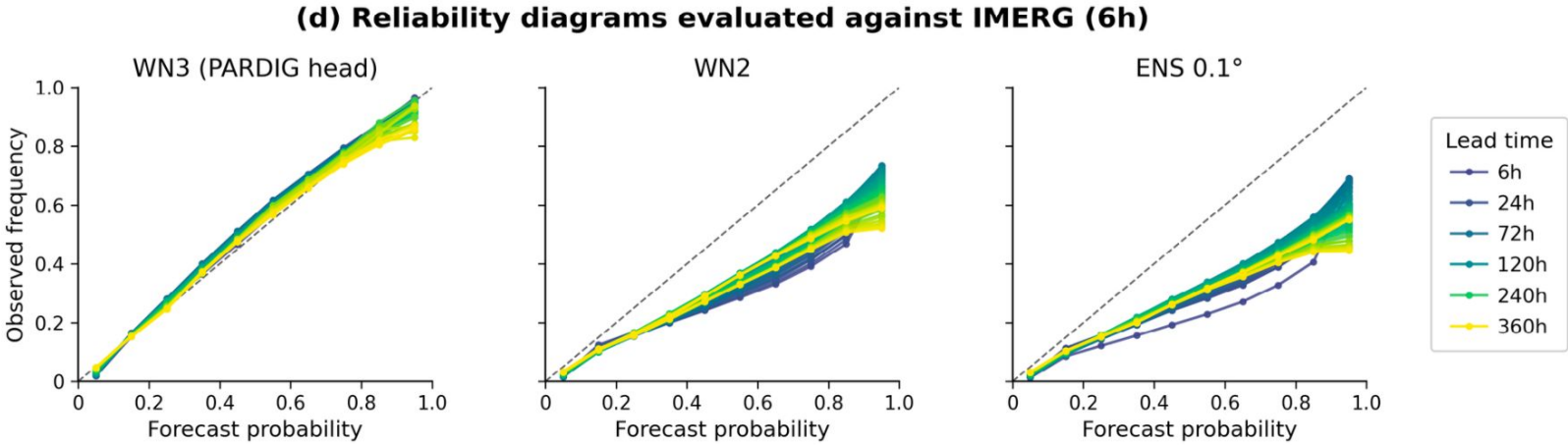


Figure 4 | **Precipitation results.** (a) CRPS of precipitation forecasts evaluated against IMERG (6h accumulation), MRMS (6h) and rain gauges (24h). Computed for 00/12 UTC initializations. (b) Brier score scorecard for the WN3 PARDIG head with respect to WN2 for different thresholds evaluated against IMERG, MRMS and rain gauges for 6h accumulation. Computed for 00/06/12/18 UTC initializations. (c) Latency-adjusted CRPS comparing hourly with 6-hourly initializations evaluated against IMERG for 1h accumulations. x-axis shows operational lead time (see Section 3.3). (d) Reliability diagrams of the WN3 PARDIG head, WN2 and ENS for 1mm / 6h accumulation evaluated against IMERG. Reliability diagrams computed for 00/12 UTC initializations.

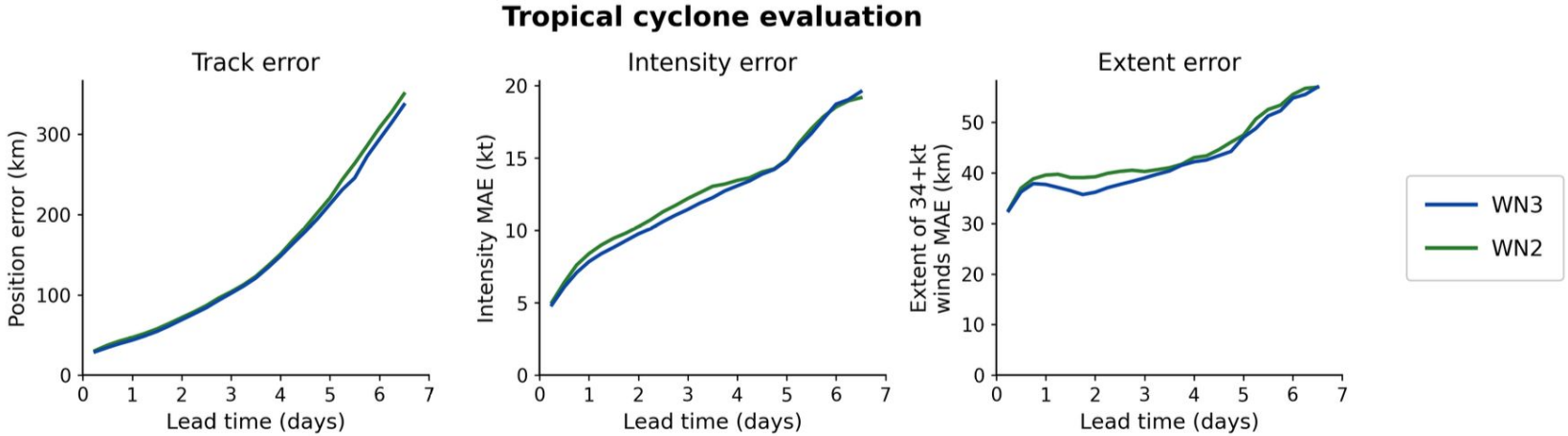


Figure 5 | **Cyclone results.** Ensemble mean track, intensity and extent error for WN2 and WN3. Intensity is defined as the maximum 1-minute sustained wind speed. Extent error is the mean absolute error of 34kt wind radii forecasts averaged over all quadrants. Evaluations for 00/06/12/18 UTC initializations in 2024. Forecasts are homogenized across WN2 and WN3, and include cyclones across all basins across the globe. We also restrict to data points classified as tropical or subtropical in nature in IBTrACS (Knapp et al., 2010; Gahtan et al., 2024), excluding extratropical or unclassified systems.

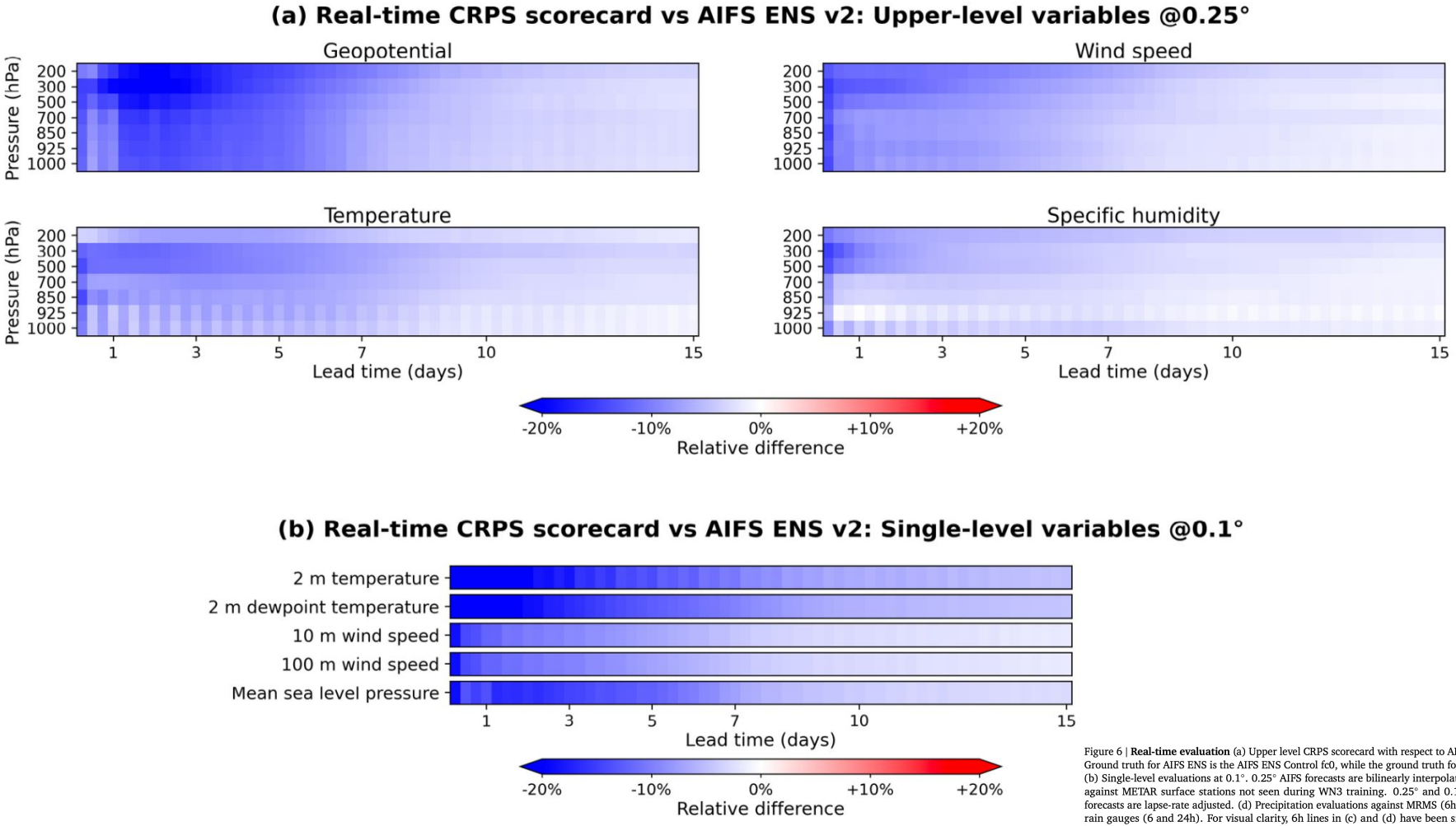


Figure 6 | **Real-time evaluation** (a) Upper level CRPS scorecard with respect to AIFS ENS v2 at 0.25°. Ground truth for AIFS ENS is the AIFS ENS Control fc0, while the ground truth for WN3 is HRES-fc0; (b) Single-level evaluations at 0.1°. 0.25° AIFS forecasts are bilinearly interpolated. (c) Evaluations against METAR surface stations not seen during WN3 training. 0.25° and 0.1° 2m temperature forecasts are lapse-rate adjusted. (d) Precipitation evaluations against MRMS (6h accumulation) and rain gauges (6 and 24h). For visual clarity, 6h lines in (c) and (d) have been smoothed with a 24 hour running average. All evaluations are for July 1 to August 11 on 00/12 UTC initializations.

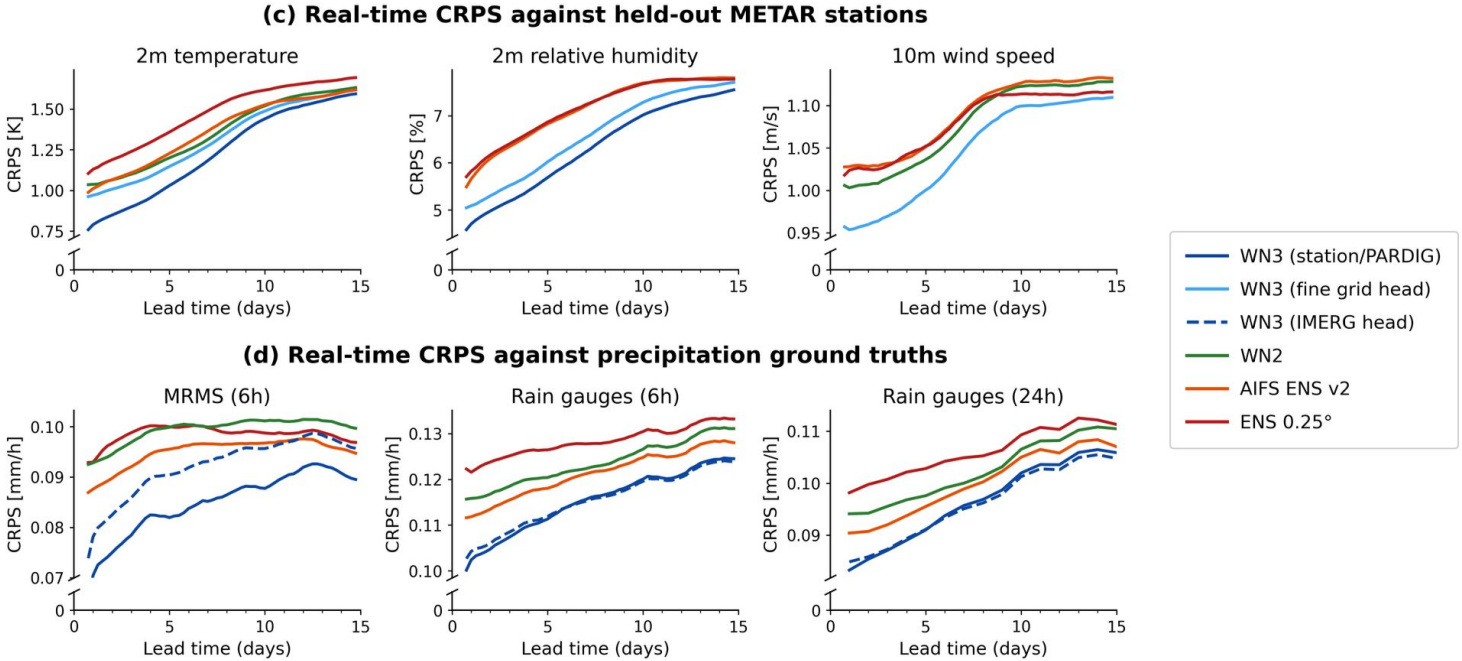


Figure 6 | **Real-time evaluation** (a) Upper level CRPS scorecard with respect to AIFS ENS v2 at 0.25°. Ground truth for AIFS ENS is the AIFS ENS Control fc0, while the ground truth for WN3 is HRES-fc0; (b) Single-level evaluations at 0.1°. 0.25° AIFS forecasts are bilinearly interpolated. (c) Evaluations against METAR surface stations not seen during WN3 training. 0.25° and 0.1° 2m temperature forecasts are lapse-rate adjusted. (d) Precipitation evaluations against MRMS (6h accumulation) and rain gauges (6 and 24h). For visual clarity, 6h lines in (c) and (d) have been smoothed with a 24 hour running average. All evaluations are for July 1 to August 11 on 00/12 UTC initializations.

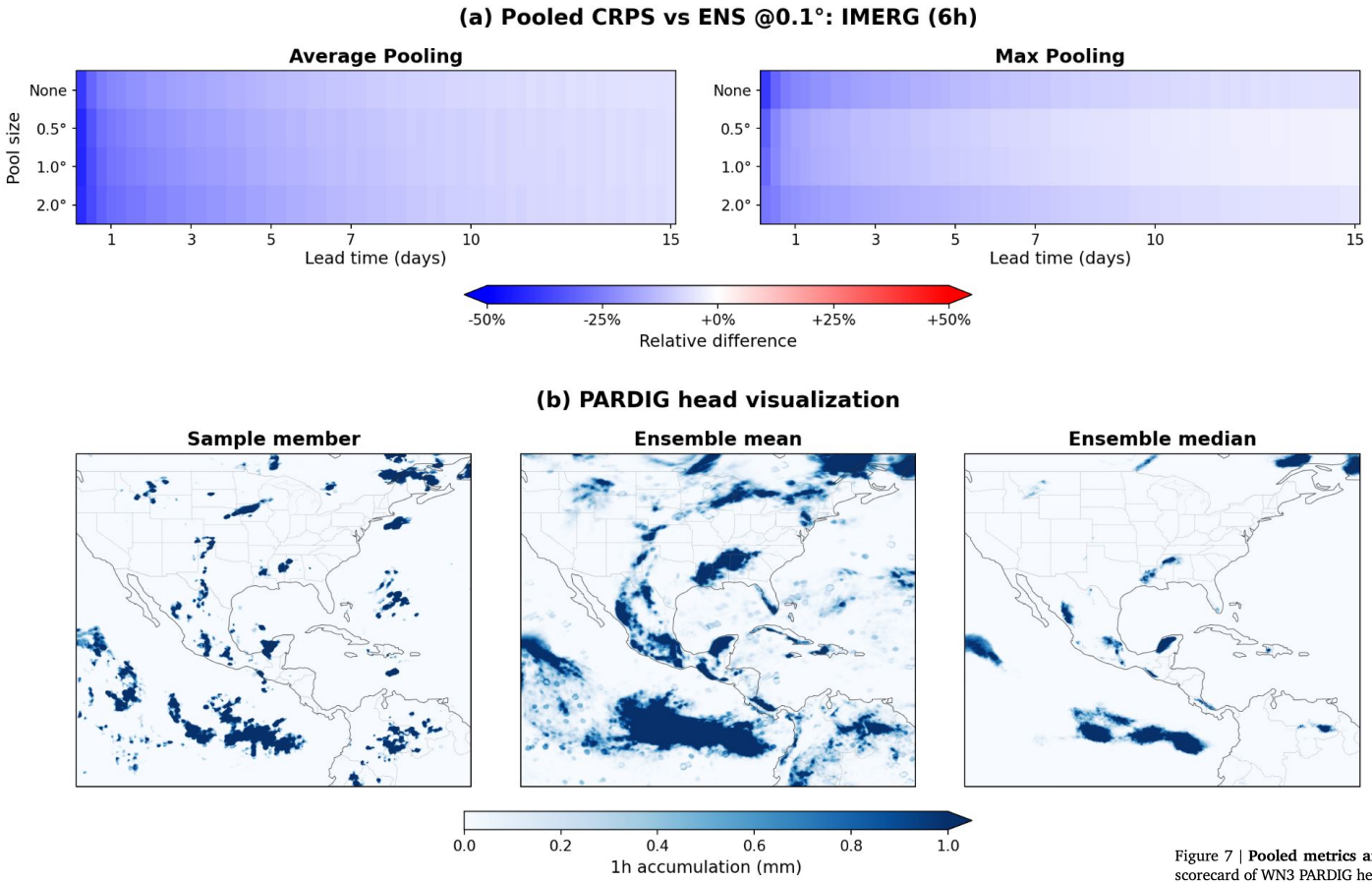


Figure 7 | **Pooled metrics and artifact visualization.** (a) 6h precipitation accumulation CRPS scorecard of WN3 PARDIG head with respect to ENS evaluated against IMERG at different pooling sizes for average and max pooling. (b) Example of artifacts in 24h PARDIG head forecasts (1h accumulation) for a single member, the 64-member ensemble mean and median for 24h forecast lead time (note that the color bar saturates at 1mm for the purposes of visualization). (c) Visualization of hourly temporal differences in the WN3 station head for 2m temperature. 12 to 13h exemplifies the transition between outer 6h time steps; (top two rows) 2 individual ensemble members; (bottom three rows) Temporal differences in ensemble quantiles ($q = 0.5$ and 0.9). All examples for 2026-07-01T00.

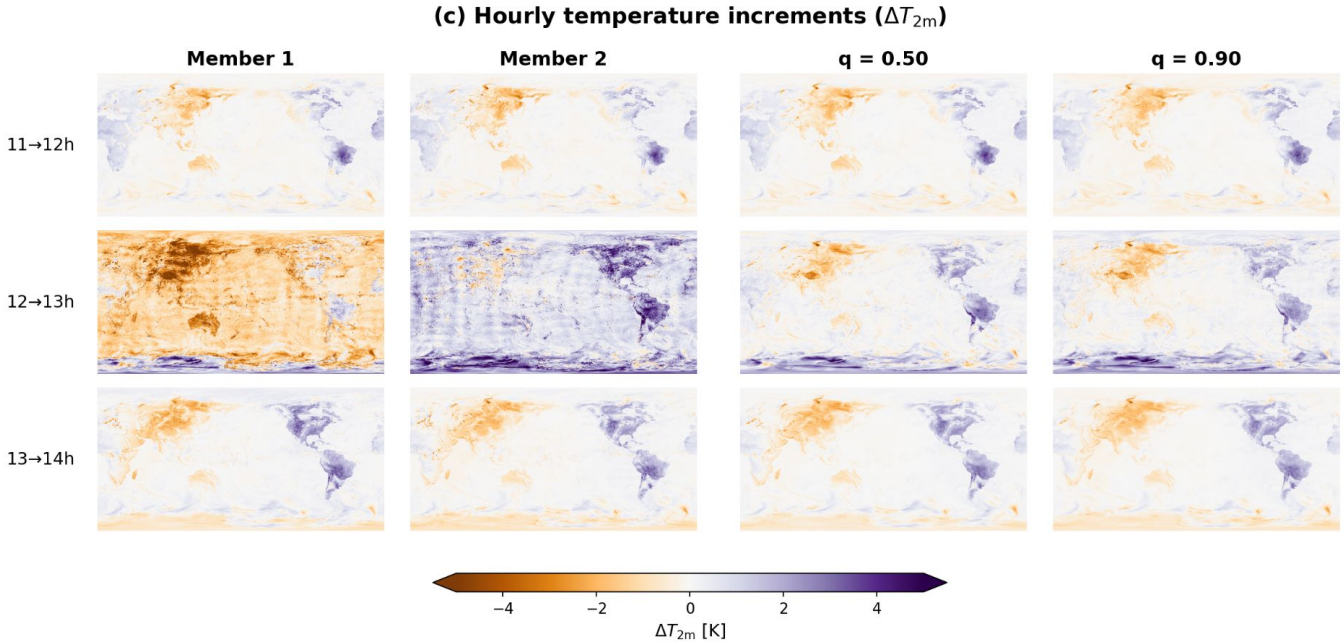


Figure 7 | **Pooled metrics and artifact visualization.** (a) 6h precipitation accumulation CRPS scorecard of WN3 PARDIG head with respect to ENS evaluated against IMERG at different pooling sizes for average and max pooling. (b) Example of artifacts in 24h PARDIG head forecasts (1h accumulation) for a single member, the 64-member ensemble mean and median for 24h forecast lead time (note that the color bar saturates at 1mm for the purposes of visualization). (c) Visualization of hourly temporal differences in the WN3 station head for 2m temperature. 12 to 13h exemplifies the transition between outer 6h time steps; (top two rows) 2 individual ensemble members; (bottom three rows) Temporal differences in ensemble quantiles ($q = 0.5$ and 0.9). All examples for 2026-07-01T00.

Atmos ML Journal Club

August Posch 5 Nov 2026

Google DeepMind

2026-9-3

WeatherNext 3: Increasing resolution and performance of global weather models with raw observations

Stephan Rasp[1*], Boris Babenko[2*], Dominic Masters[1*], Andrew El-Kadi[1*], Samier Merchant[2*], Guy Shalev[1*], Ilan Price[1*], Fred Zyda[2], Remi Lam[1], Sasha Shysheya[1], Matthew Willson[1], Stratis Markou[1], Shreya Agrawal[1], Suhani Vora[1], Mohammed Alewi Hassen[1], Sunny Mak[1], Tom R. Andersson[1], Megan Bela[3], Akib Uddin[2], Nofar Peled Levi[1], Ben Gaiarin[1], Ferran Alet[1*], Aaron Bell[2], Peter Battaglia[1] and Alvaro Sanchez-Gonzalez[1*]

[*]Equal contribution, [1]Google DeepMind, [2]Google Research, [3]Google