

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

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

\*Equal contribution, <sup>1</sup>Google DeepMind, <sup>2</sup>Google Research, <sup>3</sup>Google

State-of-the-art AI weather models have shown impressive medium-range forecast skill and computational efficiency, but suffer two key shortcomings: their forecasts have lower spatial and temporal resolution than the best physics-based models and they are exclusively initialized with and trained on analysis data. As a result, they cannot directly make use of observations, and any biases in the analysis are inherited by the forecast. WeatherNext 3 addresses these shortcomings and establishes a new state-of-the-art for probabilistic medium-range forecasting skill. First, WeatherNext 3 generates new forecasts every hour (rather than every 6 hours like traditional global models) by ingesting low-latency geostationary satellite data. Second, WeatherNext 3's temporal and spatial resolution are on par with physics-based global models, with hourly time steps and 0.1° resolution for single-level variables, including solar radiation and cloud cover. Third, WeatherNext 3 moves beyond traditional analysis variables by learning to predict satellite-derived precipitation estimates, as well as tropical cyclone and station observations. Modelling sparse station data allows WeatherNext 3 to make 2m temperature and dewpoint predictions at any location and time, conditioned on local geographical features, with substantially lower error than competing global models, even when evaluated against unseen stations. Together, WeatherNext 3's capabilities move operational AI-based weather forecasting beyond emulating the traditionally distinct stages of data assimilation, forecasting and post-processing, which helps to further push the frontier of performance and granularity for global weather prediction.

## 1. Introduction

AI weather models are revolutionizing weather forecasting, surpassing traditional models on many measures of skill, while also being faster and cheaper to run (Lam et al., 2023; Bi et al., 2023; Rasp et al., 2024; Price et al., 2025; Kossaifi et al., 2026). Several AI models now run operationally and are gaining widespread adoption (Alet et al., 2026; Lang et al., 2026). So far, most state-of-the-art global AI models have been initialized with and trained against (re-)analysis data, which are gridded estimates of the state of the atmosphere produced by assimilating a wide range of observations using a traditional numerical weather prediction (NWP) model (Hersbach et al., 2020; Soci et al., 2024).

Analyses have proven very effective for training AI models, but have shortcomings that are inherited by models trained exclusively on this data. First, analyses have known biases, particularly for key variables like precipitation or surface temperature (Lavers et al., 2022; Liu et al., 2024). Second, global operational analyses are only produced every 6 hours, assimilating observations up to 3 hours past the analysis date. Combined with a 3 hour latency in producing and publishing the analysis, this means that the most recent analysis state is always between 6 and 12 hours stale (ECMWF, 2026b). Third, some observations, such as those coming from geostationary satellites, are only indirectly assimilated (Lean and Bormann, 2023), resulting in a loss of information.

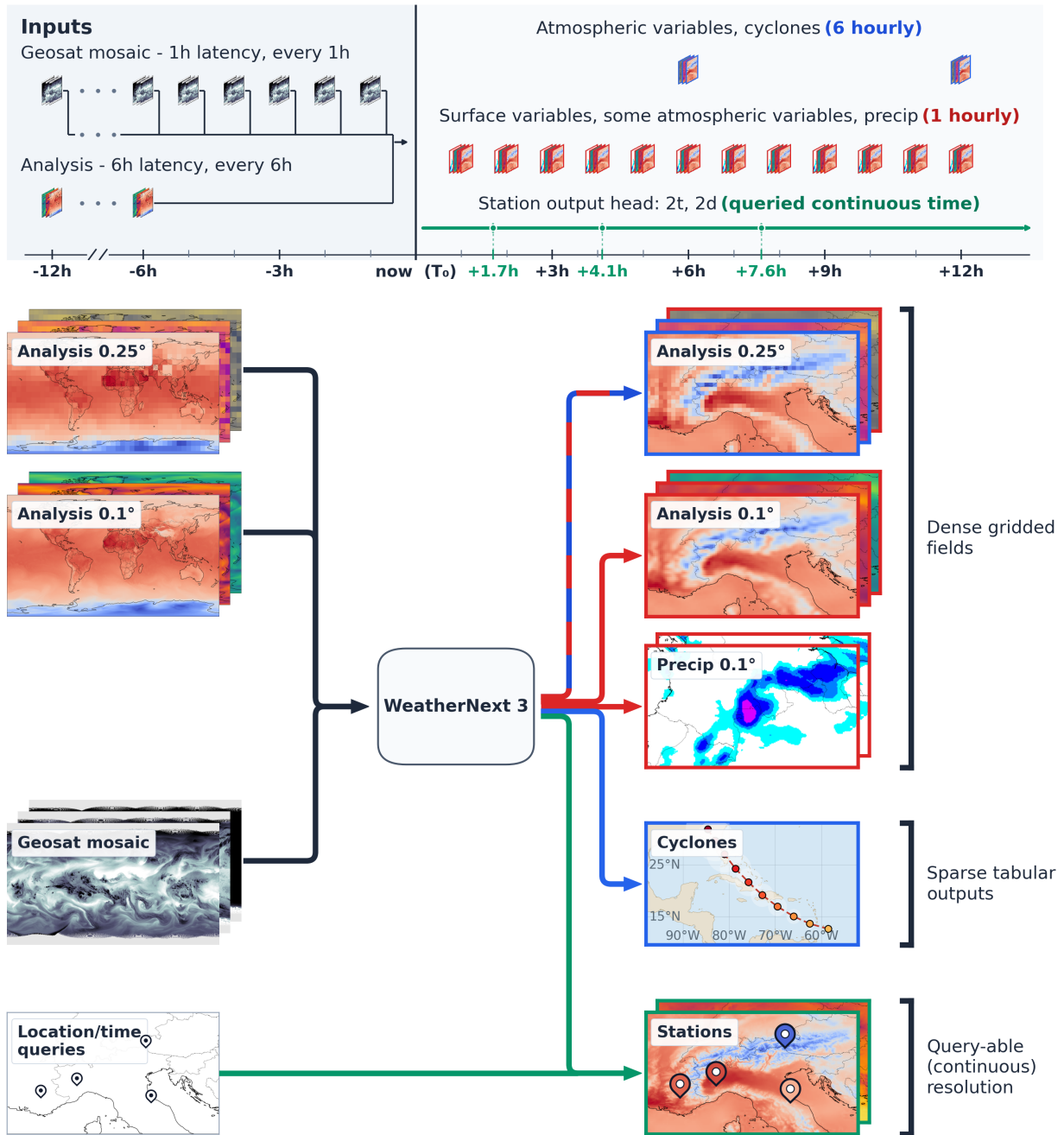


Figure 1 | **Model overview.** Schematic showing the different input and output modalities, and spatiotemporal resolutions of WeatherNext 3 (WN3). WN3 takes two sources of input: (a) two analysis frames (6 hours apart), with an operational latency of 5 hours, and (b) the 12 most recent geostationary 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.

There has been a growing interest in directly leveraging observational data in AI weather models to address these shortcomings. This was first done in regional models that ingest and predict observational data directly (Sønderby et al., 2020; Ravuri et al., 2021; Andrychowicz et al., 2023; Zhao et al., 2024, 2026). Such models have proven particularly successful at precipitation nowcasting but, because of their limited spatial context, are restricted to short-range prediction. More recently, there has also been emerging research into AI-based global data assimilation, either going from observations to analysis (Vaughan et al., 2024; Ni et al., 2025; Gupta et al., 2026; WindBorne Systems, 2026), or directly from observations to predicted observations (Pinnington et al., 2026). Results are promising, but so far most global models are not operational or have a lower spatial resolution than traditional NWP systems. We recently published WeatherNext-Cyclones (Alet et al., 2026), which combines training on global analysis with observation cyclone data, resulting in a new state-of-the-art in operational tropical cyclone prediction during the 2025 hurricane season (Cangialosi and Martinez, 2026).

WeatherNext 3 (WN3) further builds on these ideas by directly ingesting and predicting raw observations, setting a new state-of-the-art on analysis and observation benchmarks. Specifically:

- WN3 produces a new forecast every hour, ingesting the latest geostationary satellite imagery.
- WN3 produces hourly,  $0.1^\circ$  output of all single-level variables including variables that were not available in WeatherNext 2 (WN2), such as solar radiation and cloud cover.
- WN3 is trained to predict global hourly precipitation products derived from satellite data.
- WN3’s station head makes geographically-informed predictions of 2m temperature and dewpoint at any location on the globe, and any point in time, as opposed to predefined grid points and time steps in traditional models.
- WN3 substantially improves the  $0.25^\circ$  analysis and cyclone specific forecasts of WeatherNext 2 Cyclones (Alet et al., 2026).

WeatherNext 3 is running operationally. For more information on how to access the forecasts, see <https://deepmind.google/science/weathernext/>.

## 2. WeatherNext 3

### 2.1. Problem formulation.

The weather forecasting problem we address is to learn a distribution over future weather trajectories conditional on analysis and recent observations. The atmosphere at time  $t$  and spatial location  $s$  is represented by a set of  $N$  spatio-temporal fields (indexed by  $i$ ),

$$\mathbf{W}_t = \{f_i(s, t) : \mathcal{S}_i \times \mathcal{T}_i \rightarrow \mathcal{V}_i\}_{i=1}^N, \quad (1)$$

where each field  $f_i$  denotes a physical variable over a continuous or discrete spatial domain  $\mathcal{S}_i \subseteq \mathbb{S}^2$ , continuous or discrete time subset  $\mathcal{T}_i \subseteq \mathbb{R}$ , and value space  $\mathcal{V}_i$ . Whereas WN2 was exclusively modelling a single data modality, namely analysis variables at  $0.25^\circ$ , WN3 has several modality groups, within which individual variables and vertical levels are separate fields sharing a common spatial grid and temporal cadence (Table 1).

We seek to learn the conditional distribution

$$p(\mathbf{W}_{(0,T]} \mid \mathbf{W}_{\leq 0}) \quad (2)$$

over weather trajectories from the present to lead time  $T$ , where subscripts denote continuous-time intervals. We discretise the forecast into  $T/\Delta$  outer steps of width  $\Delta = 6$  h. Modelling the atmosphere as a 2nd-order Markov process, the trajectory distribution factorises as

$$p(\mathbf{W}_{(0,T]} \mid \mathbf{W}_{\leq 0}) = \prod_{t=0}^{T/\Delta-1} p(\mathbf{W}_{(t\Delta, (t+1)\Delta]} \mid \mathbf{W}_{((t-2)\Delta, (t-1)\Delta]}, \mathbf{W}_{((t-1)\Delta, t\Delta]}). \quad (3)$$

Each factor predicts the weather over a  $\Delta$ -wide window given the two preceding windows. Because some fields  $f_i$  have temporal supports finer than  $\Delta$  (e.g. hourly single-level analysis fields or precipitation estimates), the model simultaneously predicts different weather time steps within each window by treating each inner time step as its own variable.

Some of the fields are modelled as autoregressive: these are fed back as input to the next step. The remainder are *target-only*: predicted but not re-ingested (Table 1). This detail is omitted from the simplified mathematical formulation above to avoid clutter in the exposition.

## 2.2. Key features.

WN3 is based on an encode-process-decode architecture and leverages the Functional Generative Networks (FGN) approach which already underpins WN2 (Alet et al., 2025). It produces 15-day 64-member ensemble forecasts, where individual members are sampled by injecting noise into the model. The training objective is to minimize the marginal (i.e. per-modality, per-level, per-location and per-time step) continuous ranked probability score (CRPS). WN2 used this approach to achieve state-of-the-art probabilistic medium-range weather prediction at  $0.25^\circ$  spatial and 6 hourly temporal resolution. ECMWF’s AIFS ENS (Lang et al., 2026) uses a related approach with some architectural and loss differences. WN3 extends this framework and scales up the training with respect to WN2 (see Figure 1). This section describes the key features of WN3; for details refer to Section A.1.

**Multiple data modalities and resolutions.** WN3 natively handles data modalities with different spatial resolutions, temporal cadences and operational latencies. Architecturally, different spatial resolutions (such as  $0.25^\circ$  and  $0.1^\circ$  grids) are handled by separate encoders and decoders that map the data from its native resolution to the shared icosahedral processor mesh and back again. This design allows each modality to interact at its natural resolution without requiring all data to be regrided to a common grid. At the same time, it also enables dynamics to be propagated on the processor mesh independently of the native data resolution or data availability.

To fill in the intermediate hourly frames of the 6 hourly outputs, WN2 learned a separate  $6\text{h} \rightarrow 1\text{h}$  upsampler model with the same architecture but different parameters<sup>1</sup>. In contrast, WN3 predicts hourly temporal resolutions directly, making it a native feature of the model.

**Observational input and hourly initialization.** Leveraging its multimodal capabilities, WN3 now ingests geostationary satellite data directly, in addition to analysis. Because satellite data is available more frequently than the 6 hourly HRES analysis, we can generate a new forecast every hour (Section 2.3). We find that this rapid refresh yields skill improvements at short lead times, particularly for fast-evolving variables like precipitation (Section 4.3).

**Continuous output in time and space via latent interpolation.** To model sparse, i.e. non-gridded data sources, such as in-situ surface observations, we decode predictions as a continuous query. This

<sup>1</sup><https://developers.google.com/weathernext/guides/model-specs-vmg>



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

Table 1 | Table of model inputs and targets. See Section A.3 for a list of variable abbreviations. Newly added features with respect to WN2 highlighted in **bold**. (\*) Model trained until the end of 2023 for 2024 evaluations. Production model trained until June 30 2026. (†) Note that while WN3 was trained to output *cdir* as one of its output variables, during development it was observed that skill for this variable was an outlier and substantially worse than expected. We thus do not make WN3 *cdir* forecasts available operationally, and so do not include evaluations thereof in this paper.

is done by interpolating the 0.1° latent to the specific sparse observation location and combining it with relevant metadata, namely the exact station elevation, whether it is a land (station) or sea (buoy or ship) observation and the relative time offset to the outer 6 hour time step (Section A.1), before making the 2m temperature and dewpoint predictions. Since the output head is trained globally and metadata can easily be obtained for any location on Earth, this allows WN3 to make predictions at any arbitrary location and time at inference. In production inference and for evaluation in this paper, we create predictions at a 0.05° grid and at hourly intervals. Technically, it would be possible to make predictions at precise station locations, however, in practice this only made a marginal difference in our evaluations.

**Scaling model size and multi-stage training curriculum.** WN3 increases the latent size from 768 to 1024 and the mesh transformer depth from 24 to 32 layers with respect to WN2. To fit this larger

model on available accelerators, we introduce spatial sharding of the processor mesh and grid-to-mesh and mesh-to-grid gather/scatter operations (Section A.1). Further, we modify the training curriculum by progressively increasing resolution from  $1^\circ$  to  $0.25^\circ$  to  $0.1^\circ$ , followed by a frozen finetuning stage for the sparse station head. We also make use of ERA5 dating back to 1959 (from 1979 for WN2) and all of its hourly initializations. See Section A.1.3 for a detailed list of training stages including which data modalities are added at which stage.

**Epistemic dropout.** WN3 uses functional perturbation ensembles (Alet et al., 2025) to capture both aleatoric and epistemic uncertainty. Aleatoric uncertainty is modelled by perturbing the normalisation layers with a low-dimensional noise vector. In WN2, epistemic uncertainty was modelled with deep ensembles by training 4 independent model seeds. WN3 only uses 2 seeds, but adds *epistemic dropout*.

Dropout has been successfully used as a model perturbation during CRPS training to capture aleatoric uncertainty (Cachay et al., 2026; Diaconu et al., 2026). Epistemic dropout instead builds on the framing of dropout as sampling from an exponential family of networks (Gal and Ghahramani, 2016), and combines it naturally with our aleatoric uncertainty training. During inference, the dropout mask is independently sampled for each ensemble member and time step. During training, however, a single dropout mask is sampled and shared across ensemble members for that step. This prevents the samples from exploiting dropout variance, and instead effectively trains one of the dropout subnetworks. This naturally expands the spread of evaluation samples relative to training, compensating for the fact that the model learns to calibrate its ensemble spread based on its training skill, which in practice is better than its evaluation skill. Given the increased model capacity and smaller number of model seeds, this proved important in preserving ensemble spread.

### 2.3. Model inputs and targets

A list of all data used in WN3 is provided in Table 1.

**Atmospheric analysis.** We use a combination of ERA5 (Hersbach et al., 2020) and HRES (ECMWF, 2023) for training WN3. In prior work, we used HRES-fc0 which is available 6 hourly. For hourly initialization and hourly targets we use the most recent HRES forecasts in-between the 6 hour intervals. This is denoted as HRES-fc0-5 throughout the text. For the  $1^\circ$  and  $0.25^\circ$  training stages, we use ERA5 from 1959 to 2015, and HRES-fc0-5 from 2016 onward. For the  $0.1^\circ$  variables, we train exclusively on HRES-fc0-5 starting in 2016 (details in Section A.1.3). Note that the model treats the  $0.25^\circ$  analysis (3D and single-level fields) and the  $0.1^\circ$  analysis (only single-level fields) as separate modalities. Single-level variables are predicted at hourly temporal resolution as separate channels with time offsets  $\{-5, -4, -3, -2, -1\}$  h relative to the 6-hourly analysis prediction. In addition, some pressure level  $0.25^\circ$  variables (300/500 hPa  $t/z$ ; 1000hPa  $u/v$ ) are also predicted at hourly resolution.

For the model used throughout most of the evaluations in this paper, we train until the end of 2023. For the production model, we train until June 30 2026. ECMWF released a new model cycle (50r1) on May 12 2026<sup>2</sup>. To provide WN3 with as much data of the new cycle as possible, we also use the experimental 50r1 data available from January 20 to May 11 2026.

**Geostationary satellite mosaics.** WN3 uses an 11-channel  $0.1^\circ$  hourly mosaic of geostationary satellite imagery produced by combining data streams of several satellites (Section A.3). Geostationary satellites cover the infrared and visible spectrum, providing important information on water vapor,

<sup>2</sup><https://www.ecmwf.int/en/newsletter/185/earth-system-science/upgrade-ifs-cycle-50r1>

temperature and the location of clouds. Our mosaic has an operational latency of roughly 1 hour, giving us data which is at least 5 hours more recent than the most recent analysis. To exploit this we feed geostationary satellite data with time offsets  $\{0, 1, 2, 3, 4, 5\}$  h with respect to the analysis state to the model. In order for the model to be able to propagate this information over many rollout steps, we predict the geostationary channels autoregressively.

As an example of how these time offsets play out in practice, at 14 UTC, the latest geostationary mosaic is from 13 UTC, so the model input contains data from  $\{08, 09, 10, 11, 12, 13\}$  UTC. The latest available HRES analysis at this point is from 06 UTC, so we input the 2h HRES forecast, valid at 08 UTC.

**Precipitation estimates.** In addition to analysis precipitation, which was the sole precipitation product predicted by WN2, we added two additional, more specialized precipitation products to WN3: a novel experimental precipitation estimate, called PARDIG, that is based on a separate AI model trained to predict sparse space-borne radar data from the GPM core observatory<sup>3</sup>; and the 'Final' version of NASA's Integrated Multi-satellitE Retrievals for GPM (IMERG), a widely used global precipitation product (Huffman et al., 2023). Our internal analysis shows that PARDIG matches independent observations from ground-based radar and rain gauges more closely than IMERG. Nevertheless, due to the popularity of IMERG we include IMERG-based predictions as well. Both products have a spatial resolution of  $0.1^\circ$ . PARDIG is available from 1999, while IMERG availability starts in June 2000 (currently IMERG Final is only available until September 2025). See Section A.3 for a detailed description of both products.

**In-situ surface observations.** We combine three surface observation datasets for model training, starting in 2001: METAR and Mesonet (both from MADIS<sup>4</sup>) and ICOADS (Freeman et al., 2017). METAR is a set of roughly 5,000 airport weather stations, while Mesonet is a collection of high-density regional station networks, primarily in Europe and North America. The number of Mesonet stations changes with time. In 2024, it was roughly 15,000. ICOADS is a dataset of ship and buoy measurements. The quantity of observations fluctuates greatly over time but, on average in 2024, they number around 3000 per hour. See Section A.3 for more details. We hold out a random but temporally consistent 5% of METAR and Mesonet stations from training, which we can then use for independent evaluation of spatial generalization. The station head is trained to predict 2m temperature and dewpoint. 10m wind speed observations are used for evaluation only.

Figure A.1 highlights the substantial gaps in spatial coverage of surface in-situ observations. As a result we noticed strong biases in our model predictions in sparsely observed, out-of-distribution areas, such as the Andes or Himalayas, and some high-latitude ocean regions. To alleviate these issues, we add "pseudo-station" data by randomly sampling 2,000 locations per hour and filling these with interpolated values from the corresponding variable in ERA5 or HRES-fc0-5. This did not result in a degradation of the predictions when evaluated against hold-out stations but greatly reduced biases.

**Tropical cyclones.** Our model directly learns to predict the existence, position and attributes of tropical cyclones, as done in Alet et al. (2026). We similarly convert centre positions, maximum sustained wind speeds  $V_{\max}$ , and 34/50/64-knot wind radii from the International Best Track Archive (IBTrACS, Knapp et al., 2010) into discrete gridded targets. We make only very minor modifications to this process, documented in Appendix A.3.8. Additionally, we interpolate IBTrACS to hourly resolution before griddification to obtain cyclone targets for training with hourly initializations.

---

<sup>3</sup><https://gpm.nasa.gov/missions/GPM/core-observatory>

<sup>4</sup><https://madis.ncep.noaa.gov/>

### 3. Evaluation methodology

Our evaluation philosophy is to identify the current state-of-the-art model for each task and to use this as a primary baseline for our scorecards. This follows our previous work (Price et al., 2025; Alet et al., 2025, 2026). Since WN3 has an extended set of capabilities, no single baseline model covers all evaluations. Similarly, for each task we identify the appropriate independent ground truth(s) against which to evaluate. When dealing with different forecast resolutions, we interpolate lower resolution forecasts to the ground truth resolution. While this introduces some subtleties, we believe that this ultimately is the best representation of actual forecast skill. Evaluations are done for the full year of 2024, using a model trained until the end of 2023, except for the real-time evaluation, which uses a model trained until June 30 2026. Aside from the latency-adjusted evaluation, all reported lead times are relative to nominal initialization (e.g. for WN3 this corresponds to the nominal time of the input analysis).

#### 3.1. Ground truths

**Analysis.** To evaluate the  $0.25^\circ$  and  $0.1^\circ$  analysis predictions we use HRES-fc0 as ground truth, which is also what WN3 is initialized from. One exception is the evaluation of AIFS ENS, for which we use the AIFS ENS control fc0 as ground truth, which derives from the same raw data as HRES-fc0 but has differences in interpolation (see Section A.3 for details). We note that ECMWF typically evaluates against HRES-an, which involves an additional post-processing step that mainly affects surface temperature. Evaluating directly against weather stations as an independent ground truth allows us to bypass this ambiguity. Given the absence of hourly HRES-fc0 and HRES-an, we can only evaluate WN3’s hourly analysis forecasts against HRES-fc0-5, which is done for  $\leq 2$  day lead times in Figure A.3.

**Weather stations.** We use the set of METAR and Mesonet stations not seen during training for independent evaluation. To be consistent, this also applies to wind speed measurements, which were not explicitly used for training. We use bilinear interpolation to go from model grid to station location. This also applies to the WN3 station head predictions which were queried at a  $0.05^\circ$  grid and then interpolated. Using the nearest grid point led to worse results, with bigger differences for the coarser forecasts. Following ECMWF, for 2m temperature we apply a lapse rate adjustment to account for differences between the model grid and station elevation (Ingleby, 2015). However we only apply this to  $0.25^\circ$  and  $0.1^\circ$  forecasts. For our  $0.05^\circ$  station head predictions, we only saw a negligible effect on metrics and decided to show the raw scores. We also evaluate relative humidity, which is computed from the temperature and dewpoint forecasts and observations.

**Precipitation.** We evaluate against three precipitation ground truths: IMERG Final, the Multi Radar Multi Sensor (MRMS) Quantitative Precipitation Estimate Pass-2 product (Smith et al., 2016) and rain gauges. Evaluating against a range of independent precipitation products is important since no single product is perfect and they disagree a fair amount (Moazami and Najafi, 2021). By choosing three datasets with different trade-offs, we aim to provide a representative picture of precipitation forecast performance.

IMERG provides a precipitation estimate based on data from the GPM satellite constellation. It has the benefit of being available globally at  $0.1^\circ$  resolution. However, for WN3 it is not an independent ground truth since IMERG is used as a model target. This naturally applies to the IMERG output head but could also indirectly influence the PARDIG head since both are co-trained end-to-end. Note that IMERG Final is unavailable for the real-time evaluation.

MRMS is primarily based on ground radar in the contiguous United States. The Pass 2 precipitation product is additionally calibrated against rain gauge measurements. MRMS is considered a high-quality dataset, more accurate than e.g. IMERG, with the obvious downside of limited geographic coverage. MRMS natively has a resolution of  $0.01^\circ$ . For computational reasons we subsample to  $0.05^\circ$  for evaluation.

Rain gauges are yet another independent data source. We obtain quality controlled rain gauge data from Synoptic’s derived precipitation service<sup>5</sup>. Rain gauges provide reliable observations, yet suffer, just like weather stations, from an unequal global distribution (Figure A.1). For all evaluations against rain gauges the predictions were interpolated bilinearly to the station locations.

### 3.2. Baselines

**WeatherNext 2.** For analysis evaluation, our key baseline is WeatherNext 2 (WN2, Alet et al., 2025) which represents the previous state-of-the-art in probabilistic medium-range forecasting. For the 2024 evaluations, we use a model version that has been trained until the end of 2023. For the real-time evaluation we use the publicly available WeatherNext 2 forecasts, which are based on a model trained until the end of 2025.

**ECMWF ENS.** ENS has long been the gold standard in physics-based, ensemble medium-range prediction. Here we use ENS primarily as a baseline for variables not available in WN2. ENS has a resolution of  $0.1^\circ$ . We downloaded different subsets of ENS forecasts from different sources, and have noticed some discrepancies in performance depending on the source (see Section A.3.1 for details). This slightly penalizes ENS in some evaluations. However, since the gap between ENS and the AI-based models is generally quite large, we do not expect this to affect the key conclusions of our evaluations. For the real-time evaluations we use data from the ECMWF open-bucket available at  $0.25^\circ$ . This is clearly denoted in the legends and captions.

**ECMWF AIFS ENS v2.** AIFS ENS (Lang et al., 2026) is another strong global AI-weather model. It has been upgraded to v2 on May 12, along with the switch to the 50r1 cycle (ECMWF, 2026a). AIFS ENS consists of 50 members initialized from the perturbed ENS initial conditions, plus a control member initialized from the unperturbed initial conditions. For our real-time evaluation we use data from the ECMWF open-bucket that has been regridded from AIFS ENS’s native N320 reduced Gaussian grid to  $0.25^\circ$ . Note that an hourly interpolation model was recently released for AIFS (Ingstad et al., 2026). However, these forecasts are not yet readily available, so we were not able to evaluate them.

### 3.3. Metrics

We use the continuous ranked probability score (CRPS, Ferro et al., 2008) as our main evaluation metric. Lower CRPS values indicate better forecasts. See Section A.2.1 for details.

**Precipitation evaluation.** Precipitation evaluation is a complex topic without a single agreed-upon best score to determine the "best" precipitation forecasts. CRPS serves as our lead metric, similar to ECMWF’s headline metrics (Haiden et al., 2025). To measure probabilistic skill at different precipitation rates, we additionally use the Brier Score (Glenn et al., 1950). Note that we cap our evaluations at moderate accumulations of 4 mm/6h. For more extreme values, true positives become

---

<sup>5</sup><https://synopticdata.com/>



very sparse, making reliable evaluation difficult. For this reason, we leave the evaluation of extreme precipitation for future work. We also compute reliability diagrams (Murphy and Winkler, 1977), which measure how well predicted probabilities correspond, on average, with the actual occurrence frequency of events. Well calibrated forecasts should lie on the 1:1 line. Forecasts below the diagonal represent underdispersive, overconfident forecasts. Finally, to assess precipitation skill for larger spatial regions, we use a pooled version of CRPS, where we either compute the score for the average or maximum values in a given region. This is similar to the pooled evaluations in Price et al. (2025) and Alet et al. (2025) but the pooling method differs slightly (Section A.2.2).

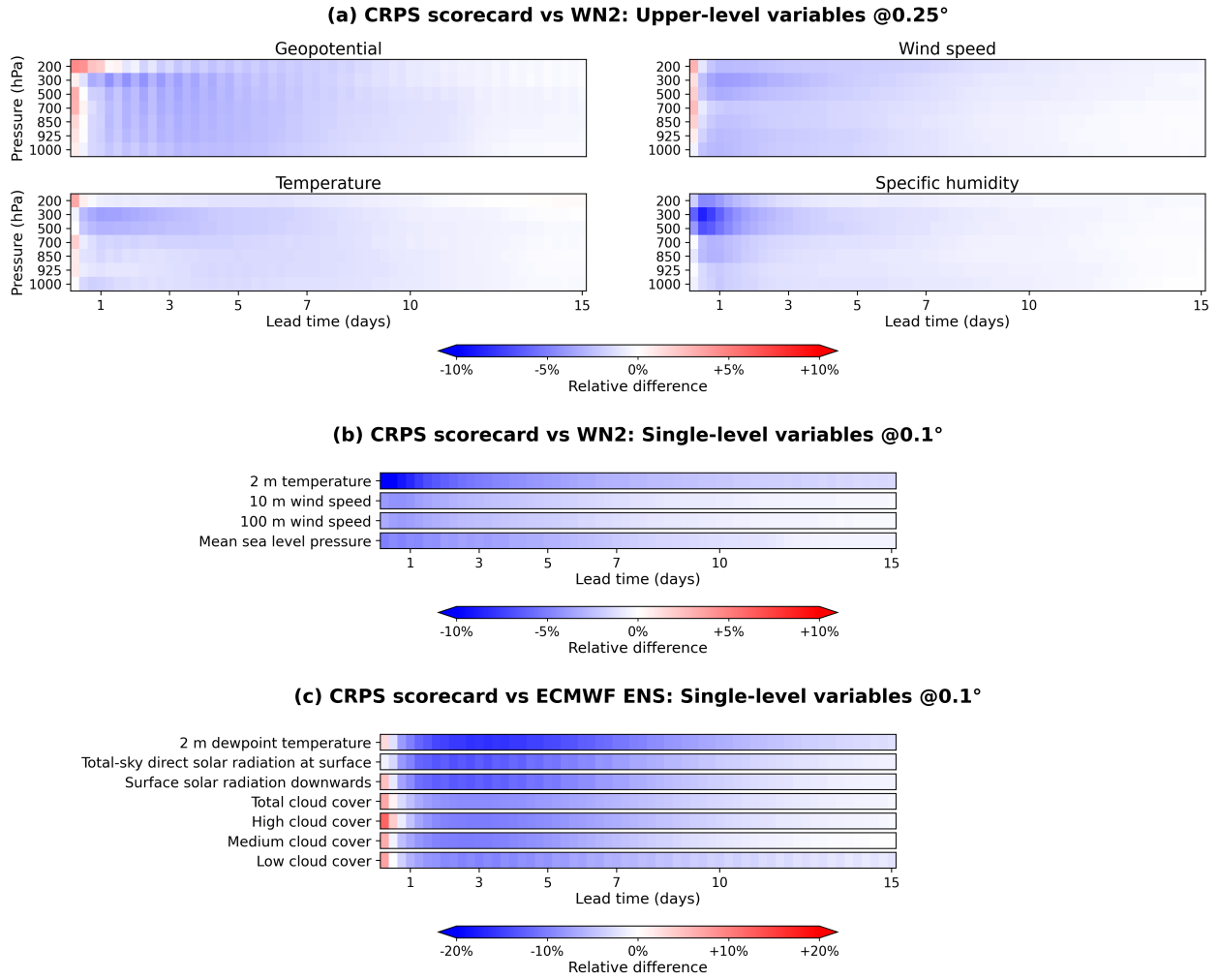
**Latency-adjusted evaluation.** WN3 produces a new forecast every hour, increasing update cadence and reducing operational forecast age compared to conventional models that run on 6-hour cycles. To quantify the skill improvement enabled by this higher frequency, we evaluate precipitation performance against an identical model constrained to 00/06/12/18 UTC initializations (denoted "6-hourly inits" in Figure 4c). Here, rather than looking at lead times relative to nominal initialization time, we look at operational lead times, where  $t = 0$  denotes the time at which the forecast is operationally available. We assume a latency of 7 hours (6 hours for raw input availability plus  $\leq 1$  hour for model execution and dissemination) for WN3.

For example, consider an operational lead time of 2 hours at a decision time of 16 UTC (targeting valid time 18 UTC). The most recent WN3 forecast has a nominal initialization at 09 UTC (which takes as input HRES-fc3 initialized at 06 UTC, as well as geostationary satellite mosaics at 09 through 14 UTC), so the operational lead time of 2 hours corresponds to the model's nominal lead time of 9 hours. On the other hand, assuming 6-hourly initializations, the most recent forecast has a nominal initialization at 06 UTC, requiring a nominal lead time of 12 hours to target valid time 18 UTC. In this example, therefore, the hourly-initialized WN3 has a 3 hour advantage.

**Cyclones evaluation.** For our cyclones evaluation we evaluate against the IBTrACS database as ground truth (Knapp et al., 2010; Gahtan et al., 2024). WN3 and WN2 are both global models, so our evaluation includes cyclones from all basins. Following the evaluation protocol of Alet et al. (2026) and the National Hurricane Center verification report (Cangialosi and Martinez, 2026), we: (i) perform a homogeneous evaluation, whereby we only include a datapoint if both WN3 and WN2 predict it, and it also exists in the ground truth; (ii) include only datapoints where the cyclone nature was classified as tropical or subtropical, excluding disturbances, extratropical and unclassified systems. In addition, we account for the 6-hour lateness of the models' forecasts by performing the same 'earlification' process from Alet et al. (2026), taking the forecast initialized from the previous 6-hour cycle, and applying a correction term to it based on operationally available TCVitals data (NCEP, 2025).

## 4. Results

WN3 shows state-of-the-art performance across a variety of ground truths. In this section, we first focus on evaluations spanning a full year, 2024, starting with analysis scorecards, followed by evaluations against observational targets: weather stations, precipitation and tropical cyclones. Then, to evaluate the skill of our operationally running model version against baselines such as AIFS ENS, we show evaluations for a 6 week period in 2026. Finally, we discuss evaluations of the spatial-joint forecast distribution and artifacts in the model output.



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

#### 4.1. Analysis evaluation

WN3 outperforms the previous state-of-the-art global weather model WN2 on a majority of upper-level targets (Figure 2a). Particularly large improvements can be seen for mid-to-upper-level humidity which hints at the additional information provided by the geostationary satellite imagery. For single-level variables, we evaluate against  $0.1^\circ$  data, which WN3 directly predicts, while we bilinearly interpolate the  $0.25^\circ$  WN2 forecasts. WN3 outperforms WN2 across all surface variables, with the largest improvements seen for 2m temperature (Figure 2b). WN3 still outperforms WN2 when only evaluated on common ( $0.5^\circ$ ) grid points, albeit by a smaller margin (Figure A.2). This shows that the improvement comes from both better general model skill and increased resolution. Comparisons against ENS show even larger improvements (Figure A.2). For variables that were not available in WN2, we use ENS as our primary baseline (Figure 2c). Here as well, WN3 shows substantial improvements in CRPS skill across all variables. Note that for cumulative variables (ssrd and fdir) we are using the 6h WN3 variables for this evaluation (Section A.1.1).

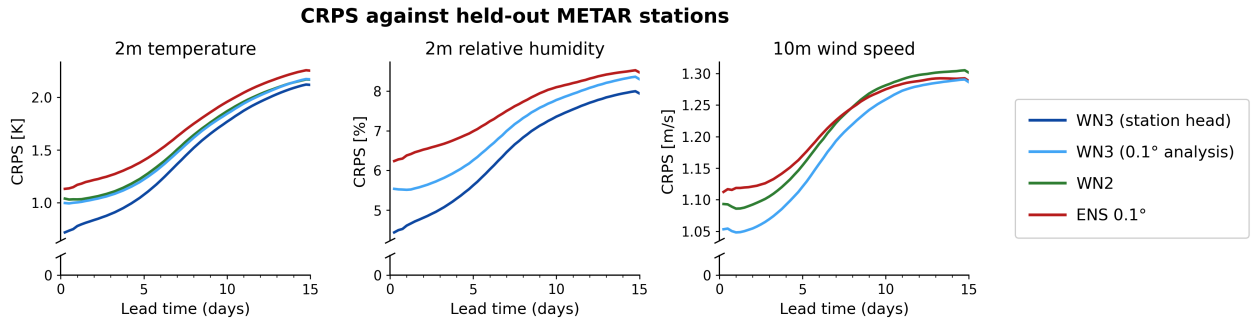


Figure 3 | **Station evaluation results.** CRPS of 2m temperature, 2m relative humidity and 10m wind speed computed against METAR weather stations that have not been used during training. WN3 station head forecasts were queried at  $0.05^\circ$ . All forecasts were bilinearly interpolated to the station locations.  $0.25^\circ$  and  $0.1^\circ$  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.

The only noteworthy exceptions to WN3’s superior performance on analysis forecasting are the 6h (and occasionally 12h) lead times for a subset of variables. Further research and ablations are needed to definitively understand the causes of this phenomenon. However, we note that due to the roughly 7h latency in global forecast systems, nominal 6-hour forecasts are never actually used operationally, and so 6h performance has little effect in practice.

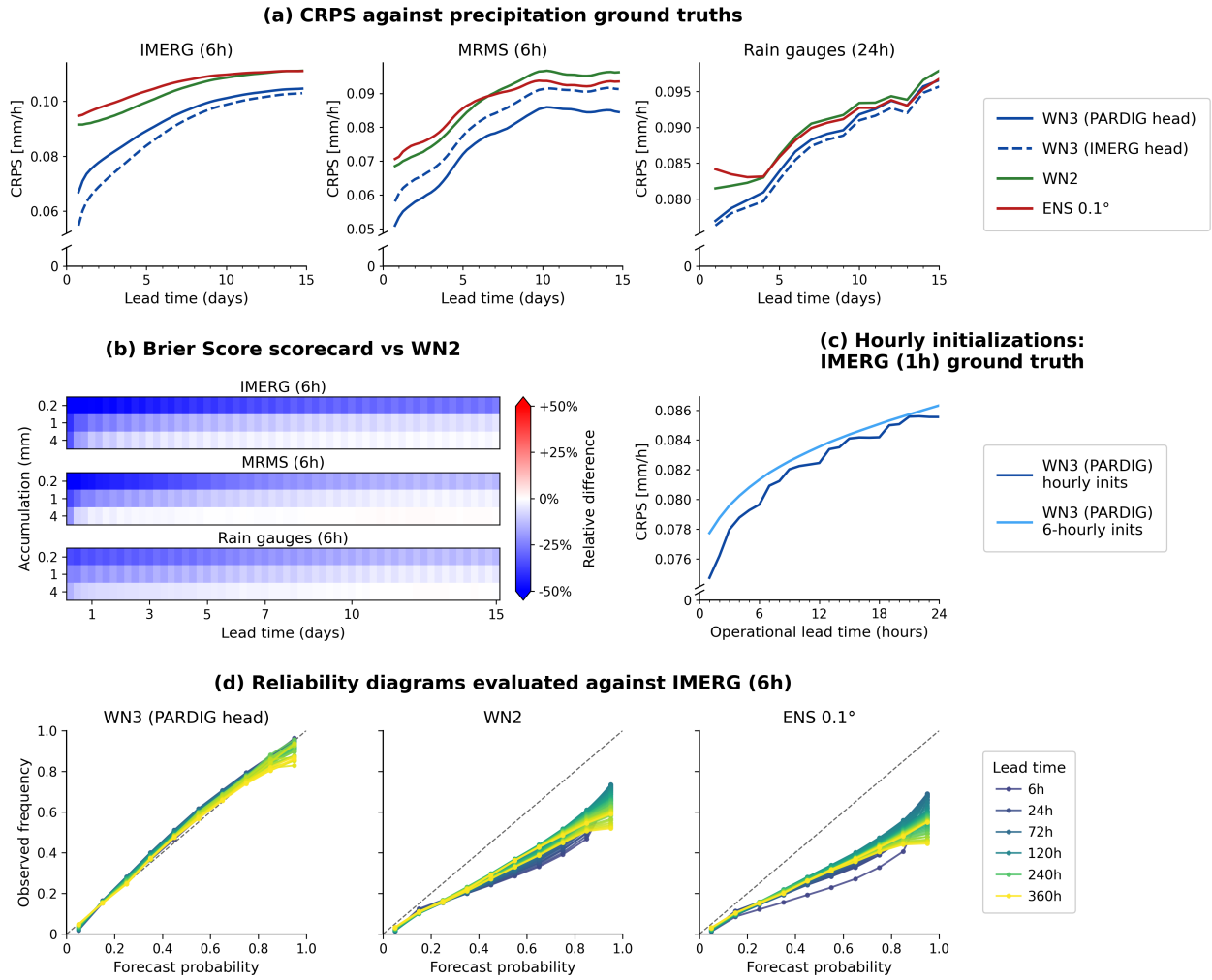
The medium-range improvements of roughly 5% on upper-level variables with respect to WN2 approximately translate to 6 hours of additional forecast lead time at the same level of skill. Traditionally, weather forecast skill has progressed by one day of forecast skill per decade (Bauer et al., 2015). This improvement over WN2, published a bit more than a year ago, demonstrates the continued accelerated pace of progress enabled by AI-weather models. However, one of WN3’s major contributions is to go beyond a pure focus on analysis-based medium-range skill towards observations.

## 4.2. Weather station evaluation

Evaluation against global METAR weather station data shows that WN3’s station head predictions have substantially smaller error compared to analysis-based forecasts (Figure 3, and Figure A.4 for evaluation of hourly lead times  $\leq 2$  days). For short lead times, the station head improves CRPS for 2m temperature by up to 30% compared to WN2 and 40% compared to ENS. Similar improvements with respect to ENS hold for 2m relative humidity (WN2 does not predict 2m dewpoint temperature). All evaluations shown here are based on stations not seen during training. The station head does overfit to training stations but only by a few percent, indicating strong spatial generalization (Figure A.5).

The WN3  $0.1^\circ$  analysis predictions also show strong performance, demonstrating the benefit of increased resolution. For 10m wind speed, WN3 leads to a 5% reduction in CRPS at early lead times. Note that because of the noise in the wind speed observations, the base error is quite high compared to the increase in the forecast error with lead time.

As mentioned in Section 3.1, we apply a lapse rate-adjustment to the  $0.25^\circ$  and  $0.1^\circ$  2m temperature predictions. This improves the coarse forecasts the most. For our  $0.05^\circ$  station head predictions, however, this additional post-processing turned out to be unnecessary, which highlights the ability of WN3 to combine traditionally separate stages of the weather forecasting pipeline into a single model.



**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.

### 4.3. Precipitation evaluation

WN3's PARDIG head predictions greatly improve precipitation skill when evaluated against a range of ground truths. Evaluations against global probabilistic baselines WN2 and ENS show a reduction in CRPS by up to 60% for IMERG, 30% for MRMS and 10% for rain gauge measurements for early lead times (Figure 4a). The IMERG head naturally shows lower errors when using IMERG as a ground truth, but performs worse against MRMS, and similarly against 24h rain gauge measurements. More fine-grained analysis using the Brier Score for different precipitation rates in Figure 4b shows that improvements are most pronounced for low precipitation rates around the rain/no-rain threshold; Figure A.6 shows comparisons that include ENS and both WN3 precipitation heads. For higher rates absolute Brier Score values become very small and less reliable at longer lead times, highlighting the

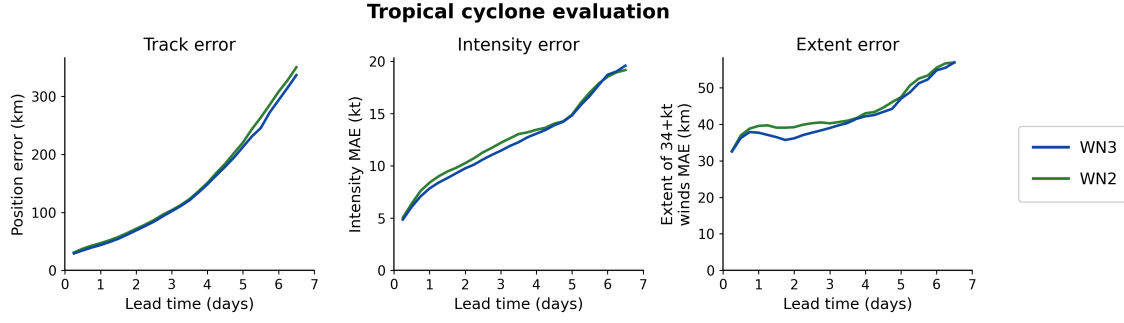


Figure 5 | **Cyclone results.** Ensemble mean track, intensity and extent error for WN2 and WN3. Intensity is defined as the maximum 1-minute sustained wind speed. Extent error is the mean absolute error of 34kt wind radii forecasts averaged over all quadrants. Evaluations for 00/06/12/18 UTC initializations 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.

need for more specific precipitation metrics.

For precipitation forecasts, it is especially important that forecasts are well calibrated, meaning that predicted probabilities of rain, on average, match the actual occurrence frequency. Reliability diagrams for the WN3 PARDIG head, as well as WN2 and ENS, show that WN3 is better calibrated against IMERG compared to previous global models, which are overconfident in their predictions (Figure 4d). Notably, calibration also remains stable throughout the 15 day forecast. The pronounced underdispersion of ENS at 6 hours lead time is likely a symptom of model spin-up.

Finally, we focus on the impact of hourly initializations enabled by using low-latency geostationary satellite observations (Figure 4c and Figure A.7). For this we compare the latency adjusted skill of the 6-hourly WN3 PARDIG head predictions to the hourly updates. The hourly forecasts show improved skill, particularly for early lead times. The lead time gained is 2–3 hours, which is in line with expectations since the average latency advantage of hourly over 6-hourly initializations is roughly 3 hours. This gain is important for quickly developing storms.

#### 4.4. Tropical cyclone evaluation

We follow the evaluation protocol of Alet et al. (2026) and compare WN3 against WN2 in a global homogeneous evaluation. WN2 is a modified version of WeatherNext Cyclones (WN-C) from Alet et al. (2026) with the same training protocol and nearly identical training data, except it also uses 100-metre wind as inputs and targets. Apart from this difference, the performance of WN2 and WN-C are extremely similar on both atmospheric and cyclone variables, and were the state-of-the-art in AI-based cyclone forecasting in the 2025 season. Figure 5 shows relatively small, but consistent improvements in mean absolute error of both the ensemble mean track and intensity forecasts. Extent improvements are more sizeable between 1- and 3-day lead times. At the same time, we also observed that WN3 tended to be generally more under-spread than WN2 on both intensity and extent (Figures A.8 and A.9). This under-spreading may be due to the larger size of the model which could induce some amount of overfitting on the cyclone variables.



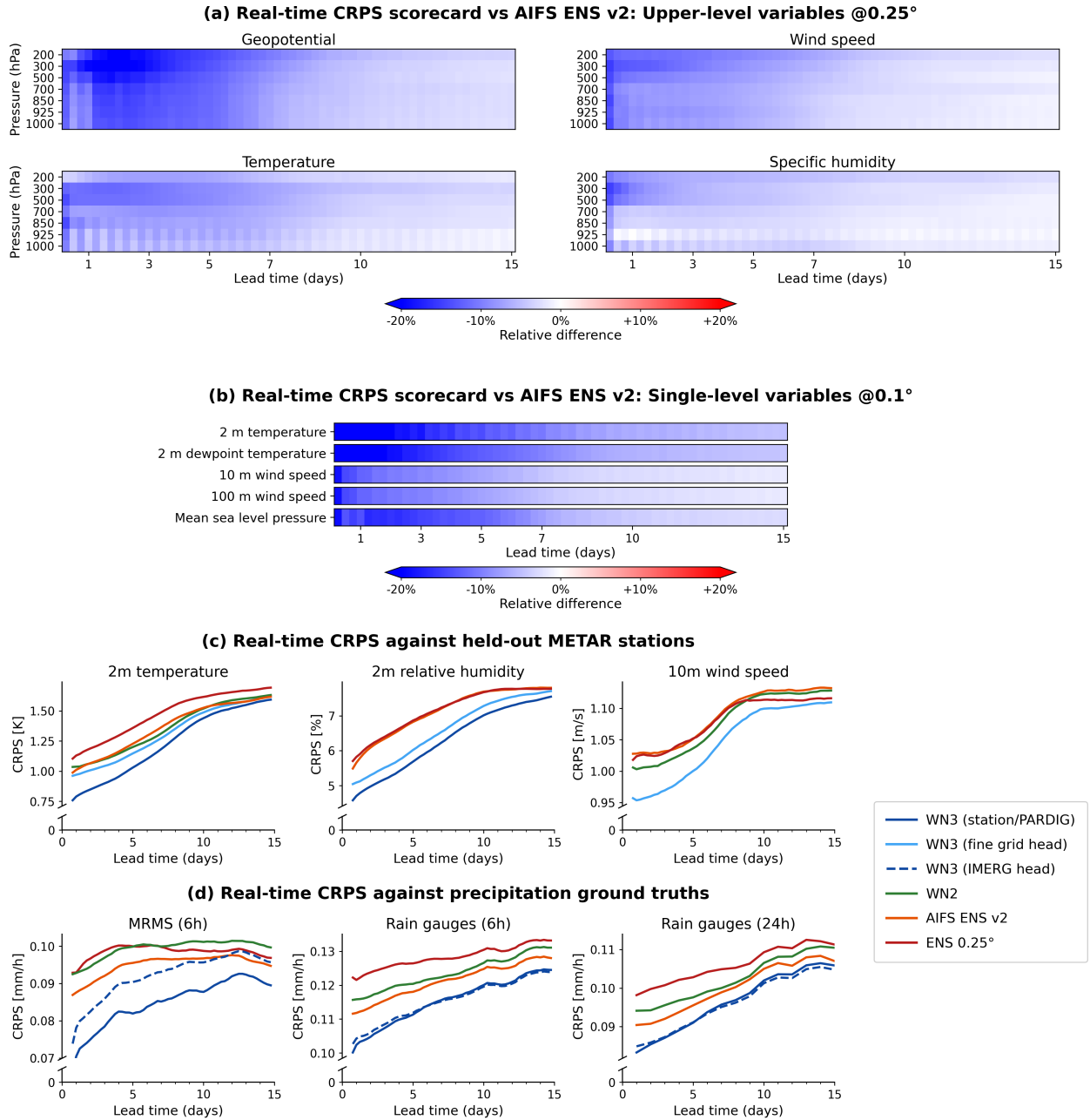


Figure 6 | **Real-time evaluation** (a) Upper level CRPS scorecard with respect to AIFS ENS v2 at 0.25°. Ground truth for AIFS ENS is the AIFS ENS Control fc0, while the ground truth for WN3 is HRES-fc0; (b) Single-level evaluations at 0.1°. 0.25° AIFS forecasts are bilinearly interpolated. (c) Evaluations 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.

#### 4.5. Real-time evaluation

To fairly compare the actual operational skill of the leading AI weather models, we run a quasi-real time evaluation using the operational data feeds of WN3, WN2 and AIFS ENS for the 6-week period July 1 to Aug 11 2026. This time period was chosen to allow us to evaluate our production model,

which was trained until June 30 2026 in order to include 50r1 data. While 6 weeks make for a relatively small sample size, clear trends are still visible but it is important to not over-interpret smaller signals. WN2’s operational checkpoint was trained until the end of 2024, and thus does not include 50r1, while both WN3 and AIFS ENS v2 include some of it in their training data.

WN3 outperforms AIFS ENS across all upper level variables with an average improvement of roughly 10% in the first forecast week (Figure 6a). For single-level variables, we evaluate against 0.1° HRES-fc0 ground truth, interpolating the 0.25° AIFS ENS forecasts. There, WN3 has a substantial advantage. Part of this is due to differences in interpolation and the choice of ground truth (see Section A.3 and Figure A.10). Even when accounting for these differences, however, WN3 outperforms AIFS ENS on the majority of variables and lead times. On station evaluations the WN3 station head shows a substantial reduction in 2m temperature and relative humidity error with respect to AIFS ENS (Figure 6c). The WN3 0.1° analysis predictions also perform strongly when evaluated against stations with a 5% improvement in wind speed CRPS over the baselines. AIFS ENS is roughly on par with WN2 for this period on 2m temperature and slightly worse on 10m wind speed.

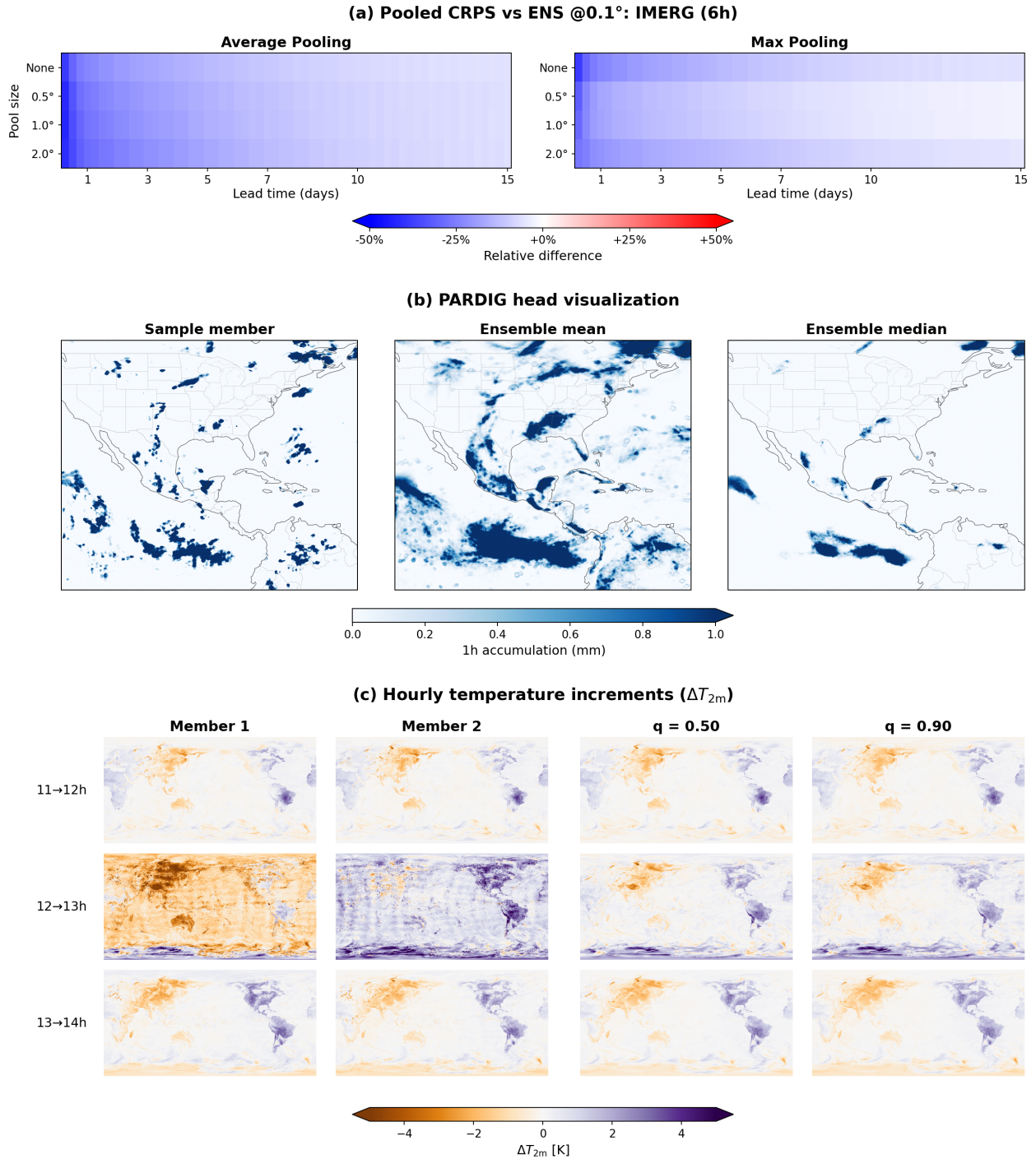
For precipitation, we evaluate against MRMS (6h) and rain gauges (6 and 24h). IMERG Final was not available for this real-time period. Because of the limited sample size, the precipitation evaluations are relatively noisy, especially for MRMS which only has limited geographic coverage. For MRMS, the PARDIG head has the lowest CRPS compared to the IMERG head and other baselines by some margin. For the rain gauge evaluations the PARDIG and IMERG heads are roughly on par, with a substantial gap to AIFS ENS, which itself performs strongly against WN2 and ENS.

#### 4.6. Joint forecast skill and artifacts

Despite being trained to optimize marginal skill, FGN is able to produce largely realistic joint structures (Alet et al., 2025). This is evident in its ability to produce reliable forecasts of cyclones (Alet et al., 2026), derived (wind speed and relative humidity) and accumulated variables (6 or 24h precipitation) as well as realistic-looking forecast visuals (Figure A.11). Further, spatial metrics show strong performance across spatial scales (see Fig. 3 in Alet et al., 2025). Here, we perform a spatially pooled evaluation of the WN3 PARDIG head precipitation forecasts (Figure 7a) for 6h accumulations. No systematic degradation of forecast skill is evident for larger pooling sizes. This is important for making reliable probabilistic predictions of cumulative precipitation over larger areas, used for example in flood forecasting.

Nevertheless, spatial and temporal artifacts are visible in individual WN3 samples. These issues arise because optimizing for marginal CRPS allows the model to ‘cheat’ on the covariance structure to maximize point-wise skill. One manifestation of this behaviour, first noted in WN2 (Fig. 5 of Alet et al., 2025), is the appearance of distinct hexagonal patterns that reflect the model’s underlying grid mesh structure. While visible to some degree across most variables, they are most severe in precipitation and station outputs. Figure 7b illustrates these patterns in the PARDIG head, where we have chosen a colormap with a deliberately low maximum hourly accumulation of 1mm in order to show the artifacts clearly (see Figure A.13 for station head examples). These artifacts are still visible in the ensemble mean, but less so in the ensemble median. We find that the PARDIG head has more pronounced artifacts than the IMERG head and believe this is due to PARDIG having a sharper distribution, leading to higher overall skill for the PARDIG head but making individual member artifacts more evident (Figure A.12).

Further, temporal discontinuities across the 6h outer time step boundaries are evident in the station output head, and to some degree the PARDIG and IMERG output heads. (Note that this does not apply to the hourly analysis output, which has no noticeable discontinuities.) The station output



**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.

head also exhibits a per-ensemble-member global bias, where one sample will tend to be warmer or colder globally compared to the ensemble mean. This global bias changes with each 6 hour time step as a new noise vector is drawn. Figure 7c shows the hourly temporal difference in 2m temperature station head predictions  $\Delta T_{2m}$ , which clearly shows the large jumps across 6h boundaries (12h to 13h). However, marginal statistics, such as quantiles, are largely artifact-free (except in Antarctica which is badly constrained in the station head due to the lack of observation data). Since many downstream applications rely primarily on ensemble statistics like median or threshold-exceedance probabilities, these forecasts remain highly usable.

## 5. Discussion

WeatherNext 3 offers substantial improvements over previous operational global weather models in forecast skill and capabilities. In terms of raw medium-range forecast skill, it sets a new state-of-the-art for probabilistic weather prediction, outperforming both its predecessor WN2 and ECMWF’s AIFS ENS v2 on a vast majority of analysis metrics. WN3 also produces output of single-level variables at  $0.1^\circ$ , matching the resolution of the best traditional global NWP system, ECMWF ENS. In addition, leveraging low-latency satellite input allows WN3 to be initialized every hour, rather than every 6 hours as is typical for most global weather models, allowing the model to adapt to rapidly changing weather phenomena, which delivers a boost in precipitation skill in early lead times.

Another jump in performance comes from directly training against high-quality observation data, circumventing known biases and inaccuracies in analysis data. For precipitation, we use our own precipitation reanalysis, PARDIG, derived from global space-borne radar data. This results in improved and more reliable precipitation forecasts, evaluated against a range of ground truths. For on-the-ground temperature and humidity, weather stations serve as the gold standard observations. Traditionally, sparse data sources have been hard to use in machine learning models. WN3’s architecture allows us to train directly against sparse data, yet still produce predictions anywhere on the globe provided easily available metadata, such as elevation and a land/sea mask. These predictions show a substantial reduction in error, even when evaluated against unseen locations. It is important to highlight that machine learning-based post-processing approaches for correcting global model output to station observations have been an established part of the weather forecasting toolkit for many decades (Glahn and Lowry, 1972; Gneiting et al., 2005; Hewson and Pilloso, 2021). By directly incorporating station calibration into the model, WN3 has the advantage of not needing a separate calibration stage. This reduces processing time and allows the station head to directly tap into the core model latent. Another feature of WN3’s station head is its strong spatial generalization. Post-processing methods that generalize beyond the training stations are an active area of research but, again, require training, fine-tuning and running a separate model (Vannitsem et al., 2021).

We believe that WN3 represents a major step forward for AI-based weather predictions by going beyond relying purely on analysis and utilizing information-dense, low-latency observation data. Physics-based analyses provide an uninterrupted, dense dataset perfect for training AI models. The existence of ERA5, for instance, has been a major contributor to the rapid advancement of AI weather models. However, it has also long been acknowledged that analysis only represents a best guess of the atmosphere, unable to use all observation sources to their full extent and exhibiting biases, especially for small scale processes. Here we demonstrate that augmenting an analysis-based model with observations in input and output allows us to alleviate these shortcomings and combine the best of both worlds.

## Acknowledgments

In alphabetical order, we thank Abby Bullock, Alex Wilkins, Aliyah Bond, Avinatan Hassidim, Bradley Goldstein, Christian Wagner, Daniel Rothenberg, Daniel Worrall, Darshan Prajapati, David Landry, Devaja Shah, Drew Bollinger, Drew Purves, Elinor Kruse, Emily Morris, Emma Yousif, Gaby Pearl, Hans Mohrmann, Jash Rana, Jenny Shepard, Jessica Sapick, Juanita Bawagan, Julian Green, Kasia Mohammed, Kunal Shah, Leen Verburgh, Marc Deisenroth, Marcus Trail, Natalie Williams, Piyush Ingale, Rachel Stigler, Rahul Mahrsee, Raia Hadsell, Reuven Sayag, Ryan Keisler, Shail Parekh, Shubham Joshi, Shubham Kumar, Stephanie Sanchez, Thomas Turnbull, Zach Hynes, and Zoubin Ghahramani.

We also thank ECMWF and NOAA for providing invaluable datasets to the research community, and the experts at NHC, UKMet and WNI.

## Disclaimer

**Experimental Forecasts:** WeatherNext is an automated, experimental AI system under active research. Predictions are provided "as-is" for informational purposes and are not official severe weather warnings. For protection of life and property, always defer to official alerts from your local emergency authorities and national meteorological services. Use of WeatherNext and its output are subject to our [Terms of Service](#).

This document is based on data and products of the European Centre for Medium-Range Weather Forecasts (ECMWF).



## References

- S. Agrawal, M. A. Hassen, E. A. Brempong, B. Babenko, F. Zyda, O. Graham, D. Li, S. Merchant, S. H. Potes, T. Russell, et al. An operational deep learning system for satellite-based high-resolution global nowcasting. *arXiv preprint arXiv:2510.13050*, 2025.
- F. Alet, A. K. Jeewajee, M. B. Villalonga, A. Rodriguez, T. Lozano-Perez, and L. Kaelbling. Graph element networks: adaptive, structured computation and memory. In *international conference on machine learning*, pages 212–222. PMLR, 2019.
- F. Alet, I. Price, A. El-Kadi, D. Masters, S. Markou, T. R. Andersson, J. Stott, R. Lam, M. Willson, A. Sanchez-Gonzalez, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.
- F. Alet, T. R. Andersson, I. Price, S. Markou, A. El-Kadi, D. Masters, A. Li, S. Merchant, N. Williams, G. Thornton, et al. Operational tropical cyclone forecasting with ai. *Nature*, pages 1–3, 2026.
- M. Andrychowicz, L. Espeholt, D. Li, S. Merchant, A. Merose, F. Zyda, S. Agrawal, and N. Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023.
- P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.
- K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.
- P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, et al. The thorpe interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91(8):1059–1072, 2010.
- S. R. Cachay, D. Watson-Parris, and R. Yu. U-cast: A surprisingly simple and efficient frontier probabilistic ai weather forecaster. *arXiv preprint arXiv:2604.09041*, 2026.
- J. P. Cangialosi and J. Martinez. National hurricane center forecast verification report: 2025 hurricane season. Technical report, National Hurricane Center (NHC), 2026. URL [https://www.nhc.noaa.gov/verification/pdfs/Verification\\_2025.pdf](https://www.nhc.noaa.gov/verification/pdfs/Verification_2025.pdf).
- J. J. Danielson and D. B. Gesch. Global multi-resolution terrain elevation data 2010 (GMTED2010). Technical Report 2011-1073, U.S. Geological Survey, 2011. URL <https://doi.org/10.3133/ofr20111073>.
- C. Diaconu, J. Scholz, A. Shysheya, S. Markou, P. Mukhopadhyay, M. Cranmer, and R. E. Turner. Otter weather: Skillful and computationally efficient medium-range weather forecasting. *arXiv preprint arXiv:2606.26421*, 2026.
- ECMWF. ECMWF High Resolution (HRES) Analysis Dataset, 2023. URL <https://www.ecmwf.int>. Available under Creative Commons Attribution 4.0 International (CC BY 4.0).
- ECMWF. Implementation of AIFS v2. *ECMWF Newsletter*, 187:16–22, 2026a. URL <https://www.ecmwf.int/en/newsletter/187/news/implementation-aifs-v2>.
- ECMWF. Dissemination Schedule. <https://confluence.ecmwf.int/spaces/DAC/pages/272310483/Dissemination+schedule>, 2026b. Accessed: 2026-09-02.



- C. A. Ferro, D. S. Richardson, and A. P. Weigel. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 15(1):19–24, 2008.
- E. Freeman, S. D. Woodruff, S. J. Worley, S. J. Lubker, E. C. Kent, W. E. Angel, D. I. Berry, P. Flick, R. Freeman, P. G. Hutchins, R. Kolper, R. W. Reynolds, and S. R. Smith. ICOADS Release 3.0: A major update to the historical marine climate record. *International Journal of Climatology*, 37(5): 2211–2232, 2017. doi: 10.1002/joc.4775.
- J. Gahtan, K. R. Knapp, C. J. Schreck, H. J. Diamond, J. P. Kossin, and M. C. Kruk. International best track archive for climate stewardship (ibtracs) project, version 4r01. NOAA National Centers for Environmental Information. doi:10.25921/82ty-9e16, 2024.
- Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- H. R. Glahn and D. A. Lowry. The use of model output statistics (mos) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, 11(8):1203–1211, 1972.
- W. B. Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.
- T. Gneiting, A. E. Raftery, A. H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly weather review*, 133(5):1098–1118, 2005.
- M. Grecu, W. S. Olson, S. J. Munchak, S. Ringerud, L. Liao, Z. Haddad, B. L. Kelley, and S. F. McLaughlin. The gpm combined algorithm. *Journal of Atmospheric and Oceanic Technology*, 33(10):2225 – 2245, 2016. doi: 10.1175/JTECH-D-16-0019.1.
- A. Gupta, A. Subramaniam, M. S. Pritchard, K. Kashinath, S. Frolov, K. Lieberman, C. Miller, N. Silverman, and N. D. Brenowitz. Healda: Highlighting the importance of initial errors in end-to-end ai weather forecasts. *arXiv preprint arXiv:2601.17636*, 2026.
- T. Haiden, M. Janousek, F. Vitart, F. Prates, M. Maier-Gerber, C. W. Y. Li, and M. Chevallier. Evaluation of ecmwf forecasts, 09/2025 2025.
- H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.
- T. D. Hewson and F. M. Pilloso. A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth & Environment*, 2(1):132, 2021.
- G. J. Huffman, D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, C. Kidd, E. Nelkin, and P. Xie. Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (IMERG) version 07. *Algorithm theoretical basis document (ATBD) version*, 47, 2023.
- B. Ingleby. Global assimilation of air temperature, humidity, wind and pressure from surface stations. *Quarterly Journal of the Royal Meteorological Society*, 141(687):504–517, 2015.

- M. S. Ingstad, M. C. Clare, O. Ersland, V. Gahlen, H. H. Haugen, O. Miralles, E. M. Nordhagen, T. N. Nipen, I. A. Seierstad, J. B. Bremnes, et al. Hourglass: A probabilistic data-driven temporal downscaler for global and regional weather forecasting. *arXiv preprint arXiv:2607.11457*, 2026.
- K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. doi: 10.1175/2009BAMS2755.1.
- D. Kochkov, J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klöwer, J. Lottes, S. Rasp, P. Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.
- J. Kossaiifi, N. Kovachki, M. Mardani, D. Leibovici, S. Ravuri, I. Shokar, E. Calvello, M. S. Abbas, P. Harrington, A. Subramaniam, et al. Demystifying data-driven probabilistic medium-range weather forecasting. *arXiv preprint arXiv:2601.18111*, 2026.
- R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.
- S. Lang, M. Leutbecher, and P. Maciel. A multi-scale loss formulation for learning a probabilistic model with proper score optimisation. *arXiv preprint arXiv:2506.10868*, 2025.
- S. Lang, M. Alexe, M. C. Clare, C. Roberts, R. Adewoyin, Z. Ben Bouallègue, M. Chantry, J. Dramsch, P. D. Dueben, S. Hahner, et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2(1):18, 2026.
- D. A. Lavers, A. Simmons, F. Vamborg, and M. J. Rodwell. An evaluation of era5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148(748):3152–3165, 2022.
- K. Lean and N. Bormann. Using model cloud information to reassign low-level atmospheric motion vectors in the ecmwf assimilation system. *Journal of Applied Meteorology and Climatology*, 62(3):361–376, 2023.
- R. Liu, X. Zhang, W. Wang, Y. Wang, H. Liu, M. Ma, and G. Tang. Global-scale era5 product precipitation and temperature evaluation. *Ecological Indicators*, 166:112481, 2024.
- S. Moazami and M. Najafi. A comprehensive evaluation of gpm-imerg v06 and mrms with hourly ground-based precipitation observations across canada. *Journal of Hydrology*, 594:125929, 2021.
- A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47, 1977.
- NCEP. NCEP, cited 2011: Format of the tropical cyclone vital statistics records, 2025. URL [http://www.emc.ncep.noaa.gov/mmb/data\\_processing/tcvitals\\_description.htm](http://www.emc.ncep.noaa.gov/mmb/data_processing/tcvitals_description.htm).
- Z. Ni, J. Weyn, H. Zhang, Y. Xiang, J. Bian, W. Jin, K. Thambiratnam, Q. Zhang, H. Dong, and H. Sun. Huracan: A skillful end-to-end data-driven system for ensemble data assimilation and weather prediction. *arXiv preprint arXiv:2508.18486*, 2025.
- NOAA National Severe Storms Laboratory and NCEP. Multi-Radar Multi-Sensor (MRMS) System Quantitative Precipitation Estimation (QPE). NOAA NSSL and NCEP, 2024. URL <https://www.nssl.noaa.gov/projects/mrms/>.

- E. Pinnington, P. Lean, M. Alexe, E. Boucher, S. Lang, P. Laloyaux, G. Mertes, T. Kral, P. de Rosnay, M. Chantry, et al. Aifs-dop: End-to-end medium-range weather prediction from observations alone with machine learning. *arXiv preprint arXiv:2606.19093*, 2026.
- I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.
- S. Rasp, S. Hoyer, A. Merose, I. Langmore, P. Battaglia, T. Russell, A. Sanchez-Gonzalez, V. Yang, R. Carver, S. Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.
- S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.
- Research Data Archive at NCAR and NOAA NCEI. International Comprehensive Ocean-Atmosphere Data Set (ICODAS) Release 3, Individual Observations. NSF National Center for Atmospheric Research and NOAA, 2016.
- T. M. Smith, V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, et al. Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97(9):1617–1630, 2016.
- C. Soci, H. Hersbach, A. Simmons, P. Poli, B. Bell, P. Berrisford, A. Horányi, J. Muñoz-Sabater, J. Nicolas, R. Radu, et al. The era5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society*, 150(764):4014–4048, 2024.
- C. K. Sørnderby, L. Espeholt, J. Heek, M. Dehghani, A. Oliver, T. Salimans, S. Agrawal, J. Hickey, and N. Kalchbrenner. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.
- S. Vannitsem, J. B. Bremnes, J. Demaeyer, G. R. Evans, J. Flowerdew, S. Hemri, S. Lerch, N. Roberts, S. Theis, A. Atencia, et al. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681–E699, 2021.
- A. Vaughan, S. Markou, W. Tebbutt, J. Requeima, W. P. Bruinsma, T. R. Andersson, M. Herzog, N. D. Lane, M. Chantry, J. S. Hosking, et al. Aardvark weather: end-to-end data-driven weather forecasting. *arXiv preprint arXiv:2404.00411*, 2024.
- WindBorne Systems. What’s new in weathermesh-6, June 2026. URL <https://windbornesystems.com/blog/introducing-wm-6>.
- D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.-E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino. ESA WorldCover 10 m 2021 v200, 2022. URL <https://doi.org/10.5281/zenodo.7254221>.
- P. Zhao, J. Bian, Z. Ni, W. Jin, J. Weyn, Z. Fang, S. Xiang, H. Dong, B. Zhang, H. Sun, et al. Omg-hd: A high-resolution ai weather model for end-to-end forecasts from observations. *arXiv preprint arXiv:2412.18239*, 2024.

P. Zhao, S. Xiang, W. Jin, Z. Ni, J. Bian, Z. Fang, H. Sun, B. Zhang, R. E. Turner, J. Weyn, et al. Skillful high-resolution weather forecasting independent of physical models. *arXiv preprint arXiv:2605.28153*, 2026.

## Appendix

### A.1. Model details

#### A.1.1. Model formulation

To restate our notation from Section 2, the atmosphere at timestep  $t$  and spatial location  $s$  is represented by a set of  $N$  spatio-temporal fields (indexed by  $i$ ),

$$\mathbf{W}_t = \{f_i(s, t) : \mathcal{S}_i \times \mathcal{T}_i \rightarrow \mathcal{V}_i\}_{i=1}^N, \quad (\text{A.1})$$

where each field  $f_i$  denotes a physical variable over spatial domain  $\mathcal{S}_i \subseteq \mathbb{S}^2$ , continuous or discrete time subset  $\mathcal{T}_i \subseteq \mathbb{R}$ , and value space  $\mathcal{V}_i$ .

We parameterise the single-step conditional with a learnable operator  $\Phi_\theta$ :

$$\mathbf{W}_{((t)\Delta, (t+1)\Delta]} = \Phi_\theta(\mathbf{W}_{((t-2)\Delta, t\Delta]}, \boldsymbol{\xi}_t), \quad (\text{A.2})$$

where  $\boldsymbol{\xi}_t$  is a single noise vector that parameterises aleatoric forecast uncertainty via conditional normalisation layers, and  $(t)\Delta$  indexes the  $t$ -th  $\Delta$ -wide time interval.

WN3 builds on FGN (Alet et al., 2025), which underpins WeatherNext 2.

**Latent grids and processor mesh.** The operator  $\Phi_\theta$  follows an encode–process–decode structure. There are two input grids, one input at each resolution, indexed by  $j \in \{0, 1\}$ . Denote each input as  $\mathcal{E}_{j,t} \in \mathbb{R}^{a_j \times b_j \times c_j}$  with spatial resolution  $(a_j \times b_j)$  and  $c_j$  channels.  $\mathcal{E}_{j,t}$  is first encoded point-wise to a latent grid  $X_{j,t} \in \mathbb{R}^{a_j \times b_j \times 1024}$ , and then encoded by a Graph Neural Network (GNN) into a 6-times-refined icosahedral mesh latent  $Z_{j,t} \in \mathbb{R}^{40962 \times 1024}$ , an operation which also produces an updated grid latent  $X_{j,t}^{\text{skip}} \in \mathbb{R}^{a_j \times b_j \times 1024}$ . A unified latent mesh  $\mathbf{Z}_t = \sum_j Z_{j,t}$  is then processed by an  $L$ -layer graph transformer, before being decoded back to the latent grid and added to  $X_{j,t}^{\text{skip}}$  yielding  $\bar{X}_{j,t}$ .

**Gridded and continuous output heads.** Gridded fields are decoded onto fixed latitude-longitude grids at their native resolutions  $(a_j \times b_j)$ , pointwise, from  $\bar{X}_{j,t}$  via a Multilayer Perceptron (MLP). Surface station predictions are decoded as a continuous query from a  $0.1^\circ$  latent grid (i.e.  $a_j = 1801, b_j = 3600$ ): for any point  $(s, \delta t) \in \mathbb{S}^2 \times [0, \Delta]$ , the prediction is

$$\mathbf{v}(s, \delta t) = \text{MLP}_{\text{stn}}\left(\text{Interp}\left(\psi([X_{j,t}^{\text{skip}} \parallel \bar{X}_{j,t}], s) \parallel \mathbf{m}(s, \delta t)\right)\right), \quad (\text{A.3})$$

where  $\text{Interp}(\cdot, s)$  bilinearly interpolates a grid latent to  $s$  after the latent is processed by a 4-layer, width-768 convolutional neural network (CNN)  $\psi$ . The input to this CNN is the concatenation of the input and decoded latents  $[X_{j,t}^{\text{skip}} \parallel \bar{X}_{j,t}]$ ,  $\mathbf{m}(s, \delta t) = \text{MLP}_{\text{mdata}}([\text{elevation}(s), \text{is\_land}(s), \delta t]) \in \mathbb{R}^{768}$  encodes local surface features alongside the normalized relative observation query time  $\delta t \in [-1, 0]$  with a 1-hidden layer width-768 MLP,  $\text{MLP}_{\text{stn}}$  is a 2-hidden-layer width-768 MLP, and  $\parallel$  indicates concatenation along the feature axis. This follows the continuous querying approach of Alet et al. (2019) from a common processor mesh, enabling training on sparse, irregularly located observations and inference at arbitrary spatial and temporal query points.

**Training objective.** The model is trained end-to-end on a composite loss accumulated over  $R$  autoregressive rollout steps. We use the fair Continuous Ranked Probability Score (CRPS) with two sampled trajectories  $f(\cdot, \xi_{1:t}^1), f(\cdot, \xi_{1:t}^2) =: f^{1,2}$  and corresponding target  $\hat{f}$ . For field  $i$ , the loss  $\ell_i(\cdot, \cdot)$  averages over that field’s own set of spatial positions  $S_i \subset S^2$  and temporal samples  $\mathcal{T}_i$ , so that the per-modality weights  $\lambda_i$  (see Table A.3) control relative importance independently of grid size or sampling density:

$$\mathcal{L} = \sum_{i=1}^N \lambda_i \operatorname{avg}_{(s,q) \in S_i \times \mathcal{T}_i} \ell_i(\hat{f}_i(s,q), f_i^{1,2}(s,q)). \quad (\text{A.4})$$

For stations, the average runs over the variable-sized set of available reports within each  $\Delta$ -wide window. For gridded cyclone output fields, the average runs over all grid points for which the gridded target is not NaN (see Alet et al. (2026)). This design ensures that sparse modalities (stations, cyclones) are not overwhelmed by the much larger gridded fields.

Within the atmospheric analysis fields, each variable–pressure-level pair  $(v, p)$  carries its own weight  $\lambda_{v,p}$ :

$$\mathcal{L}_t^{\text{atm}} = \sum_{v=1}^C \sum_{p=1}^P \lambda_{v,p} \operatorname{avg}_{(s,h) \in S_v \times \mathcal{T}_v} \text{CRPS}(\hat{f}_{v,p}(s,h), f_{v,p}^{1,2}(s,h)). \quad (\text{A.5})$$

where pressure level weighting is as in Alet et al. (2026) and relative variable weightings are specified in Table A.3. For all variables specified as input and output in Table 1, with the exception of geostationary satellite fields, the target above is the residual with respect to the last input frame. Similarly, analysis variables defined on subtime channels target the residual with respect to the corresponding non-subtime variable on the outer frames. This includes the pressure-level slices, for which the corresponding level from the outer frame is used.

### A.1.2. Additional model modifications

Apart from the key features mentioned in the main text, we add the following model changes:

**Global mean loss.** For precipitation and cloud cover variables we observed that models trained solely on local (point-wise) marginal CRPS exhibited substantial per-sample bias across individual ensemble members. To mitigate this, we incorporated a globally pooled CRPS component into the training loss, weighted at a factor of 0.3 to the corresponding base variable’s loss weight. We found this to be a simple-but-effective fix, related to spherical harmonic losses (Kochkov et al., 2024) or Gaussian smoothing (Lang et al., 2025). While this pooled objective effectively eliminates per-sample bias for some variables, we found that applying it to sparse station variables exacerbated mesh-scale artifacts; we therefore omitted the pooled loss for station targets.

**Cumulative variables.** WN3 uses HRES-fc0-5 as both an input and a training target. We model cumulative variables (*cdir*, *fdir*, *ssrd* and *tp*) as the accumulation within a given time period (e.g. for an initialization time of 03 UTC and a lead time of 1 hour it would be the accumulation from 03 UTC to 04 UTC). In HRES these variables are defined as accumulations from time 0, so we "de-accumulate" them. However, for HRES-fc0 this leaves us with a constant value of 0. In this case, during training, we populate these variables using the HRES-fc6 values from the prior initialization time as necessary. For these cumulative variables, we predict both 1-hour and 6-hour accumulations.



**Input dropout for operational robustness.** Several autoregressive inputs are unavailable on the first forecast step: as mentioned previously, cumulative analysis variables have a value of 0 in HRES-fc0, while both IMERG and PARDIG products are currently not available operationally with an acceptable latency. Similarly, for older training years, some inputs (geostationary satellite imagery, IMERG and PARDIG) do not exist; in these cases, the corresponding input channels are filled with their training-set mean, to mask them out. To bridge the train–inference gap for autoregressive inputs, we apply stochastic input dropout during training: with probability 0.9 an input modality is masked as if it were the first rollout step (unavailable); with probability 0.02, as the second step (partially available from the previous output); and with the remaining 0.08 probability, the full input is provided (falling back to prior HRES-fc6 values for HRES-fc0 cumulative variables, as described above). This teaches the model to produce skilful forecasts regardless of which autoregressive inputs are available, removing the need for separate handling at initialization time.

**Sea surface temperature (sst).** As in Alet et al. (2026), missing values over land in ERA5 are replaced with the minimum value of the variable seen in a subset of the dataset. For HRES-fc0-5, we imputed this same minimum value over land for *sst*, using ERA5’s SST validity mask at 0.25°, and HRES’s land sea mask at 0.1°, to define the land mask.

**Spatial sharding.** Given the scale of WN3, we introduce spatial sharding of the processor mesh: the icosahedral mesh nodes are partitioned along a spatial axis (longitude by default), and gather/scatter operations across grid-to-mesh and mesh-to-grid edges are executed as local operations within each partition, with only the mesh-to-grid scatter requiring global communication. Sorting the grid-to-mesh edges by senders makes the most expensive sharding operation local, substantially reducing cross-device communication.

### A.1.3. Training

WN3 is trained via a multi-stage curriculum that progressively increases spatial resolution and introduces additional observation modalities. Table A.1 summarises the full stage sequence and Table A.2 lists hyperparameters used for training.

| #  | Stage                          | $\Delta$ | "Coarse" Grid                                                | "Fine" Grid                                      | Stations        | Backbone      |
|----|--------------------------------|----------|--------------------------------------------------------------|--------------------------------------------------|-----------------|---------------|
| 1  | 1° pretraining (12h)           | 12h      | 1°. < 2016: ERA5; < 2023: HRES-fc0-5, Geosats, IMERG, PARDIG |                                                  | -               | Train         |
| 2  | 1° pretraining (6h)            | 6h       | Same as Stage 1                                              |                                                  | -               | Train         |
| 3  | 0.25° pretraining              | 6h       | 0.25°. Same as Stage 1                                       |                                                  | -               | Train         |
| 4  | Cyclone warmup                 | 6h       | Same as Stage 3 + Cyclones < 2023                            |                                                  | -               | <b>Frozen</b> |
| 5  | Cyclone e2e                    | 6h       | Same as Stage 4                                              |                                                  | -               | Train         |
| 6  | 0.1° pretraining               | 6h       | 0.25°. < 2023: HRES-fc0-5, Cyclones, Geosats, IMERG, PARDIG  | 0.1°. < 2023: HRES-fc0-5, Geosats, IMERG, PARDIG | -               | Train         |
| 7  | AR finetuning (2→7)            | 6h       | Same as Stage 6                                              | Same as Stage 6                                  | -               | Train         |
| 8  | Frozen finetuning (stations) 1 | 6h       | Same as Stage 6                                              | Same as Stage 6                                  | < 2023 METAR    | <b>Frozen</b> |
| 9  | AR finetuning (8)              | 6h       | Same as Stage 6, extended to < 2024/26                       | Same as Stage 6, extended to < 2024/26           | -               | Train         |
| 10 | Frozen finetuning (stations) 2 | 6h       | Same as Stage 9                                              | Same as Stage 9                                  | < 2024/26 METAR | <b>Frozen</b> |

Table A.1 | **Training stages for WeatherNext 3** (for data start dates, see Table 1).

**Progressive increase of resolution.** The model begins training on ERA5 (<2016)/HRES-fc0-5 ( $\geq 2016$ ) with geostationary satellite imagery, IMERG and PARDIG precipitation, and hourly sub-step

| #  | Stage                          | Peak LR              | Total Steps | Gradient Norm Clipping |
|----|--------------------------------|----------------------|-------------|------------------------|
| 1  | 1° pretraining (12h)           | 4.95e-4              | 400 000     | 3.5                    |
| 2  | 1° pretraining (6h)            | 7.07e-5              | 100 000     | 5.5                    |
| 3  | 0.25° pretraining              | 7.07e-5              | 32 000      | 6.5                    |
| 4  | Cyclone warmup                 | 7.07e-5              | 5 000       | 6.5                    |
| 5  | Cyclone e2e                    | 7.07e-5              | 20 000      | 6.5                    |
| 6  | 0.1° pretraining               | 7.07e-5              | 8 000       | 5.5                    |
| 7  | AR finetuning (2→7)            | 7.07e-6, [7.07e-7]×5 | 9 000       | 11.0                   |
| 8  | Frozen finetuning (stations) 1 | 7.07e-5              | 12 000      | 5.5                    |
| 9  | AR finetuning (8)              | 7.07e-7              | 1 000       | 11.0                   |
| 10 | Frozen finetuning (stations) 2 | 1.17e-5              | 4 000       | 5.5                    |

Table A.2 | **Training hyperparameters for WeatherNext 3 stages.** Across all stages: the optimizer is AdamW with weight decay 0.1 and  $\beta_2 = 0.95$ ; a cosine decay is used for the learning rate schedule, with warmup steps  $\min(1000, 0.1 * \text{total steps})$ ; the batch size is 64 and the number of training samples in the FGN setup is 2.

analysis targets. This is initially at 1° resolution with 12 h outer steps. This then decreases to 6 h steps, after which resolution is progressively increased to 0.25° and finally 0.1°. The model is trained with two grids: "coarse" and "fine", both of which are encoded and decoded (supporting variables as input and output) at each step. All gridded variables are modelled by the coarse grid, while the fine grid separately only models the subset of variables ultimately available at 0.1° resolution. In 1° and 0.25° stages, both grids operate at the respective stage's resolution. In the final 0.1° stages, the "coarse" grid remains at 0.25° while the "fine" grid is increased to 0.1°. Each grid resolution increase is accompanied by a scaling of the sum of message vectors in the encoder GNN to preserve the scale of the incoming signal. From 1° → 0.25° this is  $\frac{1}{4}$  and from 0.25° → 0.1° this is  $\frac{1}{25}$ .

**Normalization.** All input variables are normalized to unit variance and zero mean. Targets (in their potentially residual form) are similarly normalized, with the exception of the cyclone existence field and subtime analysis variables. The former is normalized by definition, while the latter are approximately normalized using the statistics of the outer-frame differences.

**Cyclone finetuning.** Cyclone attribute maps are introduced with a two-step process: first a *warmup* stage trains only the newly added output heads with the backbone frozen, then an *end-to-end* stage unfreezes the full model for joint optimisation.

**Autoregressive finetuning.** The model is finetuned with progressively longer rollouts, from 2 to 8 autoregressive steps. These phases use all previously introduced modalities as both input and target.

**Frozen finetuning at high resolution.** The station observation heads are trained while keeping the backbone and all other heads fixed at 0.1° resolution. Frozen finetuning is interleaved with AR finetuning via a schedule: an initial frozen stage (75% of total steps) after 7AR, then the final 8AR stage, followed by a second frozen stage (25% of steps) that merges the final backbone weights from AR with the station weights from the initial frozen stage.

**Data recency.** The interleaving process above is done to enable finetuning up to different years, in a compute efficient manner, while maintaining causality for the different evaluations. Up to the end

| Target Variables                                             | "Coarse" grid |              | "Fine" grid |              |
|--------------------------------------------------------------|---------------|--------------|-------------|--------------|
|                                                              | Main (6h)     | Subtime (1h) | Main (6h)   | Subtime (1h) |
| $t, u, v, q, w$                                              | 1.0           | -            | -           | -            |
| $z$                                                          | 0.2           | -            | -           | -            |
| 300/500 hPa $t/z$ ; 1000 hPa $u/v$                           | -             | 0.1          | -           | -            |
| $2t$                                                         | 0.5           | 0.5          | 0.5         | 0.5          |
| $10u, 10v, 100u, 100v$                                       | 0.025         | 0.025        | 0.025       | 0.025        |
| $2d, msl, sst, tcc, hcc, mcc, lcc, tp\_1hr$                  | 0.05          | 0.05         | 0.05        | 0.05         |
| $tp\_6hr$                                                    | 0.05          | -            | 0.05        | -            |
| $ssrd\_1hr, cdir\_1hr, fdir\_1hr$                            | 1/60          | 1/60         | 1/60        | 1/60         |
| $ssrd\_6hr, cdir\_6hr, fdir\_6hr$                            | 1/60          | -            | 1/60        | -            |
| PARDIG 1h rain rate                                          | 0.75          | -            | 0.75        | -            |
| IMERG 1h rain rate                                           | 0.543         | -            | 0.543       | -            |
| Geosats mosaic (each 11 spectral channels)                   | 0.05          | -            | 0.05        | -            |
| Cyclone existence, $V_{max}, P_{min}$                        | 20.0          | -            | -           | -            |
| Cyclone structure: $R_{max}$ ,<br>34/50/64-kt quadrant radii | 2.0           | -            | -           | -            |

Table A.3 | **Per-variable loss weights across spatial and temporal representations for WeatherNext 3.** Hyphens (-) indicate that the variable is not evaluated on that representation. Sparse METAR variables  $2t$ ,  $2d$  are weighted at 1.0.

of the first frozen finetuning stage, the model is trained with data spanning up to the end of 2022. At 8AR, more recent data is introduced (and so is also used in the final stage of frozen finetuning). We use this to generate "splits" of the model, each trained with data up to the start of its corresponding year: <2023 (used for validation in research process), <2024 (for non-realtime evaluations), <2025, <2026 and <2026-07-01 (for the production model and realtime evaluations). For the latter two splits, IMERG Final is only available up to September 2025 so we employ the mean-masking described above for the variable's entire frame.

**Output clipping.** We noticed that in inference, some stationary pixels have increasingly unrealistic values during rollout. We traced this back to pixels going out of the training distribution, specifically for bounded variables like cloud cover. As a simple fix we clip those bounded variables during inference ( $[0, 1]$  for cloud cover,  $\geq 0$  for specific humidity, solar radiation and precipitation) at each step, before feeding the output auto-regressively. This simple solution prevents the occurrence of those pixel artifacts.

**Computational resources.** Each model seed took roughly 6.7 days of wall clock time to train. The exact hardware configuration varied depending on training stage. In total, both models required approximately 4.8 chip-years on TPUv4 and 8.9 chip-years on TPU7x chips. Generating a single ensemble member forecast takes roughly 6.3 minutes of wall clock time using 4 TPUv5p chips.

## A.2. Additional information on evaluation

### A.2.1. CRPS

We use the biased version of the CRPS because WN3 as a multi-seed ensemble violates the assumption of i.i.d. inherent in the unbiased CRPS (Ferro et al., 2008). Further, this more accurately reflects

the skill of the forecasts users would actually receive operationally rather than that of a theoretical ensemble of infinite size. We make one exception for our ENS baselines, where due to various constraints we did not have the full 50+1 member ensemble available. We note, however, that the differences between biased and unbiased metrics for ensemble sizes of  $O(50)$  are small.

### A.2.2. Pooled CRPS

Pooled CRPS computes CRPS on fields which have undergone dimensionality-preserving spatial average- or max- pooling. Pooling regions are defined on the latitude-longitude grid to be approximately equi-area patches, defined by their size in degrees at the equator (and thus warping as latitude increases or decreases towards the poles). A pooled value is computed for a patch centred at every grid point, and so latitude-based weighting is applied when computing the overall spatial average of the pooled CRPS, to account for over-counting at higher/lower latitudes.

## A.3. Additional information on datasets

| Short name | Long name                                   |
|------------|---------------------------------------------|
| $z$        | Geopotential                                |
| $t$        | Temperature                                 |
| $q$        | Specific humidity                           |
| $u$        | Zonal wind component                        |
| $v$        | Meridional wind component                   |
| $w$        | Vertical velocity                           |
| $2t$       | 2 m temperature                             |
| $2d$       | 2 m dewpoint temperature                    |
| $10u$      | 10 m zonal wind component                   |
| $10v$      | 10 m meridional wind component              |
| $100u$     | 100 m zonal wind component                  |
| $100v$     | 100 m meridional wind component             |
| $msl$      | Mean sea level pressure                     |
| $tcc$      | Total cloud cover                           |
| $hcc$      | High cloud cover                            |
| $mcc$      | Medium cloud cover                          |
| $lcc$      | Low cloud cover                             |
| $cdir^*$   | Clear-sky direct solar radiation at surface |
| $fdir^*$   | Total-sky direct solar radiation at surface |
| $ssrd^*$   | Surface solar radiation downwards           |
| $tp^*$     | Total precipitation                         |
| $tisr^*$   | Total incident solar radiation              |

Table A.4 | List of variable abbreviations and full descriptions. \* variables are cumulative and may be denoted with an accumulation period suffix ( $_{1hr}$ ,  $_{6hr}$ ) where relevant.

Here we provide additional details for some of the datasets already mentioned in the main text.

### A.3.1. ECMWF ENS and AIFS ENS baselines

For our 2024 evaluations, we download upper-level ENS forecasts for 00/12 initializations from TIGGE (Bougeault et al., 2010) in the native grid and interpolate it to  $0.25^\circ$ . For single-level variables, we download native grid TIGGE forecasts for 2t, 2d and 10u/v for 6 hour lead time multiples. For the hourly lead times in-between we download native grid data from MARS. For all other single level

variables, we download all lead times from MARS. We then interpolate those native grid data to  $0.1^\circ$ . However, we noticed that there seems to be a difference between the forecasts downloaded from TIGGE and MARS, stemming from the fact that TIGGE forecasts are stored at a reduced resolution<sup>6</sup>. This is evident in the hourly evaluations (Figure A.3 and Figure A.4). Since our ground truth is entirely downloaded from MARS, these discrepancies cause small spikes in error for those variables at 6h lead time multiples. Without further analysis, we can only estimate the magnitude of those spikes from the shape of the curves. For 2m temperature, which is the most affected variable, the effect is roughly 0.02K in CRPS for  $0.1^\circ$  analysis evaluations (note that the 6-hourly step pattern in Figure A.3 is predominantly caused by another mechanism, discussed in the figure caption). Since the gap to WN3 is generally substantially larger than 0.02K, these discrepancies should not qualitatively affect the results. In Figure 2, the only impacted variable is 2m dewpoint temperature. For station evaluations, 10m wind speed seems to be most affected, with no big difference in 2m temperature and relative humidity. For wind speed the difference is on the order of 0.02 m/s CRPS, which is much smaller than the gap to WN3.

Also note that, for download efficiency, we downloaded only 48 members for ENS (without the control forecast). We use the fair CRPS for ENS  $0.1^\circ$  in all evaluations to account for this difference.

For the real-time 2026 evaluations we obtain all data from the ECMWF open-data program<sup>7</sup>. The ENS data there is only provided at  $0.25^\circ$ . For AIFS ENS we use the AIFS ENS control member at  $t = 0$  as a ground truth. This data stems from the same raw fields as the HRES-fc0 we use but is regridded differently. While we directly convert from the native O1280 reduced Gaussian grid to  $0.25^\circ$ , for AIFS the fields are first regridded to the AIFS-native O320 grid, before being interpolated to  $0.25^\circ$ . This introduces differences that affect scores meaningfully.

### A.3.2. Geostationary satellite mosaic.

We construct our own real-time mosaic of geostationary satellite imagery by combining data streams of several satellites, following the methodology in Agrawal et al. (2025). The composite spans 11 spectral channels across visible and infrared regimes (central wavelengths at 0.64, 0.863, 1.61, 3.83, 6.20, 7.33, 8.59, 9.62, 11.20, 12.40, and  $13.30\mu\text{m}$ ). While the native spatial resolution varies across instruments and channels, we regrid them all to  $0.1^\circ$  for training WN3. The historical dataset extends back to 2016 and covers several generations of satellites and instruments; the real-time configuration currently uses the GOES-18/19, Meteosat-9/11, and Himawari-8/9 satellites.

### A.3.3. In-situ surface observations

We build a combined hourly dataset of METAR, Mesonet (available from MADIS) and ICOADS (Research Data Archive at NCAR and NOAA NCEI, 2016) for training starting in June 2001. For Mesonet, which provides some sub-hourly observations, if more than one observation is available per hour for a given station, we pick the observation closest to the hour. For METAR and Mesonet we use the quality control flags provided with the data to filter out bad-quality observations. For ICOADS, no such flags are provided. However, we noticed clear outliers. To remove those, we first enforce rough physical bounds:  $-20$  to  $40^\circ\text{C}$  for temperature and dewpoint, and  $0$  to  $50\text{ m/s}$  for wind speed. Second, we then apply a departure check from ERA5 for all three datasets, where data points that deviate more than  $5^\circ\text{C}$ , or  $5\text{ m/s}$  are removed. This only applies to training. For evaluation, we only remove data based on the provided quality control flags. We hold out 5% of METAR and Mesonet stations randomly from training, which we then use for independent evaluation.

<sup>6</sup><https://confluence.ecmwf.int/spaces/TIGGE/pages/40109876/Models>

<sup>7</sup><https://www.ecmwf.int/en/forecasts/datasets/open-data>

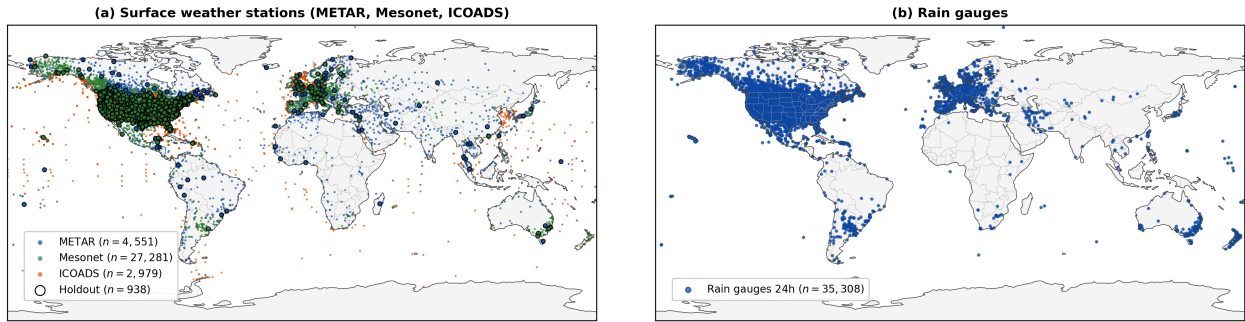


Figure A.1 | Distribution of in-situ observations for 1 January 2024 00 UTC. (a) METAR, Mesonet and ICOADS stations. Hold-out stations with outline. (b) Rain-gauge observations.

#### A.3.4. Elevation and land-sea mask

To make predictions using the trained station head at any location in the globe, we need to provide the appropriate metadata: latitude, longitude, elevation and land/sea. For elevation, we create a global  $0.05^\circ$  elevation map by reprojecting the Global Multi-resolution Terrain Elevation Data (Danielson and Gesch, 2011). For the land-sea mask we use the ESA WorldCover 10m v200 dataset (Zanaga et al., 2022). The reprojection is done by taking the mean of the high-resolution datasets, so that each  $0.05^\circ$  grid point represents the grid box average.

#### A.3.5. IMERG

NASA’s Integrated Multi-satellite Retrievals for GPM (IMERG) product (Huffman et al., 2023) (IMERG) is a widely used precipitation estimate based on microwave and infrared observations from the GPM constellation. Specifically, we use IMERG Final, which has been additionally calibrated against gauge observations. IMERG is available at  $0.1^\circ$  spatial and 30 minute temporal resolution, which we sum to hourly accumulations for model training. Note that the  $0.1^\circ$  grid used by IMERG and WN3 are offset by  $0.05^\circ$  longitudinally, requiring us to linearly interpolate the dataset. This is only done for training. For evaluation we interpolate all forecasts to the native IMERG grid. We use IMERG starting in 2000. Due to IMERG’s transition to v08, IMERG Final is currently only available until September 2025<sup>8</sup>.

#### A.3.6. PARDIG

We developed our own global precipitation estimate, called PARDIG (Precipitation AI Reanalysis - Densely Inferred GPM). This is based on a separately trained model with architecture similar to WN3, albeit with no noise injection. The input to this model is a 4 hour window of geostationary satellite data, ERA5/HRES pressure level and surface variables, and a mosaic of microwave sounder data; the target of the model is the GPM CORRA measurements (Grecu et al., 2016) at the two hour window centred within the input window. The CORRA product uses a combination of microwave and radar measurements from the core observatory satellite in the GPM program. This satellite is in low Earth orbit and thereby provides sparse data points. The precipitation estimation model uses a similar sparse output head architecture that is used for WN3’s station head (targeting a binned probability distribution). After training, this enables us to make predictions globally. We create  $0.1^\circ$  predictions of instantaneous rain rate every 15 minutes, and then sum those to produce an hourly dataset starting in 1999, which then is used as a target for WN3. In preliminary evaluations we find this precipitation estimate to be substantially more skillful when evaluated against ground-based radar or rain gauges

<sup>8</sup><https://gpm.nasa.gov/data/news/imerg-v08-transition-schedule>



compared to IMERG Final. Further details will be provided in an upcoming paper.

#### A.3.7. MRMS

For evaluation, we also use the Multi Radar Multi Sensor (MRMS) Quantitative Precipitation Estimate Pass-2 product for independent evaluation ([NOAA National Severe Storms Laboratory and NCEP, 2024](#)). This provides hourly precipitation estimates derived mainly from radar but calibrated against rain gauges. We exclude areas with a low radar quality index for evaluation to ensure we are using the highest quality data. We further restrict the evaluation to a spatial domain spanning latitudes 24°N to 50°N and longitudes 126°W to 65°W.

#### A.3.8. IBTrACS gridded fields

We follow the approach used in [Alet et al. \(2026\)](#) to convert IBTrACS data ([Knapp et al., 2010](#); [Gahtan et al., 2024](#)) into gridded targets for WN3, with two small modifications. [Alet et al. \(2026\)](#) used an imputation approach for the wind quadrant radii: for each wind threshold (34kt, 50kt and 64kt), if at least one quadrant radius is not NaN, the rest of the quadrant radii were imputed to the mean of the available quadrant radii. If all of them were NaN, they were all set to zero. We modified this approach, by removing the mean-imputation and only setting a NaN to a zero if the maximum sustained wind speed of the cyclone is below the corresponding wind threshold, i.e. if we can unambiguously determine that the NaN represents a zero quadrant radius. In addition, we perform temporal interpolation on the IBTrACS dataset, to bring it to a 1-hourly resolution, before gridding it. To interpolate the cyclone position, we perform the linear interpolation in 3D Euclidean coordinates and project back to longitude and latitude, while performing regular linear interpolation for the rest of the scalar variables. Finally, the interpolation may introduce unphysical combinations of wind speeds and quadrant radii, e.g. if at an interpolated time the wind speed drops below 34kt, then all 34kt radii should be set to zero. This happens relatively infrequently but we nonetheless account for it by enforcing physical consistency on the ground truth 1-hourly IBTrACS data, following the approach outlined in [Alet et al. \(2026\)](#).

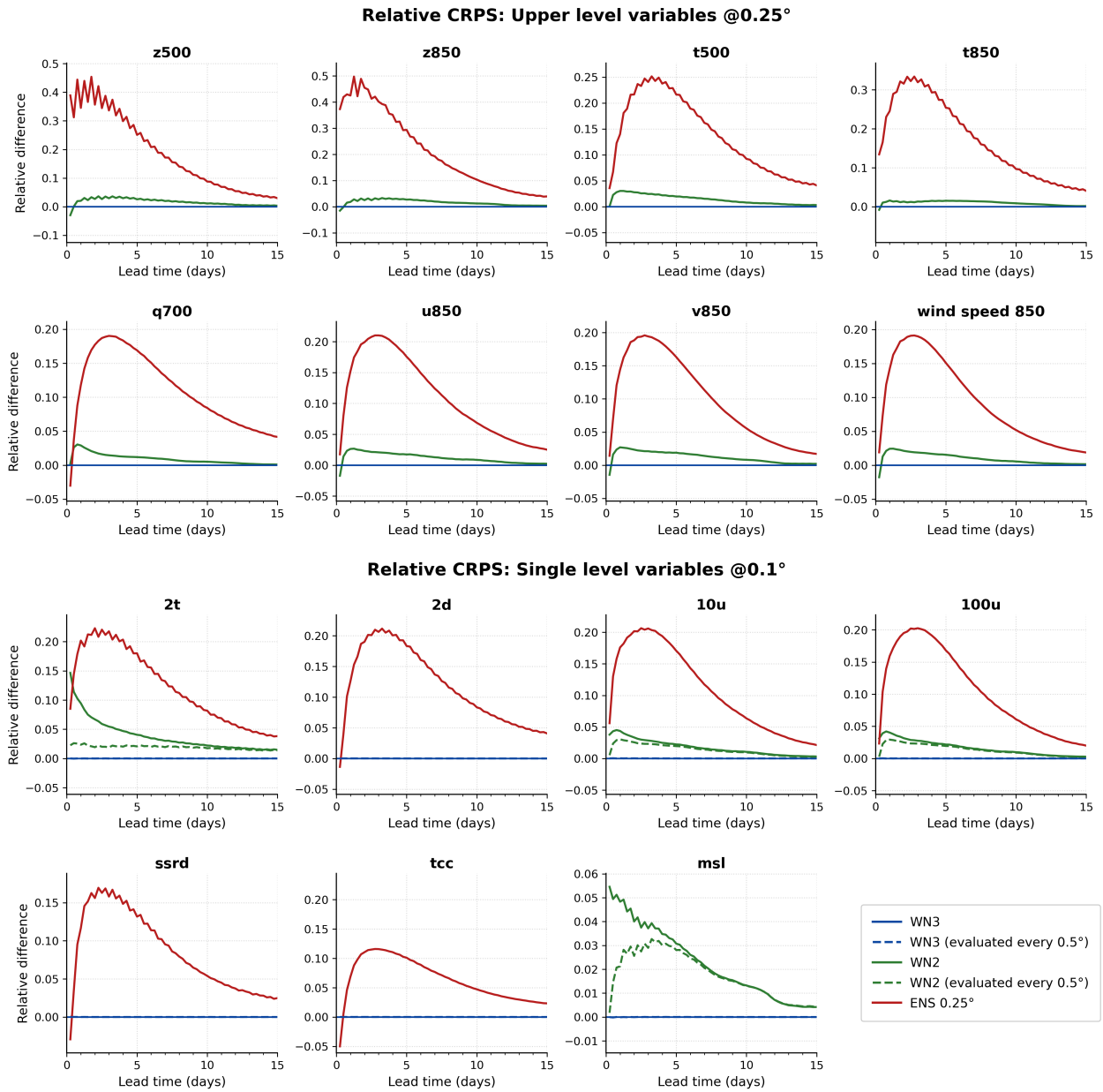


Figure A.2 | Relative CRPS difference with respect to WN3. (top) Variables evaluated at 0.25°. (bottom) Variables evaluated at 0.1°. Additionally, WN3 and WN2 evaluated at common grid points (0.5°; dashed lines) to separate the effect of regridding WN2 from 0.25° to 0.1° from pure model skill. Note that for WN3, the differences are indiscernible. For WN2 the largest differences are for 2t and msl at early lead hours, indicating that there is a substantial penalty from regridding. However, even when removing this penalty, WN3 still outperforms WN2 on those variables. Note that not all variables are available for WN2 and we did not download msl for ENS. 00/12 UTC initializations for 2024.

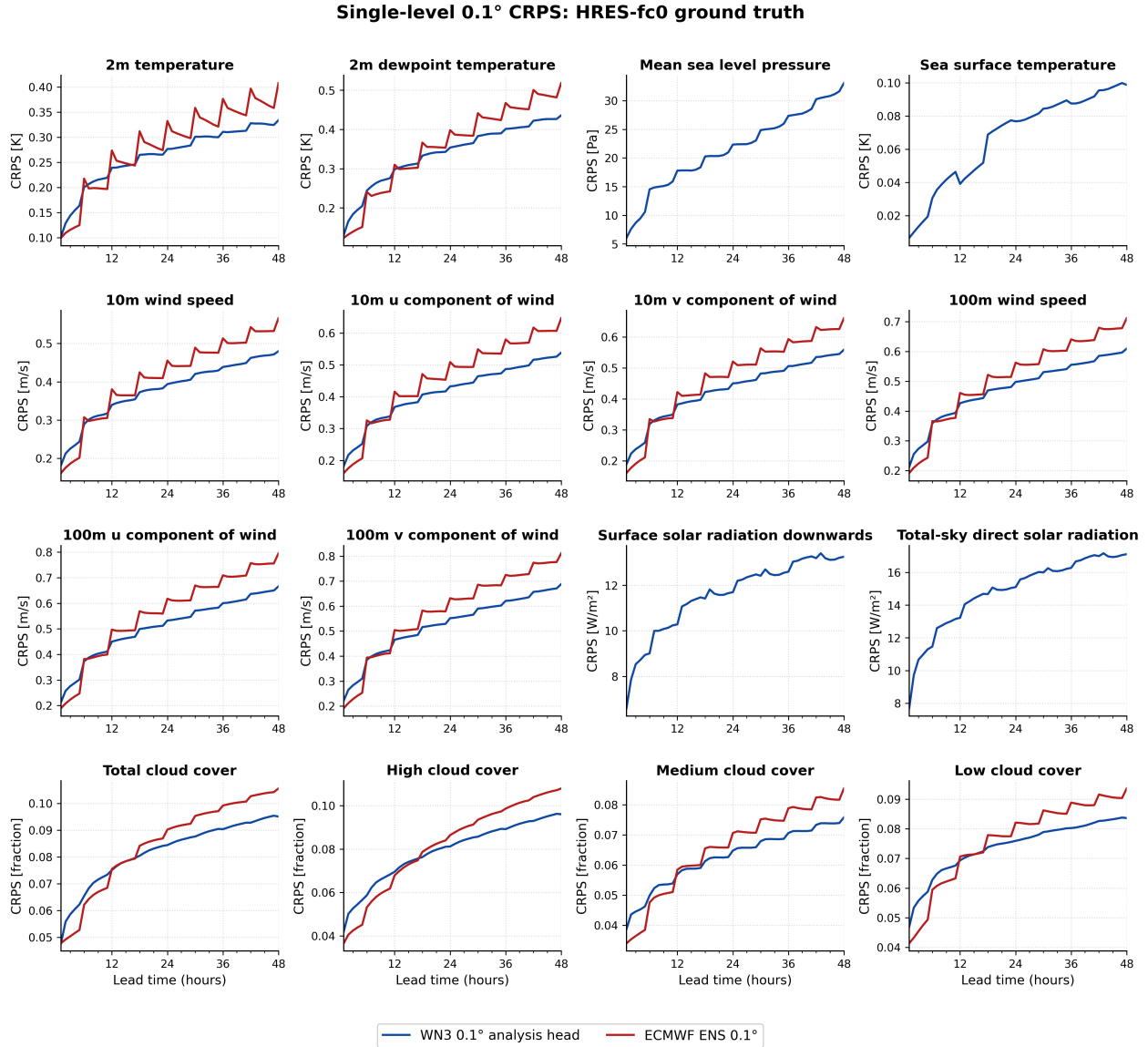


Figure A.3 | Hourly CRPS for single-level variables for the first 48h lead time with HRES-fc0-5 ground truth. The general step shape of the plots, with a substantial step change every 6 hours, is explained by the fact that the ground truth goes from being a continuous HRES forecast trajectory for the first 5 hours to a new analysis state at the next 6h multiple. For 2m temperature, 2m dewpoint, and all wind variables, there is an additional contribution to the spike at 6h multiples which is explained in Section A.3.1. We did not have ENS data available for some variables. 00/12 UTC initializations for 2024.

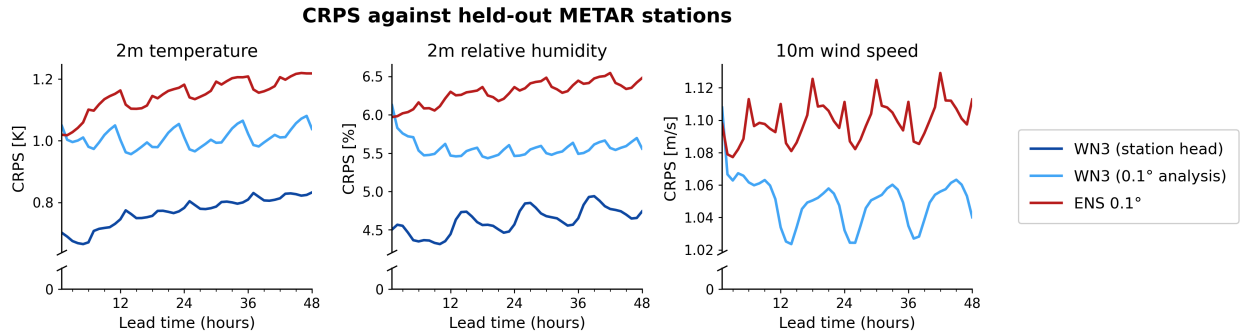


Figure A.4 | Hourly station evaluations against held-out METAR stations. ENS exhibits some spikes in CRPS at 6 hour intervals as discussed in Section A.3.1. WN3 0.1° analysis predictions show an initially large CRPS which is likely due to a mismatch between the initial conditions and the station observations. As ensemble spread is introduced, CRPS sharply improves. 00/12 UTC initializations for 2024.

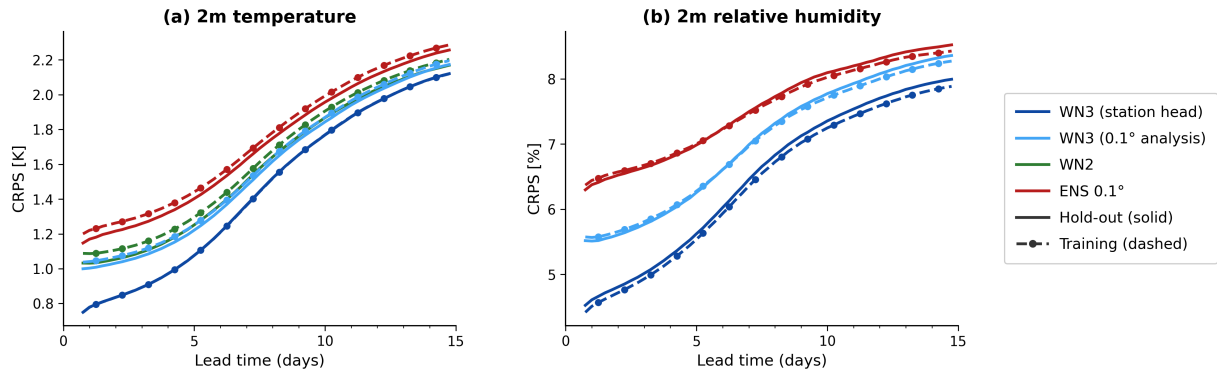


Figure A.5 | Difference between training and hold-out station sets. WN2 and ENS are not specifically optimized for stations. For this reason, we should expect no difference in skill between the two sets. The results show that for those models the training set has a slightly higher error. These differences are simply random in our limited sample size. For the WN3 station head, the two sets have almost equal CRPS for 2m temperature, and slightly lower for 2m relative humidity, indicating that WN3 does, in fact, overfit to the training set.

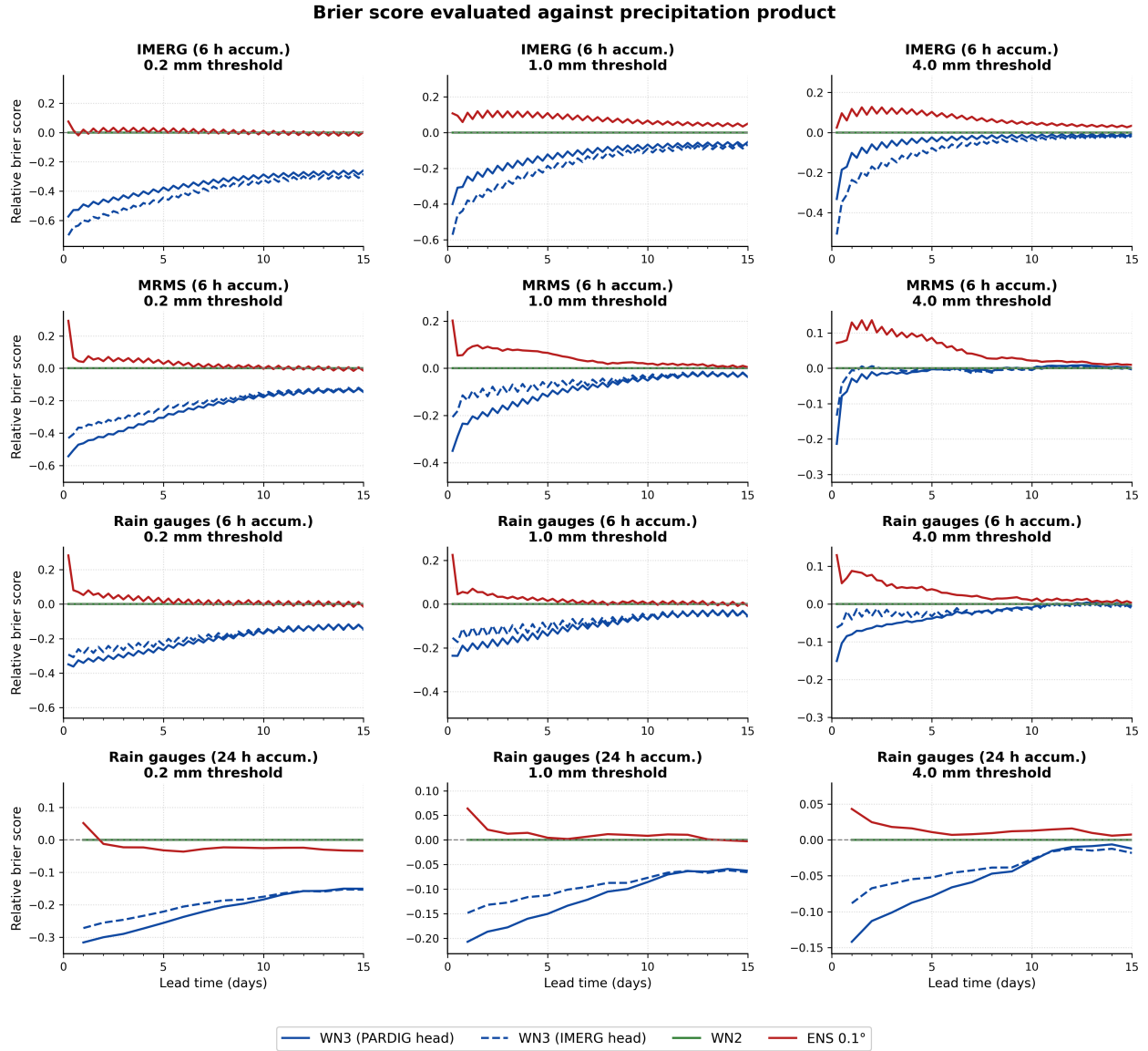


Figure A.6 | Relative precipitation Brier scores for ENS, WN3 (IMERG head), and WN3 (PARDIG head) relative to WN2 (reference at 0.0; negative is better). Results are broken down by accumulation thresholds (0.2, 1.0, and 4.0 mm; columns) and ground truth sources (IMERG, MRMS, and rain gauges; rows). Both WN3 configurations outperform WN2 and ENS across all thresholds and leads. The IMERG-trained head achieves the lowest Brier score against IMERG, whereas the PARDIG head achieves the highest skill against independent MRMS radar and rain gauge observations.

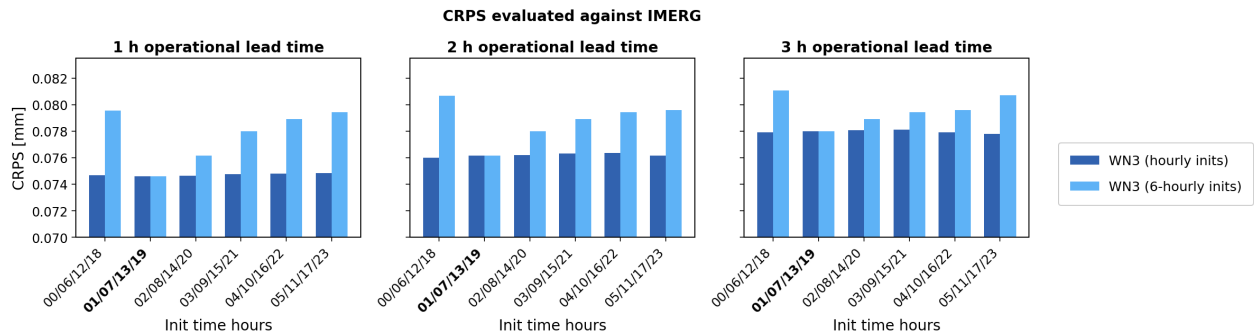


Figure A.7 | Comparison of precipitation CRPS skill for WN3 running on an hourly and 6-hourly update cycle under an assumed 7-hour operational pipeline latency. At 01, 07, 13, and 19 UTC issue times (shown in bold), both setups retrieve the identical nominal forecast (e.g., the 00 UTC run when queried at 07 UTC), resulting in identical skill. The performance divergence peaks at 00, 06, 12, and 18 UTC issue times, where the hourly model benefits from a 5-hour freshness advantage (e.g., retrieving a 05 UTC run at 12 UTC versus the 00 UTC run required by the 6 hourly model).

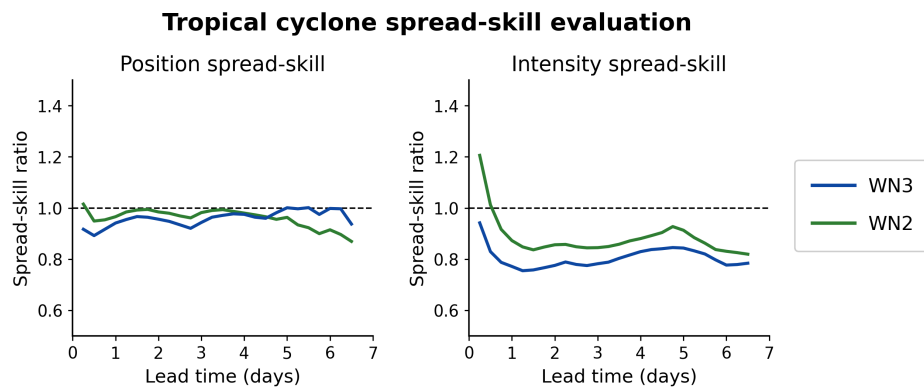


Figure A.8 | Comparison of WN3 and WN2 on global paired homogeneous evaluation, showing spread-skill scores for position and intensity. A perfectly calibrated model should have a spread-skill ratio of 1 (dashed line). In terms of position, spread skill is slightly worse at early lead times and slightly better at longer lead times, compared to WN2. For intensity, we generally observe under-spreading in WN3.



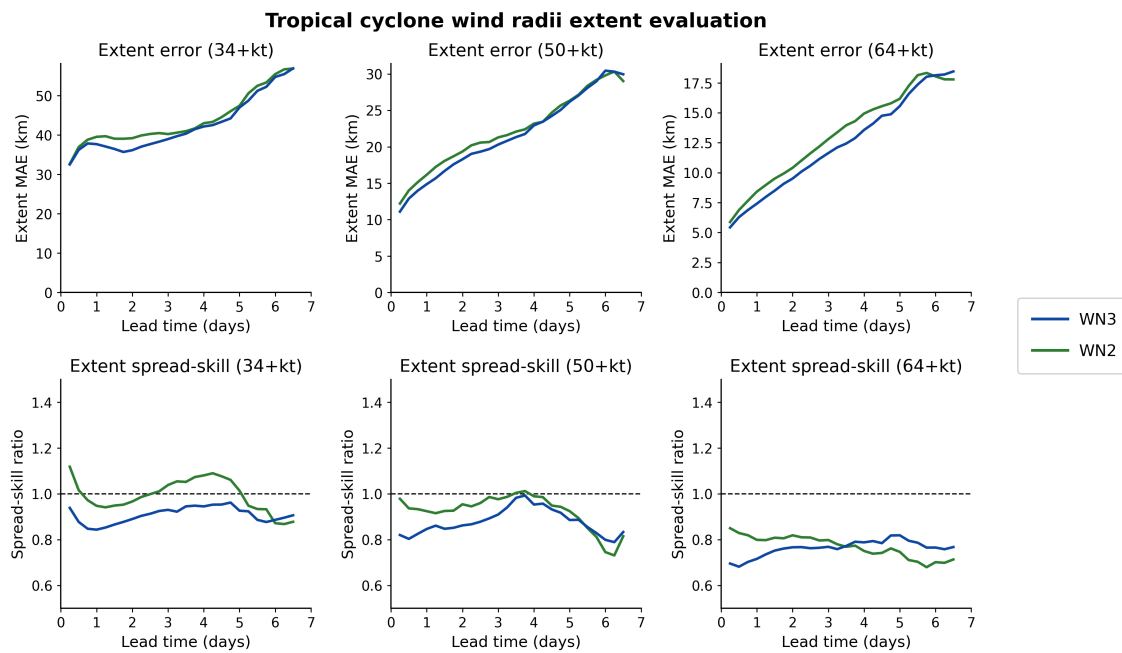


Figure A.9 | Comparison of WN3 and WN2 on global paired homogeneous cyclone forecasting evaluation, showing skill (MAE) and spread-skill scores for quadrant radii. A perfectly calibrated model should have a spread-skill ratio of 1 (dashed line). We average metrics across all four quadrants for a given threshold. WN3 shows modest improvements in terms of quadrant radius MAE across all three thresholds. However, it is also generally more under-spread for 34kt and 50kt quadrant radii predictions compared to WN2. For 64kt quadrant radii, WN3 is slightly more (less) under-spread at the earlier (later) lead times.

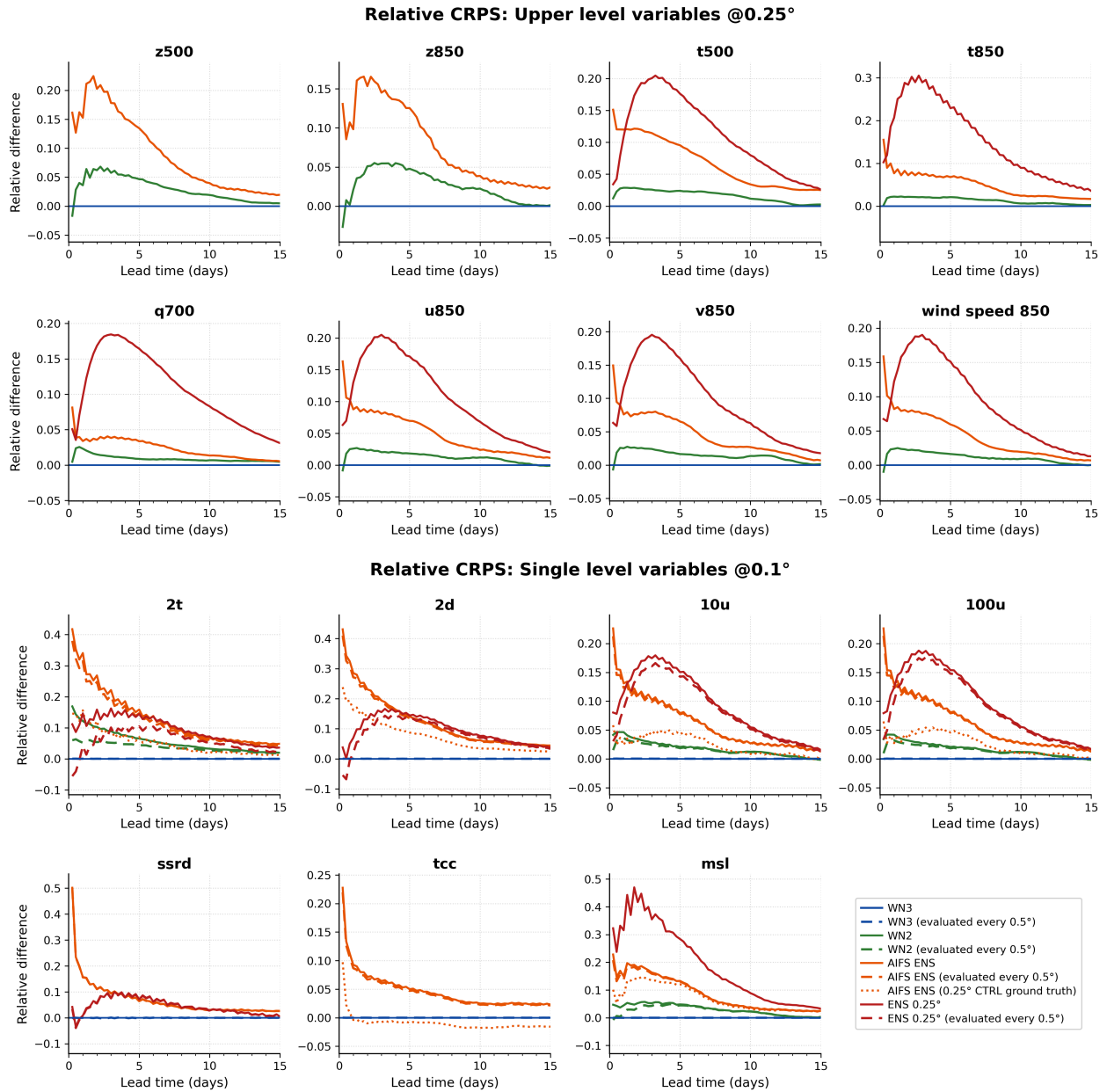


Figure A.10 | As Figure A.2 but for real-time evaluation from July 1 to August 11 2026 (00/12 UTC initializations). Note that the 0.5° evaluation (dashed) has little effect on AIFS ENS. This is likely due to the differences in interpolation described in Section A.3, meaning that the 0.1° HRES-fc0-5 ground truth used does not match the AIFS ENS initial conditions, even at 0.5° "intersecting" points. For this reason, we also added AIFS ENS evaluated against the AIFS ENS control fc0 at 0.25° as ground truth (dotted). Here, the CRPS drops substantially. Note that we did not run evaluations at 0.25° for AIFS ENS for ssrd.

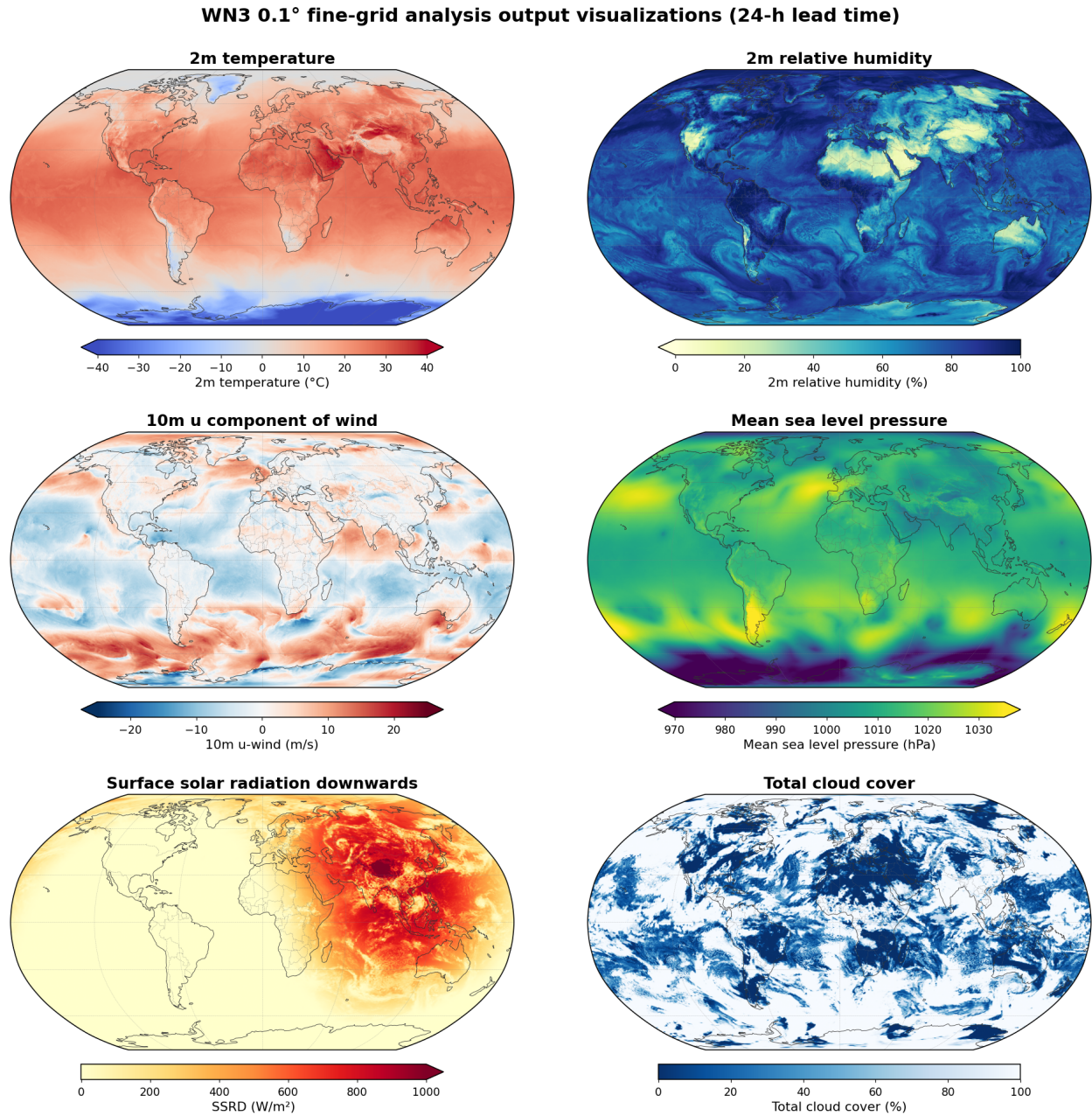


Figure A.11 | Visualizations of the 0.1° analysis output for a single sample for several single-level variables. 24h forecast initialized at July 1 2026 00 UTC.

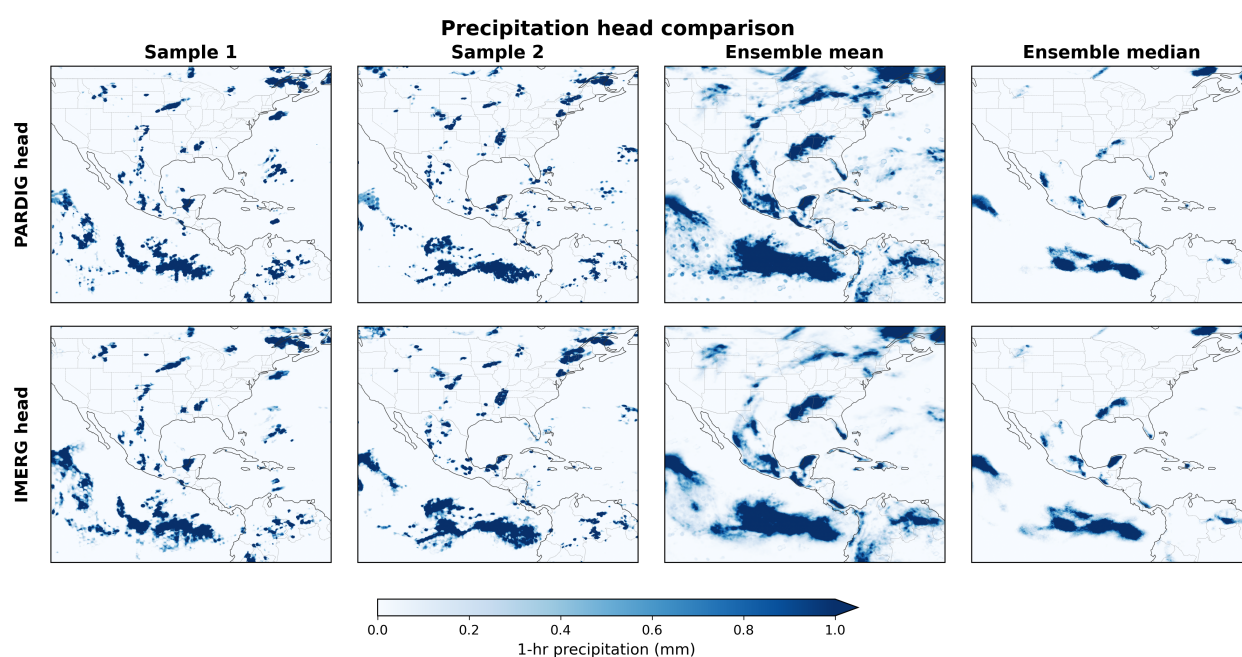


Figure A.12 | Visualizations of the PARDIG (top) and IMERG (bottom) output heads, showing two samples, the ensemble mean and the ensemble median. 24h forecast of 1h precipitation accumulation initialized at July 1 2026 00 UTC.



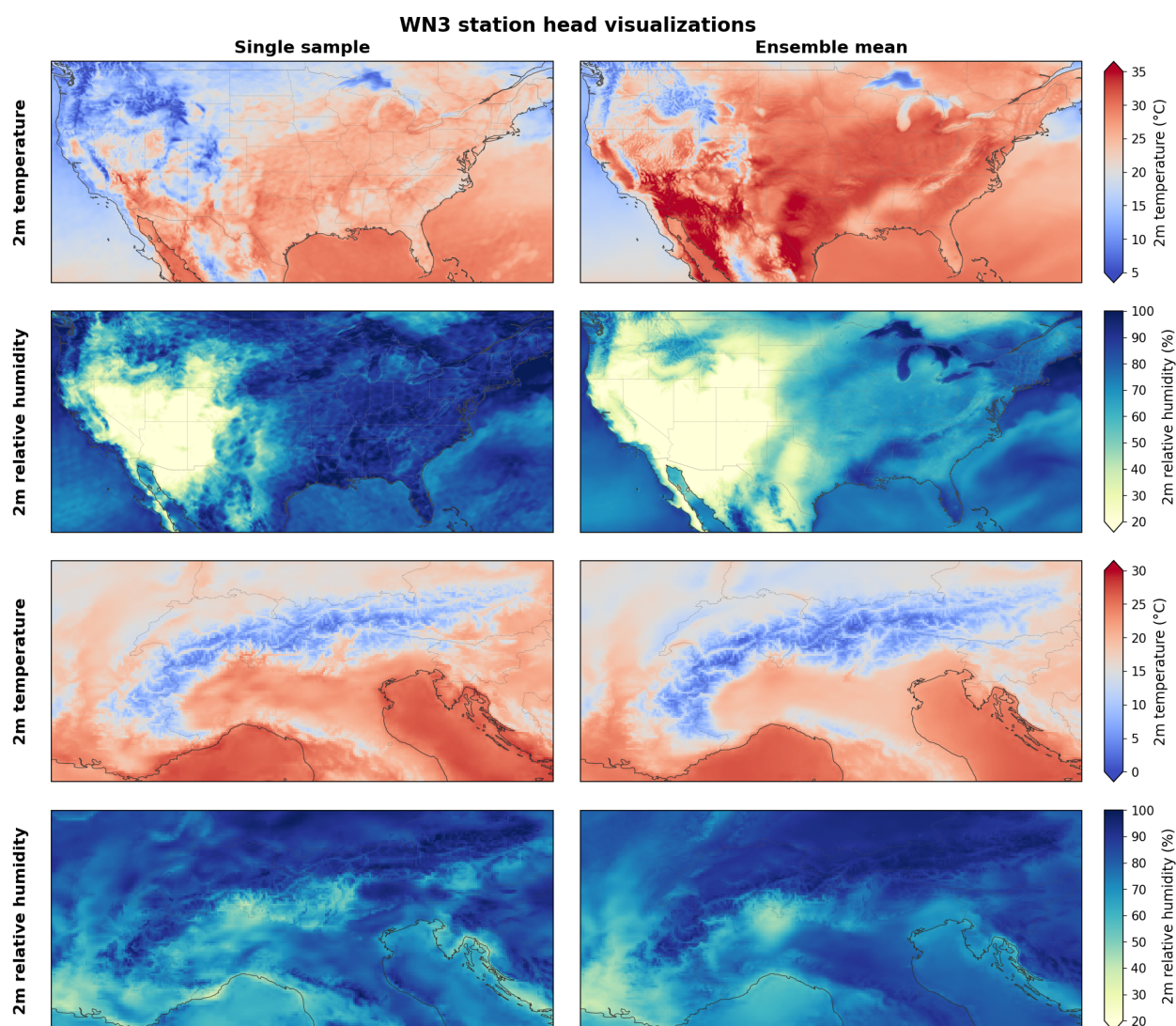


Figure A.13 | Visualizations of the 0.05° station head predictions for the United States (top two rows) and the Alpine region (bottom two rows). Left column shows an individual sample, right column shows the ensemble mean. 24h forecast initialized at July 1 2026 00 UTC.