Integrating GNSS-Derived Zenith Wet Delay into a Weather Foundation Model Improves Precipitation Forecasting

Leonardo Trentini[1], Fanny Lehmann[2,3], Laura Crocetti[1], Benedikt Soja[1,2]

[1]Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland
[2]ETH AI Center, ETH Zurich, Switzerland
[3]Computational and Applied Mathematics Laboratory, Seminar for Applied Mathematics, ETH Zurich, Switzerland

Key Points:

- Aurora learns GNSS Zenith Wet Delay as a new variable at a skill level on par with its already-pretrained variables
- Adding GNSS Zenith Wet Delay systematically improves Aurora's precipitation forecasts, with the largest gains for the most extreme events
- Including ZWD makes the predicted precipitation power spectrum more realistic at synoptic and planetary scales

Corresponding author: Leonardo Trentini, ltrentini@ethz.ch

Abstract

Global Navigation Satellite Systems (GNSS), best known for positioning, also serve weather science, as atmospheric water vapour delays their signals. This delay, the Zenith Wet Delay (ZWD), is a direct, all-weather measure of column moisture. Although assimilated into numerical weather prediction for decades, ZWD is not yet used by leading machine learning weather models (MLWM), despite addressing a known deficiency: the underestimation of severe precipitation. Here we present the first integration of GNSS-derived ZWD into Aurora, a state-of-the-art weather foundation model. Our extended Aurora learns ZWD with skill comparable to its pretrained variables. More importantly, including ZWD systematically improves forecasts when fine-tuning for six-hour accumulated precipitation. Gains grow with severity, reaching an 8.8% increase in Equitable Threat Score at the 99th percentile, while the precipitation power spectrum becomes more realistic at synoptic and planetary scales. Direct GNSS observations therefore encode information that MLWM can exploit for high-impact precipitation.

Plain Language Summary

Heavy rainfall is among the most damaging weather phenomena and also among the hardest for forecasts to predict accurately. The new generation of Artificial Intelligence weather models, which have begun to rival traditional forecasts in many respects, share this weakness when it comes to extreme rain. One reason is that they learn from datasets containing only a limited set of measurements. Here we add a new ingredient to one of these models, Aurora: water vapour information collected by ground-based Global Navigation Satellite System receivers. When these signals pass through the atmosphere, water vapor delays them, and this delay, used by weather centers for decades, reveals how much moisture sits in the air column above each receiver. With this signal added, Aurora's rainfall forecasts become more accurate, with the largest gains for the most intense rain events. The result opens a path to making Artificial Intelligence weather forecasts more reliable in exactly the situations where reliable forecasts matter most.

1 Introduction

The past years have witnessed a transformation in atmospheric forecasting driven by deep learning. Data-driven models such as FourCastNet (Pathak et al., 2022), Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023) and the probabilistic GenCast (Price et al., 2025), among others, have demonstrated global medium-range forecast skill competitive with operational Numerical Weather Prediction (NWP) at a fraction of the computational cost (Rasp et al., 2024). More recently, the paradigm has shifted toward weather *foundation models*: large systems pretrained on heterogeneous atmospheric datasets that can be fine-tuned for specific downstream tasks (Nguyen et al., 2023; Bodnar et al., 2025; Ozdemir et al., 2026; Schmude et al., 2024). Aurora (Bodnar et al., 2025), among the most capable of these, achieves state-of-the-art performance on standard benchmark variables while its Perceiver-based encoder architecture is designed to accommodate new data types without requiring modifications to the core architecture. Yet, despite this architectural flexibility, most existing weather foundation models are pretrained primarily on gridded reanalysis products such as ERA5 (Hersbach et al., 2020) and do not incorporate in-situ observations from geodetic networks. A parallel line of work is broadening the data foundation of these models, either by extending model encoders to heterogeneous observational data (Ozdemir et al., 2026) or by learning end-to-end from raw observations rather than reanalysis (Vaughan et al., 2025), underscoring the value of bringing observational streams directly into machine learning weather pipelines.

Ground-based Global Navigation Satellite System (GNSS) receivers, beyond their original purpose of providing accurate localization, can provide a unique source of atmospheric

information currently unused by leading machine learning weather models. As a GNSS signal propagates through the troposphere it accumulates a path delay proportional to the integral of refractivity along the line of sight. Mapping this delay to the zenith direction and removing the hydrostatic contribution isolates the Zenith Wet Delay (ZWD), a vertical integral of moist refractivity that depends almost entirely on water vapour. ZWD maps linearly to the Integrated Water Vapour through $\mathrm{IWV} = \Pi(T_m)\,\mathrm{ZWD}$, where the dimensionless coefficient Π varies only weakly with the column-mean atmospheric temperature T_m (Bevis et al., 1992, 1994), making ZWD a direct, weather-independent observable of column water vapour, validated against radiosondes and reanalyses (Wang & Zhang, 2008). Growing global networks now comprise thousands of continuously operating GNSS stations (Yuan et al., 2023; Blewitt et al., 2018), providing sub-hourly ZWD estimates under all sky conditions, including during active precipitation. GNSS ZWD has been assimilated into operational NWP for several decades (Vedel & Huang, 2004; Guerova et al., 2016) and used to monitor phenomena ranging from Alpine Foehn (Aichinger-Rosenberger et al., 2022), the El Niño Southern Oscillation (Crocetti, Schartner, Schindler, et al., 2024) to convective storms (Aichinger-Rosenberger et al., 2023). While GNSS-derived tropospheric delays have themselves been the target of AI foundation model approaches (Ding et al., 2024), the inverse direction - feeding GNSS-derived ZWD into a state-of-the-art atmospheric foundation model - has not been attempted to our knowledge. ZWD is moreover not represented as a state variable in ERA5.

This omission is particularly consequential for precipitation forecasting: heavy precipitation drives some of the most damaging weather hazards while remaining among the hardest forecast targets for both numerical and machine learning weather models (Olivetti & Messori, 2024). Precipitation is intermittent, strongly non-Gaussian, represented with substantial bias in ERA5 (Lavers et al., 2022), and strongly modulated by the fine-scale spatial and temporal distribution of column moisture (Lam et al., 2023; Pasche et al., 2025; Zhang et al., 2025). Because reanalysis humidity fields are smoothed by the assimilation cycle, foundation models trained on them inherit a filtered water-vapour representation and underestimate precipitation variability at convective scales - and independent verification against GNSS networks has indeed revealed systematic water vapour biases across multiple AI weather models (Ding et al., 2025). GNSS ZWD, as a direct, unfiltered integral of wet refractivity, captures column-moisture variability tied to the moisture-flux convergence and convective triggering preceding heavy rainfall - a complementary, physically grounded observable that reanalysis humidity fields partly smooth out.

Here we present the first integration of GNSS-derived ZWD into a weather foundation model and investigate whether this integration improves precipitation forecasting skill. Using ZWDX (Crocetti, Schartner, Wareyka-Glaner, et al., 2024), a global gridded ZWD product produced by an XGBoost model trained on over 19,000 GNSS stations, we extend Aurora with ZWD as a new surface variable and fine-tune it for six-hour accumulated precipitation - a variable absent from pretraining - with and without ZWD as an auxiliary input. The contrast between these two precipitation models directly isolates the contribution of GNSS ZWD to forecast skill, especially for extreme events, and addresses the broader question of how geodetic observing networks can be exploited by the emerging generation of machine learning atmospheric models.

2 Data and Methods

2.1 Aurora Weather Foundation Model

Aurora (Bodnar et al., 2025) is a deep learning weather foundation model pretrained on ERA5 (Hersbach et al., 2020), two climate models from the Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al., 2016) and operational forecasts. Given the atmospheric state at times $t - \Delta t$ and t, it predicts the state at $t + \Delta t$ on a 0.25° global grid at $\Delta t = 6$-hour intervals. Surface and pressure-level variables are encoded into a shared latent

representation, so new variables can be added by fine-tuning the model with an extended set of input and output variables.

2.2 ZWDX Global Gridded ZWD Product

To obtain the point-wise GNSS observations on a dense grid required by Aurora, we used ZWDX (Crocetti, Schartner, Wareyka-Glaner, et al., 2024; Crocetti, Schartner, Zus, et al., 2024), an XGBoost model trained on over 19,000 stations from the Nevada Geodetic Laboratory (Blewitt et al., 2018). It learns the mapping from ERA5 specific humidity, location and time to GNSS-derived ZWD. The predictor inherits the spatio-temporal fidelity of the underlying station observations, while being able to provide ZWD values at any arbitrary point, including a regular grid. Because ZWDX is XGBoost-predicted from ERA5 inputs and station coordinates, GNSS information enters Aurora through the training-target labels rather than as raw, independent observations. We exploit the ZWDX product on a 0.25° grid (identical to ERA5) at 6-hour intervals to obtain spatially complete ZWD fields aligned with Aurora's pretraining inputs. Working from a gridded product, rather than raw station data, isolates the value of the GNSS ZWD from any additional gains attributable to a station-wise-input architecture.

2.3 Precipitation Dataset

Six-hour accumulated total precipitation is sourced from MSWEP V2 (Beck et al., 2019), a globally merged dataset combining reanalysis, satellite retrievals, and gauge observations at 0.1° spatial resolution, regridded to the Aurora 0.25° grid for training by bilinear interpolation following Lehmann et al. (2025) (the regridding choice and its relation to mass conservation are detailed in Text S9). MSWEP is preferred over ERA5 precipitation because ERA5 exhibits well-documented systematic biases in precipitation intensity and regional distribution (Beck et al., 2019; Lavers et al., 2022). Total precipitation was not included in Aurora's pretraining; predicting it therefore constitutes a genuinely new downstream task for the model, providing a demanding and realistic test of the value of ZWD as auxiliary information. Precipitation is log-transformed as $\log(1 + x/\epsilon)$ where $\epsilon = 10^{-5}$, following Rasp et al. (2024).

2.4 Experimental Design

The fine-tuning protocol comprises two stages (Figure S1; per-experiment architecture diagrams in Figure S2). Step 1 is a stand-alone experiment in which pretrained Aurora is fine-tuned on the full ERA5 surface and pressure-level state augmented with ZWDX-gridded ZWD as a new surface variable, establishing that the architecture can learn ZWD at a skill comparable to its standard pretrained variables. Step 2 consists of two parallel precipitation fine-tuning experiments. *Model A ("With ZWD")* and *Model B (baseline, "Without ZWD")* are both initialised from the original Aurora pretrained checkpoint and trained with identical optimiser, learning-rate scheduler, number of training steps and data; they differ only in their variable set. In Model A, ZWD and precipitation are added as additional inputs and outputs, while only precipitation is used in Model B (both as input and output). The contrast between them therefore isolates the contribution of GNSS-derived ZWD to precipitation forecast skill. The training objective is a per-variable, latitude-weighted Mean Absolute Error (MAE),

$$\mathcal{L} = \sum_v \lambda_v \frac{1}{N} \sum_i w(\phi_i) \mathrm{MAE}(\hat{x}_v, x_v), \quad w(\phi) = \frac{\cos\phi}{\langle \cos\phi \rangle}, \tag{1}$$

where $\hat{x}_v$ and x_v are the predicted and target fields for variable v on the 0.25° global grid, λ_v is a per-variable loss weight, ϕ_i is the latitude of grid cell i, and $w(\phi)$ accounts for the decreasing area of grid cells towards the poles; the standard pretrained variables retain their pretraining weights, MSWEP precipitation enters with weight 1, and ZWD enters with a

tunable weight λ_{ZWD} set to 2 in the main analysis. Precipitation gains are largely insensitive to this weight ($\lambda_{\mathrm{ZWD}} \in \{1, 2, 3, 10\}$ agree to within $\sim$1% on most metrics; Text S4).

We use the Aurora large variant ($\sim$1.3 B parameters) for the main analysis. An alternative sequential protocol - initialising Model A from the Step 1 ZWD-augmented checkpoint - was also tested in preliminary experiments and yielded no additional skill gain. A complementary three-way ablation using the small variant ($\sim$110 M parameters) is presented in Text S6, and a direct cross-scale comparison between the two architectures in Text S7.

Hardware and step counts for all experiments are reported in Text S3.

Training data span 2010-2018, with the lower bound set by the start of ZWDX availability; January 2019-March 2020 is used for validation, and April-December 2020 is held out for testing. All results in the following are reported on the test set.

2.5 Evaluation Metrics

Standard verification metrics are used throughout: MAE, Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and their normalised counterparts for pointwise accuracy; Pearson correlation R for linear association between predicted and target fields; the Fractions Skill Score (FSS) (Roberts & Lean, 2008) at the 95th percentile of the climatological distribution for spatial structure; and the Equitable Threat Score (ETS) (Schaefer, 1990) at the 75th, 90th, 95th and 99th percentiles for threshold-based skill at increasing event severity. Spectral fidelity is assessed via the latitude-weighted zonal power spectrum and summarised by the Log Spectral Distance (LSD) (Gray & Markel, 1976; Lam et al., 2023) within four wavelength bands: planetary (5,000 - 20,000 km), synoptic (1,000 - 5,000 km), upper mesoscale (250 - 1,000 km) and lower mesoscale (10 - 250 km). Full definitions are given in Text S2.

3 Results

3.1 Fine-Tuning Brings ZWD to Parity with Already-Pretrained Variables

To assess whether Aurora can accurately predict ZWD as an auxiliary output, we compare its Step 1 skill on ZWD against the four standard surface variables seen throughout pretraining (10 m zonal and meridional wind components *u10* and *v10*, 2 m air temperature *t2m*, and mean sea-level pressure *MSLP*) and against specific humidity (q), the pressure-level moisture quantity that is ZWD's most direct counterpart. All variables in Table 1 are evaluated on the same Step 1 fine-tuned model on the held-out test set at 6-hour lead time, isolating how well a previously unseen variable is learned relative to a brief refinement of variables already pretrained.

ZWD is learned essentially at parity with specific humidity ($R = 0.998$ for both, FSS_{95} = 0.985 vs 0.986), the level-resolved quantity from which the column integral can in principle be derived. In physical units it is predicted with an MAE of 3.03 mm and an RMSE of 5.23 mm, and its relative MAE of 1.92% (Table 1) is of the same order as those of the pretrained surface fields. Its ETS (Table 1, right) is competitive at moderate thresholds (75th - 95th percentile) and degrades more steeply only at the 99th percentile, as expected from the rarity of extreme column-moisture-loading events, while its spectral skill lies well within the envelope of the pretrained variables (Text S11). ZWD is thus learned to comparable skill through Aurora's standard variable-embedding mechanism, without any bespoke modules, dedicated prediction heads, or structural modifications to the encoder, backbone, or decoder.

Table 1: Step 1 deterministic and threshold-based skill of ZWD compared to Aurora's standard pretrained variables (u10, v10, t2m, MSLP) and specific humidity (q, averaged over all 13 pressure levels), all evaluated after the same Step 1 fine-tuning on the held-out test set. Arrows indicate whether lower or higher is better.

	Deterministic			Equitable Threat Score (ETS)			
Var.	$R \uparrow$	$\mathrm{FSS}_{95} \uparrow$	Rel. MAE $\downarrow$	$\mathrm{ETS}_{75} \uparrow$	$\mathrm{ETS}_{90} \uparrow$	$\mathrm{ETS}_{95} \uparrow$	$\mathrm{ETS}_{99} \uparrow$
MSLP	0.9998	0.996	0.013%	0.964	0.960	0.953	0.941
t2m	0.9997	0.977	0.067%	0.956	0.846	0.828	0.856
u10	0.9968	0.992	4.72%	0.916	0.920	0.907	0.847
v10	0.9956	0.989	6.06%	0.894	0.886	0.884	0.861
q	0.9982	0.986	2.99%	0.9381	0.9352	0.8887	0.7083
ZWD	0.9983	0.985	1.92%	0.947	0.912	0.851	0.673

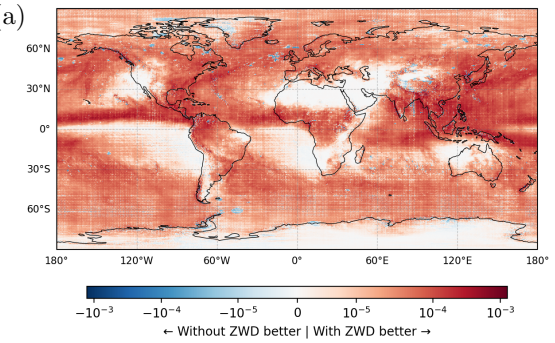
3.2 ZWD Improves Precipitation Forecasting Skill

Having established that ZWD is skillfully learned, we now assess whether its inclusion as an auxiliary target improves precipitation forecasting in the 1.3 B Aurora model. Models A and B are evaluated on the held-out test set using the metric suite introduced in Section 2.5. We focus on a 6-hour lead time in the main text, as ZWD is a short-lived signature of the instantaneous moisture column, so its predictive value is expected to be greatest at short range. For completeness, the evolution of forecast skill over longer autoregressive rollouts (out to 114 h, without any rollout fine-tuning) is characterised in Text S10, where the ZWD advantage in threshold-based skill is preserved through the early-to-intermediate lead times. Corresponding results for the 110 M small Aurora variant are reported in Text S6 (small-model ablation) and Text S7 (direct cross-scale comparison). Unless otherwise noted, every Step 2 metric reported below is the mean over ten training checkpoints saved at the same training steps for both models on the late-training plateau, quoted with sample standard deviation (s.d.) across those checkpoints where space permits. Because Models A and B share initialisation, optimiser, schedule, step budget and data and differ only in ZWD, checkpoints saved at the same step are matched pairs. Therefore, we assess the significance of the With-versus-Without-ZWD difference with a checkpoint-wise paired t-test across these matched steps, reporting alongside it the number of checkpoints at which ZWD wins (Text S8).

3.2.1 Deterministic Metrics

Figure 1b reports deterministic skill on the test period. Including ZWD improves the pointwise error metrics: MAE decreases by 0.7%, RMSE by 1.8% and MSE by 3.7%, with the largest gains in MSE-based metrics indicating a particular reduction of large errors. The FSS at the 95th percentile improves by 1.5% and the Pearson R is essentially unchanged (-0.2%), so the linear correlation of predictions with observations is preserved while their spatial and intensity structure are improved. These gains come at only a modest cost to the pretrained surface variables: averaged over the ten checkpoints, all retain $R \geq 0.997$, while only the 10 m winds degrade significantly (by $\sim$1%), while 2 m temperature, mean sea-level pressure and specific humidity stay within checkpoint-to-checkpoint noise (per-variable values and paired tests in Text S5 and Text S8). The precipitation gain, by contrast, is reproduced at 8-9 of the ten checkpoints and is significant under a paired test, an asymmetric trade-off in favour of including ZWD.

(a)



(b)

Metric	Without ZWD	With ZWD	Rel. Δ
MAE [mm]	0.195 ± 0.007	$\mathbf{0.193 \pm 0.004}$	−0.7%
RMSE [mm]	1.015 ± 0.034	$\mathbf{0.997 \pm 0.019}$	−1.8%
MSE [mm^2]	1.032 ± 0.071	$\mathbf{0.994 \pm 0.038}$	−3.7%
Pearson R	$\mathbf{0.853 \pm 0.001}$	0.851 ± 0.001	−0.2%
FSS 95%	0.917 ± 0.022	$\mathbf{0.931 \pm 0.011}$	+1.5%

Figure 1: Step 2 deterministic precipitation skill on the held-out test set. (a) Global spatial distribution of the 6-hour precipitation RMSE difference for the main-analysis checkpoint (red: "With ZWD" improves upon the baseline "Without ZWD"; blue: baseline outperforms). (b) Summary deterministic metrics for Model A ("With ZWD") and Model B ("Without ZWD"), reported as the mean ± s.d. over ten training checkpoints.

The global map of RMSE differences (Figure 1a) shows that ZWD reduces precipitation RMSE across most of the globe, with the largest gains in the tropics and mid-latitude storm tracks - regions of strong moisture-flux convergence, active deep convection and high ZWD spatial variability, where the column-moisture signal best discriminates raining from non-raining columns. The few regions where the baseline marginally outperforms are spatially confined and generally coincide with arid climatologies in which precipitation is rare and column moisture is uniformly low, so ZWD adds little discriminating information.

3.2.2 Threshold-Based Skill: ETS

The ETS scores in Figure 2a show that the relative benefit of ZWD increases with the precipitation threshold. At the 75th percentile the gain is +1.2%; at the 99th percentile it reaches +8.8%, and all four gains are significant under a checkpoint-wise paired test ($p \leq 0.013$, Text S8). Figure 2b shows the ETS at the 99th percentile as a function of time over the test period: the ZWD-enriched model outperforms the baseline at essentially every timestep, demonstrating that the improvement is not driven by a small number of outlier events, but is a robust property of the trained model. This pattern is physically intuitive: ZWD's largest excursions coincide with the high-moisture conditions that drive heavy precipitation, so it is most informative in the tail of the distribution, where moisture availability is the binding constraint on rainfall.

To confirm that this benefit manifests during individual high-impact events, and not only in the global aggregate, we examine three extreme cases from distinct basins and regimes: an active South Asian monsoon episode, Typhoon Bavi over East Asia, and Hurricane Delta over Central America (all within the test set, full per-event diagnostics in Text S12). Stratifying the mean precipitation error reduction by observed-precipitation intensity (Figure 2c-e) reveals the same signature in every case: ZWD is neutral or marginally negative at low-to-moderate intensities and strongly positive in the extreme upper tail, reaching a mean error reduction of +1.52, +2.17 and +0.71 mm in the top intensity bin for the three events, respectively. Even the hurricane, whose bulk regional RMSE is essentially unchanged, retains this tail gain (Text S12), confirming that the benefit is localized to the heavy-precipitation regime where the column-moisture signal is most informative.

(a)

Threshold	Without ZWD	With ZWD	Rel. Δ	ZWD wins	paired p
75th percentile	0.695 ± 0.016	$\mathbf{0.703 \pm 0.009}$	+1.2%	9/10	0.013*
90th percentile	0.656 ± 0.022	$\mathbf{0.669 \pm 0.012}$	+1.9%	9/10	0.012*
95th percentile	0.581 ± 0.036	$\mathbf{0.604 \pm 0.020}$	+3.9%	9/10	0.007*
99th percentile	0.395 ± 0.052	$\mathbf{0.430 \pm 0.031}$	+8.8%	8/10	0.004*

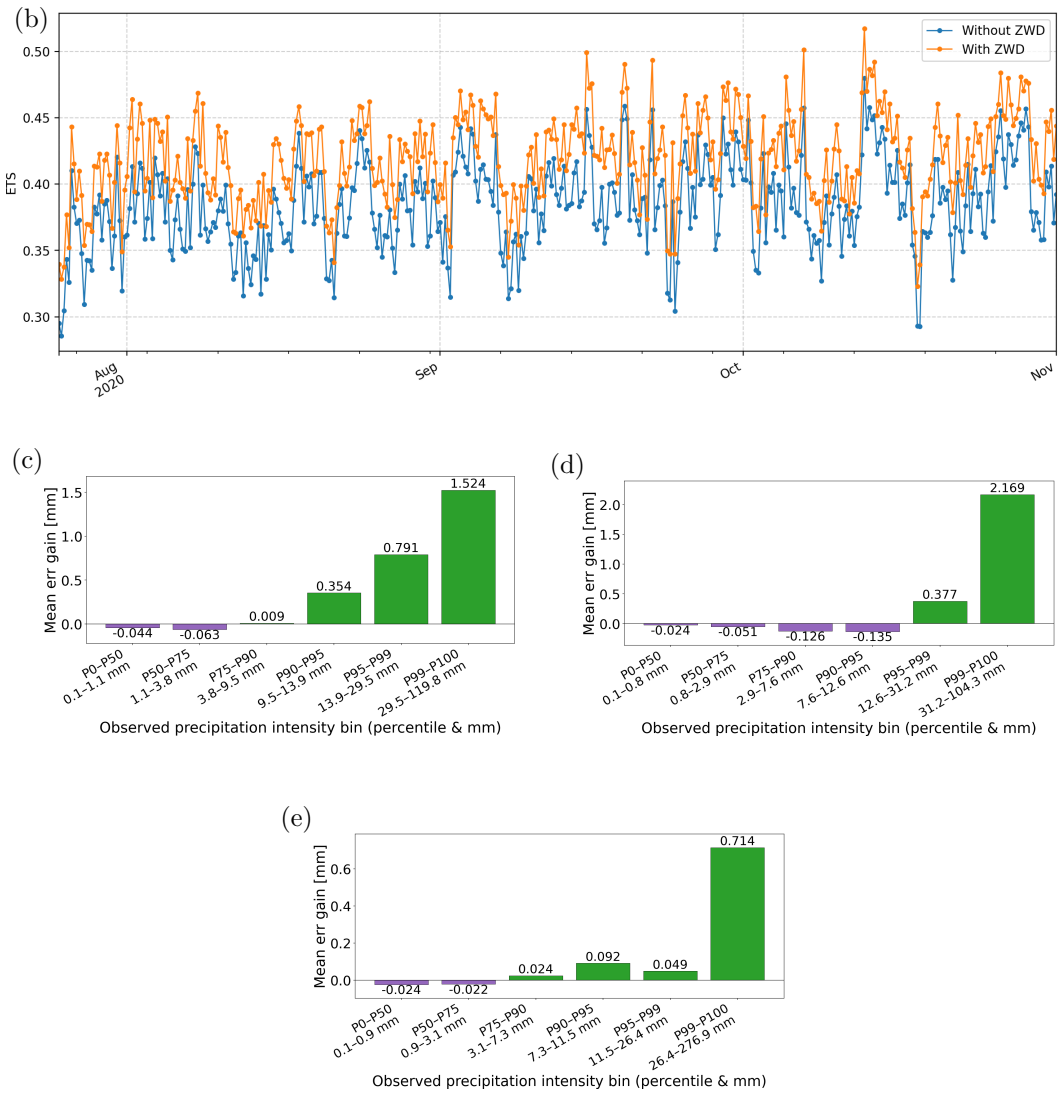
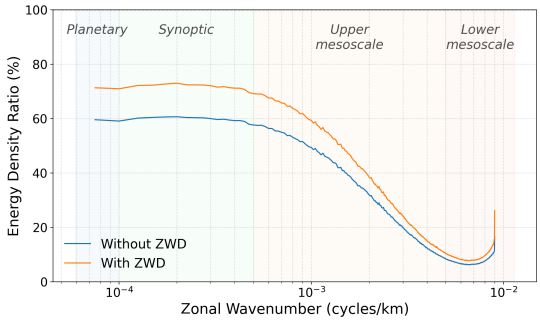


Figure 2: Step 2 threshold-based and extreme-event precipitation skill with and without ZWD. (a) ETS at four climatological percentiles (mean ± s.d. over ten checkpoints; the last two columns give the number of checkpoints favouring ZWD and the two-sided checkpoint-wise paired t-test p-value across matched steps, $^*p < 0.05$; Text S8). (b) ETS at the 99th percentile over part of the test period. (c-e) Precipitation error reduction (positive = ZWD better) versus observed-intensity bin for three extreme events: (c) South Asia monsoon, (d) Typhoon Bavi (East Asia), (e) Hurricane Delta (Central America); the gain concentrates in the upper tail in every case. Per-event diagnostics in Text S12.

(a)



(b)

Band	Without ZWD	With ZWD	Rel. Δ
Planetary	0.191 ± 0.100	$\mathbf{0.115 \pm 0.078}$	−39.9%
Synoptic	0.225 ± 0.110	$\mathbf{0.150 \pm 0.080}$	−33.6%
Upper mesoscale	0.607 ± 0.120	$\mathbf{0.526 \pm 0.092}$	−13.4%
Lower mesoscale	1.018 ± 0.120	$\mathbf{0.961 \pm 0.100}$	−5.6%
Total	0.934 ± 0.120	$\mathbf{0.876 \pm 0.098}$	−6.2%

Figure 3: Step 2 spectral precipitation skill. (a) Ratio of predicted to target latitude-weighted zonal power spectrum by wavenumber band for the main-analysis checkpoint (100% = perfect match). (b) Log Spectral Distance (LSD) for Model A ("With ZWD") and Model B ("Without ZWD"), mean ± s.d. over ten checkpoints; lower is a more faithful spectrum. The reduction is significant in every band (checkpoint-wise paired t-test $p \leq 0.023$; between eight and nine of ten checkpoints favour ZWD). *Total* is LSD integrated over the full wavenumber range, not the sum of band values.

3.2.3 Spectral Skill: Energy Spectra

Including ZWD improves the predicted precipitation power spectrum on every spatial scale (Figure 3). The spectrum ratio (Figure 3a) moves closer to unity for the ZWD model throughout the range, and the LSD decreases in every wavenumber band (Figure 3b), with a total reduction of −6.2%. The absolute LSD reduction is broadly comparable across bands (≈0.06-0.08), so although the relative gain is highest at planetary (−39.9%) and synoptic (−33.6%) wavelengths, this reflects a much smaller baseline error at those scales rather than a benefit confined to them. The model is intrinsically blurrier on the mesoscale, and ZWD corrects the spectrum there by a similar absolute amount.

The improvement is therefore not limited to extreme-event intensity: ZWD also sharpens the spatial structure of the precipitation field, counteracting the over-smoothing typical of deterministic forecasts across the full range of scales. Together with the FSS improvement above, this indicates that ZWD's added information is largely structural rather than amplitude-correcting - improving where and at what scale precipitation is organized.

4 Conclusions

We have demonstrated the first integration of GNSS-derived Zenith Wet Delay into a weather foundation model. ZWD is learned by Aurora at a skill level on par with already-pretrained variables, and including it during fine-tuning for six-hour precipitation systematically improves precipitation forecasts, with the largest gains for the most extreme events (an 8.8% increase in Equitable Threat Score at the 99th percentile) and a substantially more realistic precipitation power spectrum at planetary and synoptic scales. These gains are not only global: zooming in on three distinct high-impact events, the same benefit holds locally, with the error reduction concentrated in the heaviest-rain tail (Figure 2c-e, Text S12). A plausible reason is that ZWD provides a direct, observation-derived column-moisture signal that the model would otherwise have to reconstruct from multi-level humidity, temperature, and pressure fields - a particular advantage for precipitation, which is itself a new downstream task for Aurora. This also reconciles the two improvements. At Aurora's 0.25° resolution, even the most extreme grid-cell rainfall is dominated by organized synoptic-scale systems (atmospheric rivers, fronts, tropical waves) whose moisture supply ZWD directly

constrains. Improving the large-scale spectrum and sharpening extreme-event detection are therefore two sides of the same constraint. The trade-off paid by the standard pretrained variables is modest, their skill essentially preserved (Text S5).

These findings are robust: a loss-weight sweep identifies a clear and stable optimum (Text S4); the precipitation gain is significant against the natural checkpoint-to-checkpoint variability of each model, estimated over ten training checkpoints (Text S8); and a small-model ablation shows that the benefit of ZWD scales inversely with model capacity, so smaller architectures stand to gain even more from ZWD than the 1.3 B model demonstrates here (Text S6 and Text S7).

The present implementation relies on ZWDX, a gridded ZWD product that partially depends on ERA5 inputs and smooths individual station variability. Because its labels nonetheless carry the GNSS station-resolved moisture structure that ERA5 partly filters out (Section 2.2), the gains obtained despite this residual ERA5 dependence are a conservative lower bound on what raw, independent station data could provide. Moving to such raw, station-level data (architecturally supported by the variable-specific ESFM encoder (Ozdemir et al., 2026)) would decouple the GNSS signal from reanalysis and allow the model to exploit the dense coverage achievable with next-generation low-cost receiver networks. Further directions include physics-constrained training that enforces consistency between predicted ZWD and humidity column profiles, deeper case studies of the moisture build-up preceding high-impact events (Text S12), and additional GNSS observables such as horizontal tropospheric gradients; the sub-hourly GNSS update rate (<5 min) also matches convective time scales, motivating nowcasting.

More broadly, these findings show that variables not included in reanalysis products can carry physical information important enough to improve weather foundation models on the most challenging forecast tasks. ZWD is the example here, but the same principle would apply to other complementary observations (radar reflectivity, dense environmental sensor networks). Systematic exploitation of these observational streams is a promising frontier for next-generation atmospheric machine learning.

Acknowledgments
This work was supported under project IDs a122 and a0196 as part of the Swiss AI Initiative, through a grant from the ETH Domain and computational resources provided by the Swiss National Supercomputing Centre (CSCS) under the Alps infrastructure. The authors thank Firat Ozdemir and Yun Cheng for their help in setting up the experiments, Simon Adamov for his assistance with the evaluation. Contributions of L.T. were supported by the Swiss National Science Foundation (SNSF) under grant number 225851. Contributions of F.L. were primarily supported by the ETH AI Center through their ETH AI Center postdoctoral fellowship. The authors used Claude (Anthropic) for language editing and code debugging during manuscript preparation; the authors reviewed all output and take full responsibility for the content.

Conflict of Interest declaration

The authors declare there are no conflicts of interest for this manuscript.

Open Research Section

The global ZWDX product used in this study is described in Crocetti, Schartner, Wareyka-Glaner, et al. (2024) and is available at `https://doi.org/10.1186/s40623-024-02104-6`. MSWEP V2 precipitation data are described in Beck et al. (2019) and are available at `http://www.gloh2o.org/mswep/`. ERA5 reanalysis data are available from the Copernicus Climate Data Store (`https://cds.climate.copernicus.eu`). The weights of

the Aurora model and the inference code are available from the Microsoft Aurora repository (`https://github.com/microsoft/aurora`). Fine-tuning code and trained model checkpoints for this study are available at `https://github.com/swiss-ai/zwd-into-aurora`.

References

Aichinger-Rosenberger, M., Aregger, M., Kopp, J., & Soja, B. (2023). Detecting signatures of convective storm events in GNSS-SNR: Two case studies from summer 2021 in Switzerland. *Geophysical Research Letters*, *50*(21), e2023GL104916. doi: 10.1029/2023GL104916

Aichinger-Rosenberger, M., Brockmann, E., Crocetti, L., Soja, B., & Moeller, G. (2022). Machine learning-based prediction of Alpine foehn events using GNSS troposphere products: first results for Altdorf, Switzerland. *Atmospheric Measurement Techniques*, *15*(19), 5821–5839. doi: 10.5194/amt-15-5821-2022

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I. J. M., ... Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. doi: 10.1175/BAMS-D-17-0138.1

Bevis, M., Businger, S., Chiswell, S., Herring, T. A., Anthes, R. A., Rocken, C., & Ware, R. H. (1994). GPS meteorology: Mapping zenith wet delays onto precipitable water. *Journal of Applied Meteorology*, *33*(3), 379–386. doi: 10.1175/1520-0450(1994)033⟨0379:GMMZWD⟩2.0.CO;2

Bevis, M., Businger, S., Herring, T. A., Anthes, R. A., & Ware, R. H. (1992). GPS meteorology: Remote sensing of atmospheric water vapor using the global positioning system. *Journal of Geophysical Research: Atmospheres*, *97*(D14), 15787–15801. doi: 10.1029/92JD01517

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*, 533–538. doi: 10.1038/s41586-023-06185-3

Blewitt, G., Hammond, W. C., & Kreemer, C. (2018). Harnessing the GPS data explosion for interdisciplinary science. *Eos*, *99*. doi: 10.1029/2018EO104623

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., ... Perdikaris, P. (2025). A foundation model for the Earth system. *Nature*, *641*(8065), 1180–1187. doi: 10.1038/s41586-025-09005-y

Crocetti, L., Schartner, M., Schindler, K., Schneider, R., & Soja, B. (2024). Modelling the Troposphere with Global Navigation Satellite Systems, Meteorological Data and Machine Learning. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium* (p. 1689-1692). doi: 10.1109/IGARSS53475.2024.10640441

Crocetti, L., Schartner, M., Wareyka-Glaner, M. F., Schindler, K., & Soja, B. (2024). *ZWDX: a global zenith wet delay forecasting model using XGBoost* (Vol. 76) (No. 1). doi: 10.1186/s40623-024-02104-6

Crocetti, L., Schartner, M., Zus, F., Zhang, W., Moeller, G., Navarro, V., ... Soja, B. (2024). Global, spatially explicit modelling of zenith wet delay with XGBoost. *Journal of Geodesy*, *98*(4). doi: 10.1007/s00190-024-01829-2

Ding, J., Chen, W., Chen, J., Wang, J., Zhang, Y., Bai, L., ... Weng, D. (2025). Spatiotemporal inhomogeneity of accuracy degradation in AI weather forecast foundation models: A GNSS perspective. *International Journal of Applied Earth Observation and Geoinformation*, *139*, 104473. doi: 10.1016/j.jag.2025.104473

Ding, J., Mi, X., Chen, W., Chen, J., Wang, J., Zhang, Y., ... Tang, W. (2024). Forecasting of tropospheric delay using AI foundation models in support of microwave remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–19. doi: 10.1109/TGRS.2024.3488727

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*,

9(5), 1937–1958. doi: 10.5194/gmd-9-1937-2016

Gray, A. H., & Markel, J. D. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(5), 380–391. doi: 10.1109/TASSP.1976.1162849

Guerova, G., Jones, J., Douša, J., Dick, G., de Haan, S., Pottiaux, E., ... Bender, M. (2016). Review of the state of the art and future prospects of the ground-based GNSS meteorology in Europe. *Atmospheric Measurement Techniques*, *9*(11), 5385–5406. doi: 10.5194/amt-9-5385-2016

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. doi: 10.1002/qj.3803

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*, eadi2336. doi: 10.1126/science.adi2336

Lavers, D. A., Simmons, A., Vamborg, F., & Rodwell, M. J. (2022). An evaluation of era5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, *148*(748), 3152-3165. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4351` doi: https://doi.org/10.1002/qj.4351

Lehmann, F., Ozdemir, F., Soja, B., Hoefler, T., Mishra, S., & Schemm, S. (2025). *Finetuning a weather foundation model with lightweight decoders for unseen physical processes.* Retrieved from `https://arxiv.org/abs/2506.19088`

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th international conference on machine learning.*

Olivetti, L., & Messori, G. (2024). Advances and prospects of deep learning for medium-range extreme weather forecasting. *Geoscientific Model Development*, *17*(6), 2347–2358. doi: 10.5194/gmd-17-2347-2024

Ozdemir, F., Cheng, Y., Mohebi, S., Lehmann, F., Adamov, S., Zhang, Z., ... Salzmann, M. (2026). *Earth system foundation model (esfm): A unified framework for heterogeneous data integration and forecasting.* Retrieved from `https://arxiv.org/abs/2605.00850`

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., & Engelke, S. (2025). Validating deep learning weather forecast models on recent high-impact extreme events. *Artificial Intelligence for the Earth Systems*, *4*(1), e240033. doi: 10.1175/AIES-D-24-0033.1

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... Anandkumar, A. (2022). *FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators.*

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., ... Willson, M. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. doi: 10.1038/s41586-024-08252-9

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., ... Sha, F. (2024). Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS004019. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004019` (e2023MS004019 2023MS004019) doi: https://doi.org/10.1029/2023MS004019

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. doi: 10.1175/2007MWR2123.1

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, *5*(4), 570–575. doi: 10.1175/1520-0434(1990)005⟨0570:TCSIAA⟩2.0.CO;2

Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D. S., Singh, S., ... Ramachandran, R. (2024). *Prithvi wxc: Foundation model for weather and climate.* Retrieved from `https://arxiv.org/abs/2409.13598`

Vaughan, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., ... Turner, R. E. (2025). End-to-end data-driven weather prediction. *Nature*, *641*, 1172–1179. doi: 10.1038/s41586-025-08897-0

Vedel, H., & Huang, X.-Y. (2004). Impact of ground based gps data on numerical weather prediction. *Journal of the Meteorological Society of Japan. Ser. II*, *82*(1B), 459–472. doi: 10.2151/jmsj.2004.459

Wang, J., & Zhang, L. (2008, 05). Systematic errors in global radiosonde precipitable water data from comparisons with ground-based gps measurements. *Journal of Climate - J CLIMATE*, *21*, 2218-2238. doi: 10.1175/2007JCLI1944.1

Yuan, P., Blewitt, G., Kreemer, C., Hammond, W. C., Argus, D., Yin, X., ... Kutterer, H. (2023). An enhanced integrated water vapour dataset from more than 10 000 global ground-based GPS stations in 2020. *Earth System Science Data*, *15*(2), 723–743. doi: 10.5194/essd-15-723-2023

Zhang, Z., Fischer, E., Zscheischler, J., & Engelke, S. (2025). Numerical models outperform AI weather forecasts of record-breaking extremes. *arXiv*.

Supporting Information for

"Integrating GNSS-Derived Zenith Wet Delay into a Weather Foundation Model Improves Precipitation Forecasting"

Contents of this file

Introduction

This supporting information collects the methodological details, sensitivity analyses, robustness checks and qualitative examples accompanying the main text. All references of the form "main text" point to the companion article.

Text S1 Experimental Design Schematic

Figure S1 summarises the two-step fine-tuning protocol described in Section 2.4. Starting from the pretrained Aurora checkpoint, Step 1 fine-tunes the model on the full ERA5 state augmented with ZWDX-gridded ZWD; Step 2 trains two precipitation models - Model A (with ZWD) and Model B (baseline) - that share optimiser, learning-rate schedule, number of steps and data, and differ only in whether ZWD is included as an additional input and auxiliary target.

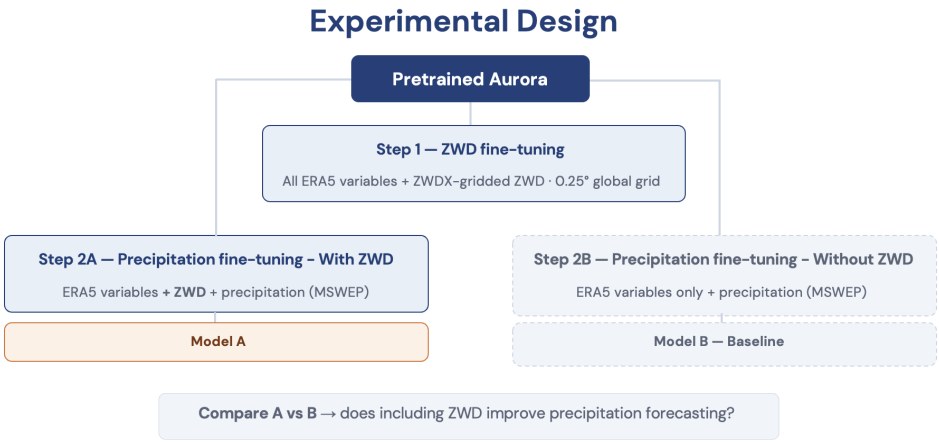


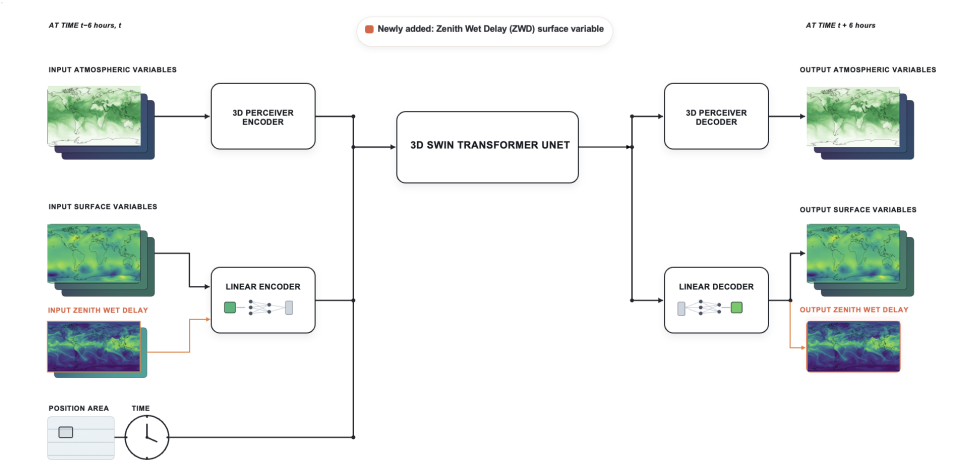
Figure S1: Schematic of the two-step fine-tuning workflow. Both Step 2 precipitation models are initialised from the same pretrained Aurora checkpoint; the only difference between Model A and Model B is the presence of ZWD.

Figure S2 shows, for each fine-tuned model, how the new variables enter the Aurora architecture. In every case the Perceiver-based encoder, 3D Swin Transformer U-Net backbone and decoder are inherited from the pretrained checkpoint; the only modification is the extension of the linear surface encoder and decoder to accommodate the additional input/output channel(s) (Section 2.1 of the main text). The Step 1 model (Figure S2a) adds ZWD as a new surface variable; the Step 2 Baseline (Model B, Figure S2c) adds total precipitation only; and the Step 2 With-ZWD model (Model A, Figure S2b) adds both ZWD and precipitation, ZWD acting as both an auxiliary input and an auxiliary prediction target.

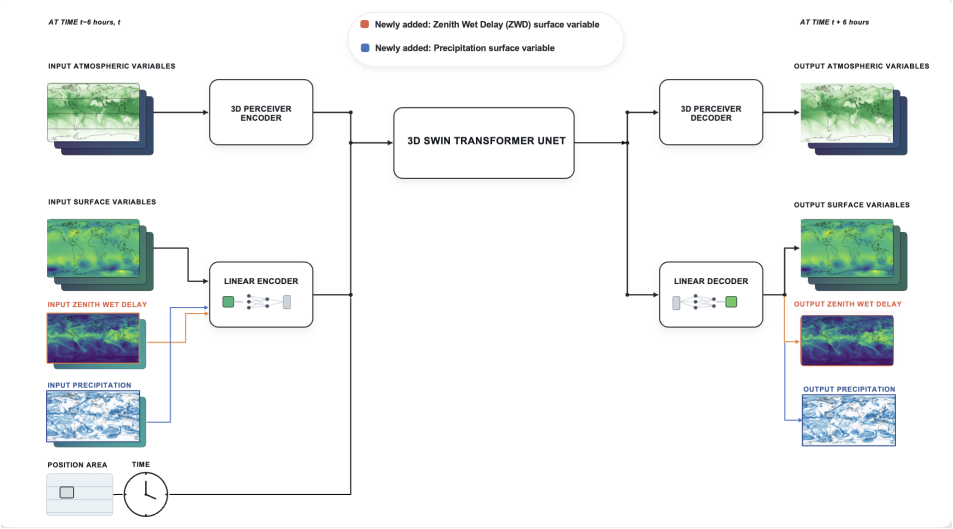
Text S2 Evaluation Metrics: Definitions

Model performance is assessed using several complementary metrics. The *Mean Absolute Error* (MAE) measures average pointwise prediction error in physical units (mm for ZWD and precipitation), and the *Root Mean Squared Error* (RMSE) and *Mean Squared Error* (MSE) further emphasise large errors. The *Pearson correlation coefficient* (R) quantifies linear agreement between predicted and target fields. The *Normalised MAE* expresses the physical MAE divided by the variable's standard deviation across the dataset, enabling comparison across variables with different physical scales, while *Relative MAE* is defined as the MAE divided by the mean absolute value of the observations; an analogous Normalized MSE is also reported.

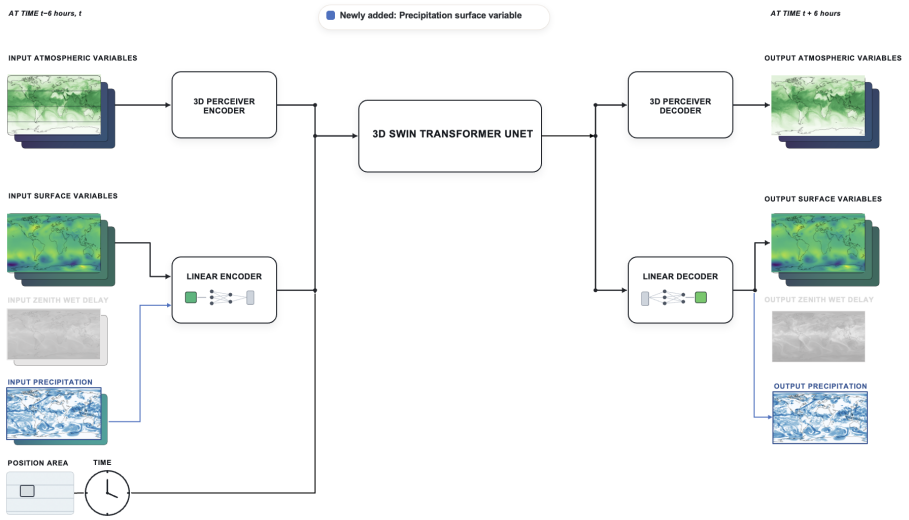
The *Fractions Skill Score* (FSS) (Roberts & Lean, 2008) is a neighbourhood-based metric that evaluates spatial structure by comparing the fraction of area exceeding a threshold between forecast and observation within windows of increasing size; we report FSS evaluated at the 95th percentile of the observed distribution, using a square neighbourhood window of side 9x9 grid points (corresponding to $\sim$2.25° at 0.25° resolution).. The *Equitable Threat*



(a) Step 1: ZWD added as a new surface variable.



(b) Step 2 Model A (With ZWD): both ZWD and total precipitation added; ZWD serves as an auxiliary input and auxiliary target.



(c) Step 2 Model B (Baseline, Without ZWD): total precipitation added as the only new variable.

Figure S2: Per-experiment model-architecture diagrams. The pretrained Aurora encoder, 3D Swin Transformer U-Net backbone and decoder are shared across all three configurations; newly added surface variables (highlighted) enter through an extension of the linear surface encoder and decoder only.

Score (ETS) is a categorical metric that measures the fraction of correctly forecast events above a given threshold while adjusting for hits expected by chance; ETS is reported at the 75th, 90th, 95th and 99th percentiles of the climatological precipitation distribution and is particularly informative for severe-event verification.

Energy spectra of precipitation are computed as the latitude-weighted (cosine of latitude) zonal power spectrum, obtained by applying a real-valued FFT along the longitude dimension and averaging $|\hat{f}(k)|^2$ across latitudes, to characterise the distribution of spatial variance across scales. Deviations of the predicted spectrum from the reference are summarised by the *Log Spectral Distance* (LSD), defined as the root-mean-square difference between the base-10 logarithms of the predicted and target power spectra,

$$\mathrm{LSD} = \sqrt{\overline{(\log_{10} S_{\mathrm{pred}} - \log_{10} S_{\mathrm{target}})^2}},$$

averaged over wavenumber within each of four bands: planetary (5,000-20,000 km), synoptic (1,000-5,000 km), upper mesoscale (250-1,000 km) and lower mesoscale (10-250 km).

When reported, the relative change is

$$\text{Rel. } \Delta = \frac{\text{With ZWD} - \text{Without ZWD}}{\text{Without ZWD}} \times 100\%,$$

computed from the unrounded checkpoint means.

Text S3 Training Details

All fine-tuning experiments use the same optimiser and warmup schedule as Aurora's pretraining (Bodnar et al., 2025) and optimise the per-variable weighted MAE objective defined in Eq. 1 of the main text.

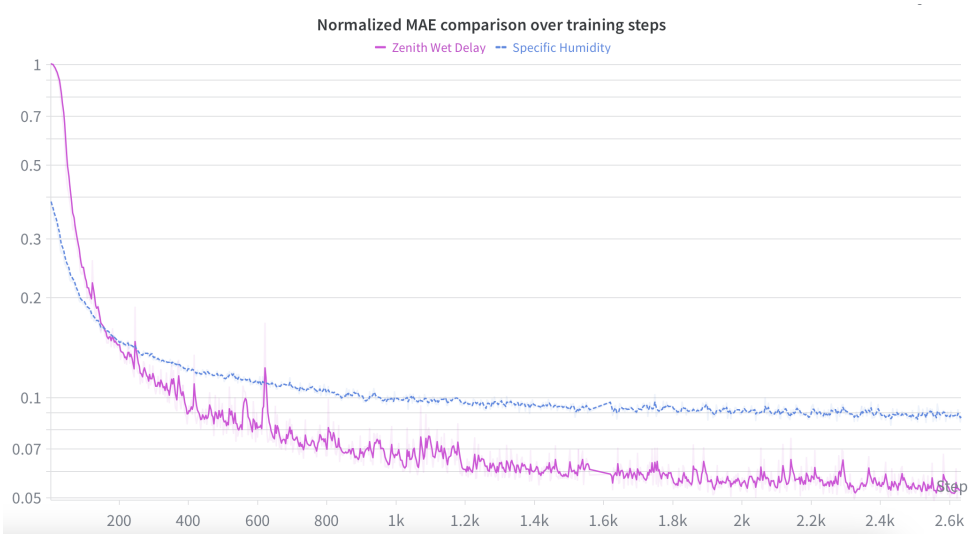
For the 1.3 B Aurora model, Step 1 (ZWD) uses a peak learning rate of 10^{-4}, 4,000 steps and 32 GH200 GPUs (8 nodes × 4 GPUs) on the Alps supercomputer at the Swiss National Supercomputing Centre (CSCS); Step 2 (precipitation, Models A and B) uses the same learning rate, ∼7,000 steps and 16 GPUs (4 nodes × 4 GPUs) for a 12-hour wall-time budget, with $\lambda_{\mathrm{ZWD}} = 2$ (sensitivity in Text S4) and precipitation weight 1.

For the 110 M small-model experiments (Text S7, Text S6), early-training instabilities required a smaller peak learning rate (5×10^{-5}) and substantially smaller new-variable loss weights ($\lambda_{\mathrm{ZWD}} = \lambda_{\mathrm{precip}} = 0.1$); each experiment runs for 6 wall-clock hours on 16 GPUs (4 nodes × 4 GPUs) for ∼5,000 steps.

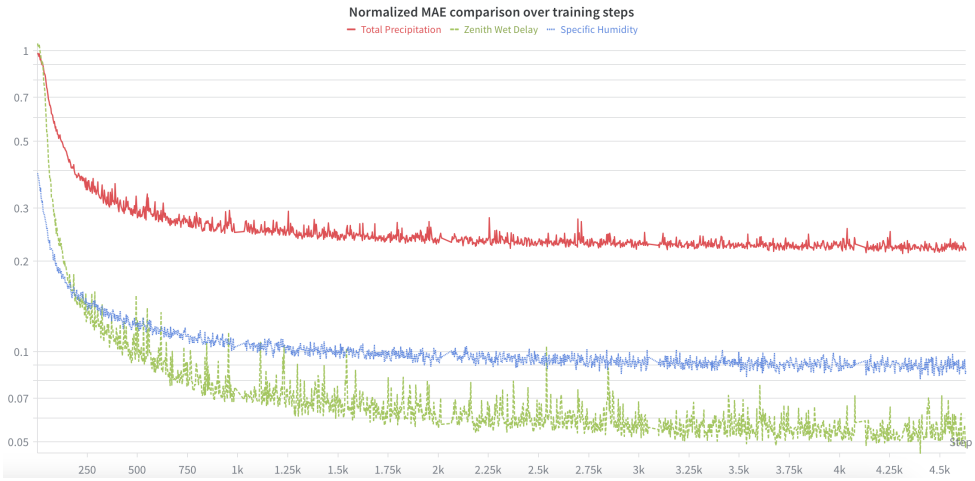
Figure S3 shows the evolution of the normalised MAE over training steps for the 1.3 B model. In Step 1 (Figure S3a), ZWD converges within the first few hundred steps to a normalised MAE comparable to that of specific humidity, the closest pretrained moisture analogue. In Step 2 (Figure S3b), total precipitation (a genuinely new task) converges to a substantially higher normalised MAE than the auxiliary ZWD and specific humidity targets, reflecting the intrinsic difficulty of precipitation, while ZWD and humidity remain at the low error levels established during Step 1.

Text S4 Loss Weight Sensitivity Analysis (1.3 B Model)

In the Step 2 precipitation fine-tuning experiment, Model A predicts both six-hour accumulated precipitation and ZWD, the latter acting as an auxiliary target whose contribution to the joint training objective is set by the loss weight λ_{ZWD}. We compare four values ($\lambda_{\mathrm{ZWD}} \in \{1, 2, 3, 10\}$) against the baseline Model B (no ZWD) on the 1.3 B Aurora; $\lambda_{\mathrm{ZWD}} = 2$ is the value adopted in the main analysis. All values in this section are evaluated at the single main-analysis checkpoint, whereas the headline metrics in the main text and in Text S8 are averaged over ten training checkpoints; the baseline and $\lambda_{\mathrm{ZWD}} = 2$ figures here therefore differ slightly from those ten-checkpoint means.



(a) Step 1: ZWD (magenta) and specific humidity (blue).



(b) Step 2 (Model A): total precipitation (red), ZWD (green) and specific humidity (blue).

Figure S3: Normalised MAE over training steps for the 1.3 B Aurora (log scale; lower is better). (a) Step 1 ZWD fine-tuning; (b) Step 2 precipitation fine-tuning with ZWD. The horizontal axis is truncated to the region where the curves vary; training continued to the full step counts given above, over which the normalised MAE remained essentially flat.

Table S1 summarises deterministic precipitation skill: $\lambda_{\mathrm{ZWD}} = 2$ achieves the best score on every metric except Pearson R (where the baseline leads marginally), and $\lambda_{\mathrm{ZWD}} = 10$ degrades to near-baseline or worse. Table S2 confirms the same picture for ETS at every threshold; at the 75th percentile $\lambda_{\mathrm{ZWD}} = 10$ drops below baseline. Spectral fidelity (Table S3) is best at $\lambda_{\mathrm{ZWD}} = 1$ and 2, with the differences between them under 1%. Finally, degradation of the pretrained variables (Table S4) is monotonic with the loss weight for the wind components and non-monotonic for t2m and MSLP; at $\lambda_{\mathrm{ZWD}} = 2$ the costs are moderate (+1.4% to +7.7% across the four variables). Across all of these criteria, $\lambda_{\mathrm{ZWD}} = 2$ is the optimal choice.

Table S1: Deterministic precipitation metrics for the 1.3 B Aurora across ZWD loss weights (lw).

Metric	No ZWD	lw=1	lw=2	lw=3	lw=10
MAE [mm]	0.192	0.190	**0.189**	0.190	0.192
RMSE [mm]	1.015	0.997	**0.991**	0.999	1.004
Pearson R	**0.853**	0.851	0.852	0.851	0.848
FSS 95%	0.915	0.929	**0.931**	0.929	0.925
Relative MAE	0.269	0.266	**0.265**	0.266	0.270
Normalised MSE	0.260	0.250	**0.248**	0.252	0.254

Table S2: ETS for precipitation at four climatological percentiles (pct) across ZWD loss weights (lw).

Threshold	No ZWD	lw=1	lw=2	lw=3	lw=10
75th pct	0.695	0.699	**0.703**	**0.703**	0.692
90th pct	0.654	0.666	**0.669**	0.668	0.657
95th pct	0.574	0.598	**0.600**	0.598	0.587
99th pct	0.386	0.420	**0.422**	0.415	0.414

Table S3: Log Spectral Distance (LSD) for precipitation across ZWD loss weights (lw).

Band	No ZWD	lw=1	lw=2	lw=3	lw=10
Planetary	0.223	0.155	**0.145**	0.165	0.156
Synoptic	0.259	0.187	**0.182**	0.203	0.189
Upper mesoscale	0.636	0.562	**0.558**	0.589	0.561
Lower mesoscale	1.076	**0.992**	1.001	1.048	1.013
Total	0.986	**0.906**	0.914	0.957	0.924

Text S5 Cost to Pretrained Variables (1.3 B Model)

The precipitation gain comes at a modest cost to the pretrained surface variables, quantified here as a mean ± sample standard deviation (s.d.) over the same ten training checkpoints used in Text S8 (Table S5). On the MAE metric used throughout the paper -

Table S4: Normalised MAE of the pretrained surface variables and of precipitation across ZWD loss weights (lw).

Variable	No ZWD	lw=1	lw=2	lw=3	lw=10
u10	**0.0363**	0.0366	0.0368	0.0372	0.0383
v10	**0.0436**	0.0440	0.0444	0.0444	0.0458
t2m	**0.0087**	0.0092	0.0090	0.0090	0.0094
MSLP	**0.0091**	0.0095	0.0098	0.0092	0.0098
Precip	0.0963	0.0952	**0.0950**	0.0953	0.0966

for which the relative change and paired p-value equal those of the normalised MAE, the two differing only by a fixed per-variable scale factor - only the two 10 m wind components degrade significantly, and by only ∼1%; 2 m temperature, mean sea-level pressure and specific humidity are all within checkpoint-to-checkpoint noise. In particular MSLP, whose degradation at the main-analysis checkpoint reaches +7.7% in normalised MAE (the largest single-checkpoint cost we observe), averages to only +1.8% over the ten checkpoints and is not statistically significant (paired $p = 0.248$), so that single-checkpoint figure sits at the upper end of the checkpoint spread rather than representing a systematic penalty. The more large-error-sensitive RMSE tells the same story but additionally flags 2 m temperature as a small (+3.9%) significant degradation. The loss-weight sweep (Text S4, Table S4) is consistent, the cost being non-monotonic in λ_{ZWD} for MSLP and t2m (falling to +1.1% for MSLP at $\lambda_{\mathrm{ZWD}} = 3$). The degradations are therefore small and largely noise-level, and they fall on variables that were already predicted with near-perfect accuracy (all retain Pearson $R \geq 0.997$ at the main-analysis checkpoint). In the small model this cost vanishes entirely (Text S6, Table S9): adding ZWD on top of precipitation introduces no degradation beyond that of precipitation alone and even improves t2m and MSLP.

Table S5: Cost of adding ZWD on the standard pretrained variables and specific humidity, as a mean ± s.d. over ten training checkpoints. Columns give the RMSE for each model, then the relative change with two-sided paired t-test p-value for both RMSE and MAE (the MAE relative change and p-value equal those of the normalised MAE). $\%\Delta > 0$ means With ZWD has the larger error; $^*\, p < 0.05$.

	RMSE (mean ± s.d.)		RMSE		MAE
Variable	Without ZWD	With ZWD	$\%\Delta$	paired p	$\%\Delta$ (p)
u10 [$\mathrm{m\,s^{-1}}$]	0.366 ± 0.001	0.371 ± 0.001	+1.3%	0.000*	+1.5% (0.000*)
v10 [$\mathrm{m\,s^{-1}}$]	0.377 ± 0.003	0.381 ± 0.001	+1.1%	0.001*	+1.2% (0.005*)
t2m [K]	0.382 ± 0.009	0.397 ± 0.011	+3.9%	0.022*	+4.8% (0.158)
MSLP [Pa]	21.49 ± 0.43	21.90 ± 0.56	+1.9%	0.118	+1.8% (0.248)
q [$10^{-4}\,\mathrm{kg\,kg^{-1}}$]	2.444 ± 0.022	2.449 ± 0.040	+0.2%	0.726	+0.2% (0.837)

Text S6 Small-Model Ablation (110 M): Surface-Only vs Without ZWD vs With ZWD

We repeat the main With-versus-Without-ZWD comparison on the small (110 M) Aurora variant, contrasting the "Without ZWD" model (precipitation, no ZWD) with the "With ZWD" model exactly as in the 1.3 B analysis. The only new element here is a third

configuration - a *Surface-only* model trained on the standard surface variables alone, with neither precipitation nor ZWD - which lets us separate the cost of introducing the precipitation task itself from the additional cost of adding ZWD on the pretrained variables, a distinction the two-model comparison cannot make on its own.

Precipitation Skill

Adding ZWD to the precipitation model improves every deterministic precipitation metric, including Pearson R, which (unlike in the 1.3 B case) does not decrease (Table S6). The FSS improvement is particularly large (+14.7%). The normalised MSE reduction (−6.5%) exceeds that of the 1.3 B model, a cross-scale contrast examined in Text S7. ETS (Table S7) shows the same threshold-dependent pattern as in the large model, but with substantially larger relative gains at every threshold: at the 99th percentile ETS improves by +50%, the largest relative gain at any threshold. The spectral improvements (Table S8) are large: the "Without ZWD" model produces fields with substantial spectral distortion (LSD > 0.45) at planetary and synoptic scales, and adding ZWD reduces this by roughly 60% at those scales, with a total LSD reduction of 38.6%.

Table S6: Deterministic precipitation metrics for the 110 M Aurora: Without ZWD vs With ZWD.

Metric	Without ZWD	With ZWD	Relative change
MAE [mm]	0.265	**0.247**	−6.5%
RMSE [mm]	1.329	**1.285**	−3.3%
MSE [mm^2]	1.767	**1.652**	−6.5%
Pearson R	0.719	**0.728**	+1.3%
FSS 95%	0.735	**0.843**	+14.7%
Relative MAE	0.370	**0.346**	−6.5%
Normalised MAE	0.133	**0.124**	−6.5%
Normalised MSE	0.445	**0.417**	−6.5%

Table S7: ETS for precipitation at four climatological percentiles (pct) for the 110 M Aurora.

Threshold	Without ZWD	With ZWD	Relative change
75th pct	0.527	**0.587**	+11.4%
90th pct	0.453	**0.533**	+17.7%
95th pct	0.354	**0.453**	+27.8%
99th pct	0.203	**0.304**	+50.1%

Figure S4 shows ETS at the 99th percentile over the test period for the two 110 M precipitation models. The "With ZWD" model outperforms the "Without ZWD" model at essentially every timestep, mirroring the behaviour of the 1.3 B model (Figure 2b of the main text) but with a substantially larger margin.

Cost to Pretrained Variables

Adding precipitation and ZWD does not materially degrade the skill of the standard pretrained surface variables (Table S9). Wind components remain at the Surface-only skill

Table S8: Log Spectral Distance (LSD) for precipitation in the 110 M Aurora. *Total* is LSD integrated over the full wavenumber range, not the sum of band values.

Band	Precip only	Precip + ZWD	Relative change
Planetary	0.457	**0.185**	−59.5%
Synoptic	0.465	**0.177**	−61.9%
Upper mesoscale	0.752	**0.418**	−44.4%
Lower mesoscale	0.606	**0.386**	−36.4%
Total	**0.631**	**0.388**	**−38.6%**

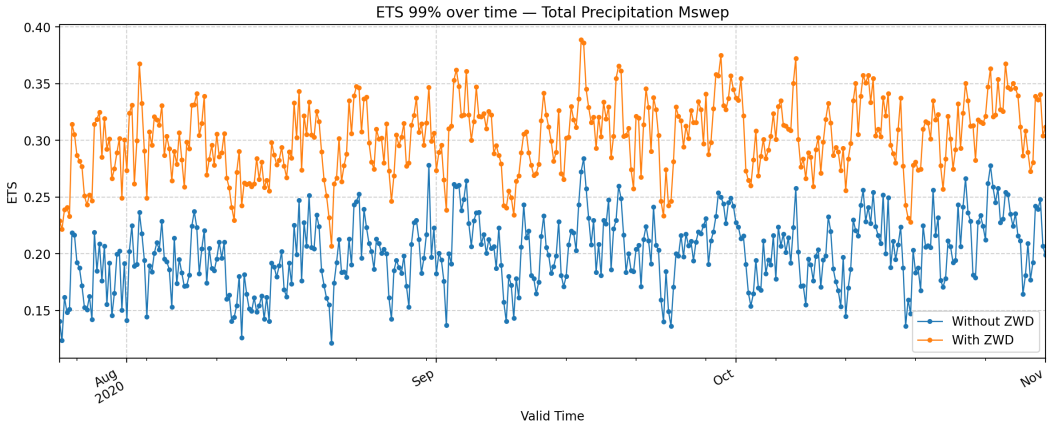


Figure S4: ETS at the 99th percentile over the test period for the 110 M Aurora, Without ZWD versus With ZWD.

level or slightly better; t2m shows a small increase in normalised MAE relative to Surface-only (+5.7%), substantially better than in the "Without ZWD" model; MSLP is comparable across all three configurations. Adding ZWD on top of precipitation therefore does not introduce meaningful additional degradation beyond that already caused by precipitation alone.

Table S9: Normalised MAE of the pretrained surface variables across the three 110 M configurations.

Variable	Surface-only	Without ZWD	With ZWD
u10	0.0488	0.0466	**0.0464**
v10	0.0576	0.0567	**0.0559**
t2m	**0.0122**	0.0161	0.0129
MSLP	0.0140	0.0145	**0.0124**

Text S7 Cross-Scale Comparison (110 M vs 1.3 B Parameters)

We place the small (110 M) and large (1.3 B) Aurora variants side by side. The small-model entries are the "Without ZWD" and "With ZWD" configurations from the ablation in Text S6; the large-model entries are Models B and A from the main analysis, evaluated

here at the single main-analysis checkpoint for consistency with the small-model figures. The large-model values therefore differ slightly from the ten-checkpoint means reported in the main text (e.g. the ETS_{99} gain is +9.3% at this checkpoint versus +8.8% averaged over the ten checkpoints); the cross-scale ranking is unaffected. Table S10 reports deterministic precipitation skill, and Table S11 the ETS at the 99th percentile.

Three patterns are evident. First, model scale dominates: the large baseline outperforms the small model with ZWD on every metric. Second, ZWD's *relative* benefit is consistently larger at the small scale, across normalised MSE, FSS and threshold skill alike (Tables S10-S11; small-model magnitudes in Text S6). Third, this scale contrast is starkest at the 99th-percentile ETS, where the gain reaches +50% in the small model against +9.3% in the large one.

Table S10: Headline deterministic precipitation metrics across model scales, with and without ZWD. The full metric set for the small model is given in Table S6.

Metric	Small (no ZWD)	Small (+ ZWD)	Large (no ZWD)	Large (+ ZWD)
MAE [mm]	0.265	0.247	0.192	**0.189**
FSS 95%	0.735	0.843	0.915	**0.931**
Normalised MSE	0.445	0.417	0.260	**0.248**

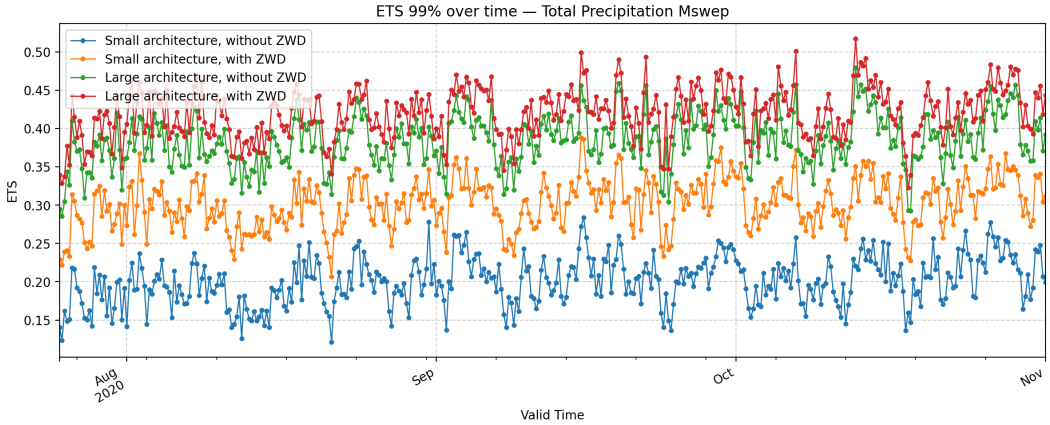
Table S11: Equitable Threat Score (ETS) at the 99th percentile, across model scales and ZWD configurations.

Configuration	ETS 99%	Relative change
Small (Without ZWD)	0.203	—
Small (With ZWD)	0.304	+50.1%
Large (Without ZWD)	0.386	—
Large (With ZWD)	0.422	+9.3%

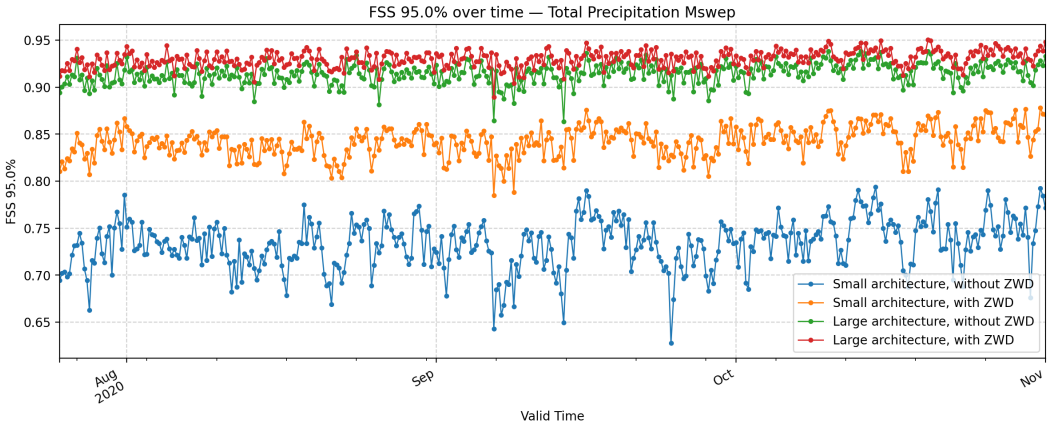
The per-timestep view in Figure S5 confirms that these conclusions hold throughout the test period rather than on the test-set average alone. At essentially every verification time the four configurations remain cleanly ordered (large (With ZWD) $\geq$ large (Without ZWD) $>$ small (With ZWD) $>$ small (Without ZWD)) for both ETS at the 99th percentile (Figure S5a) and FSS at the 95th percentile (Figure S5b). The ZWD-induced gain is visibly largest for the small model, consistent with the inverse capacity scaling reported above.

Text S8 Checkpoint Variability of the ZWD Effect (1.3 B Model, $\lambda_{ZWD} = 2$)

The deterministic and threshold-based gains in the main text (Section 3.2.1) are reported as means over ten training checkpoints, all at $\lambda_{ZWD} = 2$ and on the late-training plateau (the checkpoint used for the main-analysis maps and single-checkpoint diagnostics is one of the ten). This section documents that ten-checkpoint ensemble in full: the per-model spread, the paired significance of the With-versus-Without-ZWD difference, and the accompanying cost to the standard variables. For every metric we report the mean $\pm$ s.d. (computed with ddof= 1) across the ten checkpoints, the s.d. quantifying the intrinsic checkpoint-to-checkpoint variability of each model along its training trajectory. This analysis complements the loss-weight sweep of Text S4, which instead varies λ_{ZWD} at a fixed



(a) ETS at the 99th percentile.



(b) FSS at the 95th percentile.

Figure S5: Precipitation skill over the test period for the four model configurations (small/large × with/without ZWD). The configuration ordering is preserved at essentially every timestep.

checkpoint: here λ_{ZWD} is held fixed and the checkpoint varies, so the two together bound the sensitivity of the result to both the hyperparameter and the training trajectory.

Sampling ten checkpoints from a single training run per model is a deliberate compromise between statistical rigour and computational cost. The ideal design for separating the ZWD effect from training stochasticity would repeat each Step 2 training from several independent random seeds, yielding genuinely independent replicates; at the scale of the 1.3 B model, however, ten full retrainings per configuration is prohibitively expensive. Sampling ten checkpoints along the late-training plateau of one run per model is a far cheaper surrogate that still probes the stochastic-optimisation variability of the trajectory, and it is on this within-trajectory variability - rather than on true run-to-run variability - that the robustness assessment of this section is based.

The two models do not share weights - Model A and Model B are distinct networks - but they are trained under an identical protocol (same pretrained initialisation, optimiser, learning-rate schedule, step budget and training data; Figure S1), differing only in the presence of ZWD. A checkpoint is saved from each run at the *same* ten training steps on the late-training plateau, so checkpoint i of Model A and checkpoint i of Model B are matched by training step: they occupy the same position along an otherwise identical trajectory and differ only in ZWD. This step-matched correspondence is what defines a pair, and it is why the with-versus-without-ZWD contrast is naturally paired: differencing the two models at each matched step cancels the variation common to both runs at that step - schedule position, learning rate, cumulative data seen and the stochastic-optimisation phase - and isolates the ZWD effect. We therefore report how many of the ten checkpoints favour the ZWD model and the two-sided paired t-test p-value across checkpoints; this is the appropriate test for a per-checkpoint difference and is more sensitive than comparing the two means against their (larger) marginal spread. A paired t-test moreover remains valid under any pairing, so the residual divergence of the two trajectories once ZWD is added can only reduce its power, never inflate its false-positive rate.

Table S12 reports the precipitation deterministic skill of each model as a mean ± s.d., and Table S13 the paired head-to-head comparison for RMSE; the corresponding ETS paired results (number of checkpoints favouring ZWD and paired p-value at each threshold) are given in the main text (Figure 2a) and are not repeated here. The precipitation improvement is reproduced at almost every checkpoint (8-9 of the ten for every metric) and is statistically significant under the paired test both for RMSE (9/10 checkpoints favour ZWD, $p = 0.032$) and at every ETS threshold ($p \leq 0.013$; main text). The per-model checkpoint scatter is itself sizeable (e.g. ± 0.05 on ETS_{99}), comparable to the mean With-minus-Without gap, so the two models' ±s.d. ranges overlap; the gain is nonetheless robust precisely because it is systematic within each checkpoint, which is what the paired test captures. The With-ZWD model is moreover the more stable of the two, with a smaller s.d. on every precipitation metric.

Table S12: Precipitation deterministic skill as mean ± s.d. over ten training checkpoints (1.3 B model, $\lambda_{\mathrm{ZWD}} = 2$).

Metric	Without ZWD (Model B)	With ZWD (Model A)
MAE [mm]	0.195 ± 0.007	$\mathbf{0.193 \pm 0.004}$
RMSE [mm]	1.015 ± 0.034	$\mathbf{0.997 \pm 0.019}$
MSE [mm^2]	1.032 ± 0.071	$\mathbf{0.994 \pm 0.038}$

The accompanying cost of ZWD on the standard pretrained variables over the same ten checkpoints is reported in Text S5 (Table S5).

Table S13: Paired checkpoint comparison of precipitation RMSE (1.3 B model, $\lambda_{\mathrm{ZWD}} = 2$): mean ± s.d. for each model, relative change, number of checkpoints favouring ZWD, and two-sided paired t-test p-value (lower RMSE is better). The corresponding ETS paired results are reported in the main text (Figure 2a). $^{*}\, p < 0.05$.

Metric	Without ZWD	With ZWD	Rel. Δ	ZWD wins	paired p
RMSE [mm]	1.015 ± 0.034	$\mathbf{0.997 \pm 0.019}$	−1.8%	9/10	0.032^{*}

Across the ten checkpoints the ZWD model improves precipitation consistently (by up to +8.8% in ETS at the 99th percentile, significant under the paired test), at a largely noise-level cost to the standard variables (Text S5).

The spectral improvement is checkpoint-robust in the same way. The band-wise Log Spectral Distance, reported checkpoint-averaged in the main text (Figure 3b of the main text), decreases with ZWD at 8-9 of the ten checkpoints in every band, with a two-sided paired-test significance of $p \leq 0.023$ (planetary and synoptic $p \leq 0.001$). As for ETS, the per-checkpoint spread of the LSD is large relative to the mean With-minus-Without gap (especially at planetary and synoptic scales, where the baseline LSD is itself small), so the improvement is significant through its consistency across checkpoints rather than by separation of the marginal ±s.d. ranges.

Text S9 Regridding of the MSWEP Precipitation Target

The MSWEP V2 precipitation product is distributed at 0.1° resolution and is regridded to Aurora's 0.25° grid (Section 2.3 of the main text) by bilinear interpolation, so that each 0.25° target cell aggregates approximately $(0.25/0.1)^2 \approx 6$ native cells. We use the same regridded MSWEP product as Lehmann et al. (2025), accumulated here to six-hour totals (the sum of two consecutive three-hour fields). Because precipitation is a mass-conserved quantity, bilinear interpolation does not exactly preserve the domain water budget: conservative (area-weighted) remapping is the mass-preserving alternative, and it additionally tends to attenuate the extreme upper tail of the distribution because the value assigned to each coarse cell is an area average over its native cells rather than a point sample.

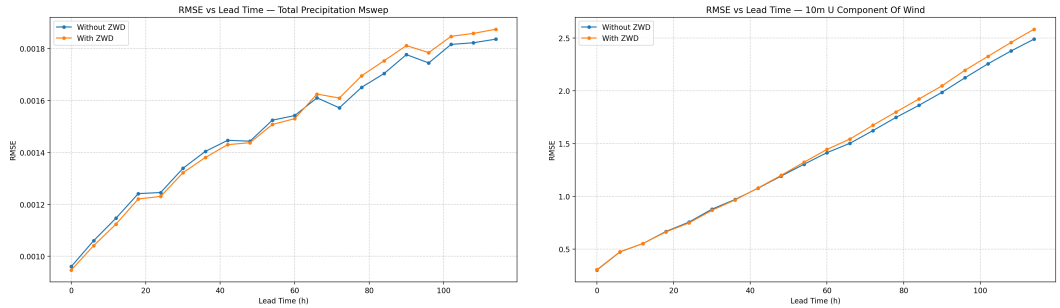
We retain bilinear interpolation. Because Model A (with ZWD) and Model B (baseline) are trained and evaluated against the *same* bilinearly regridded MSWEP target (Section 2.4), any departure of that target from the "true" 0.25° precipitation is common to both models and cancels in the with-versus-without-ZWD contrast that all reported gains measure; the regridding acts identically on both models' targets. We have not retrained under a conservative scheme, so the absolute scores and the precise magnitude of the ZWD gains may differ under a different regridding, whereas the with-versus-without-ZWD comparison does not.

A complementary direction for future work is to sidestep the coarsening altogether by operating at MSWEP's native 0.1° resolution: using the 0.1° precipitation field directly as the target and regridding the ZWDX product to that same grid, so that ZWD informs precipitation at its native scale. Unlike the fixed 0.25° Aurora used here, this would require a foundation model able to ingest and forecast fields at heterogeneous resolutions, such as the Earth System Foundation Model (Ozdemir et al., 2026).

Text S10 Rollout skill as a function of Lead Time

All metrics in the main text and in the preceding sections are reported at a single 6-hour lead time. To characterise how the two precipitation models behave when integrated forward, we evaluate them over autoregressive rollouts out to a lead time of 114 h (≈4.75 days) at 6-hourly steps. Neither model received any rollout (multi-step) fine-tuning: the forecasts are produced purely autoregressively by feeding each 6-hour prediction back as the input for the next step, so the curves below reflect the intrinsic rollout stability of the 6-hour models. Scores are averaged over a set of 10 initialisation dates spanning the test set.

We first verify that adding the ZWD objective does not compromise rollout stability, for precipitation or for the pretrained variables. Figure S6 shows deterministic RMSE over the rollout for precipitation and for the 10 m zonal wind. For both variables RMSE grows steadily and smoothly with lead time for both models, with no divergence. Consistent with the 6-hour analysis, the two models differ only negligibly in bulk RMSE: the With-ZWD model is marginally better than the Baseline at the shortest precipitation leads and marginally worse at the longest, with the difference between the curves remaining negligible throughout.



(a) Total precipitation.

(b) 10 m zonal wind.

Figure S6: RMSE as a function of lead time, Baseline (without ZWD) versus With ZWD, for (a) total precipitation and (b) the 10 m zonal wind. Error grows smoothly over the rollout for both models, and the bulk-RMSE difference between them stays negligible throughout - confirming that adding ZWD does not degrade deterministic skill.

This well-behaved decay extends to threshold-based skill. Figure S7 shows ETS as a function of lead time at all four climatological percentiles, for both the With-ZWD model (Model A, $\lambda_{\mathrm{ZWD}} = 2$; solid) and the Baseline (Model B; dotted). For both models the score decreases smoothly and monotonically with no sign of instability or error blow-up over the rollout. As expected, skill is lower in absolute terms at the more extreme thresholds (the 99th-percentile curve falls from ≈0.43 at the analysis time to ≈0.05 at 114 h, versus ≈0.70 to ≈0.26 at the 75th percentile), but the decay behaves well at all thresholds.

The same figure also shows that the ZWD advantage seen at the 6-hour lead is not confined to that first step. The ordering established at the 6-hour lead in the main text is preserved throughout the early and intermediate part of the rollout: at the most extreme thresholds (95th and 99th percentiles) Model A clearly outperforms the Baseline out to ≈60-70 h, with the largest separation at the shortest leads; at the lower thresholds (75th and 90th) the two are close throughout, with Model A matching the Baseline and never falling below it. Beyond ≈70 h all curves converge as both models lose skill, occasionally crossing within the noise. The ZWD advantage is therefore concentrated at the early-to-intermediate lead times.

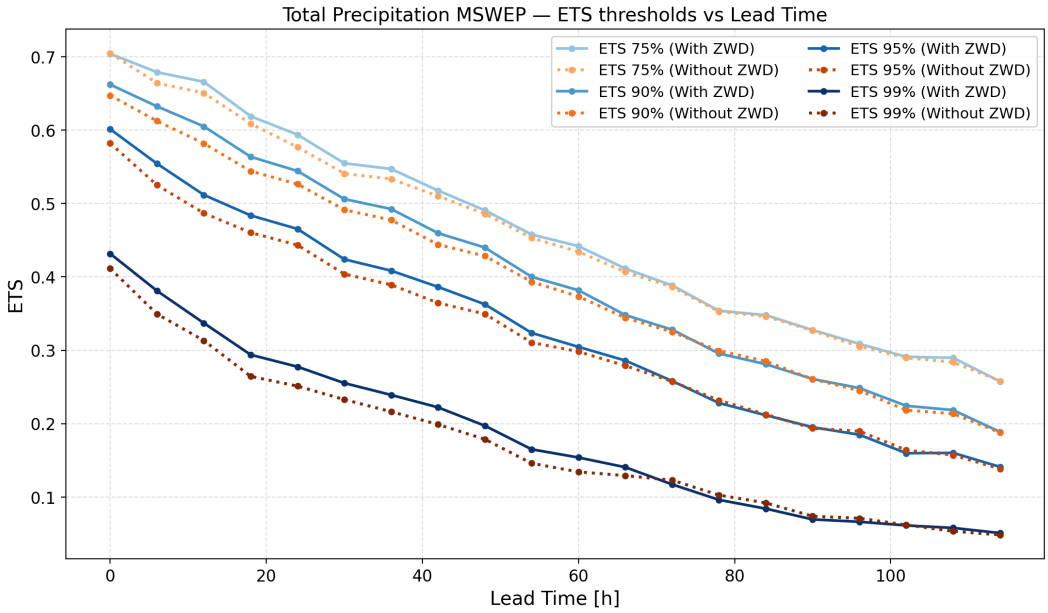


Figure S7: ETS as a function of lead time at four climatological percentiles (75th, 90th, 95th, 99th), for the With-ZWD 1.3 B model (Model A, $\lambda_{\mathrm{ZWD}} = 2$; solid) and the Baseline (Model B; dotted). For both models skill decays smoothly and monotonically over the autoregressive rollout at every threshold; the With-ZWD model outperforms the Baseline at the higher thresholds through the short-to-intermediate rollout, the two converging at long lead times.

Both models remain stable over the full 114 h horizon, and the ZWD advantage persists through the short-to-intermediate range before the two models converge at long lead times.

Text S11 Temporal Consistency and Learning-Quality Examples

This section complements the aggregate metrics with temporally resolved skill and single-case spatial illustrations for the 1.3 B model.

Figure S8 shows FSS at the 95th percentile over the full test period for Models A and B. The ZWD model lies above the Baseline at nearly every verification time, confirming that the aggregate FSS improvement reported in Section 3.2.1 reflects a temporally consistent gain rather than a few favourable dates.

Figure S9 and S10 illustrate the spatial quality of the Step 1 prediction for a representative initialisation. ZWD (Figure S9) is a variable outside Aurora's original vocabulary, yet the prediction reproduces its observed large-scale spatial structure and regional moisture gradients, rather than merely matching it in the bulk metrics of Table 1 of the main text. Specific humidity (Figure S10), an already-pretrained moisture field, is shown alongside as a reference: the newly learned ZWD attains a visual fidelity comparable to this established variable - the spatial-domain counterpart to the near-identical convergence of their training curves (Text S3, Figure S3a). Qualitative precipitation forecasts comparing target, Baseline and With-ZWD models are presented separately for three individual high-impact events in Text S12.

The spectral skill of the Step 1 ZWD prediction is reported in Table S14. The integrated LSD across all wavenumber bands is well within the envelope of the pretrained variables and

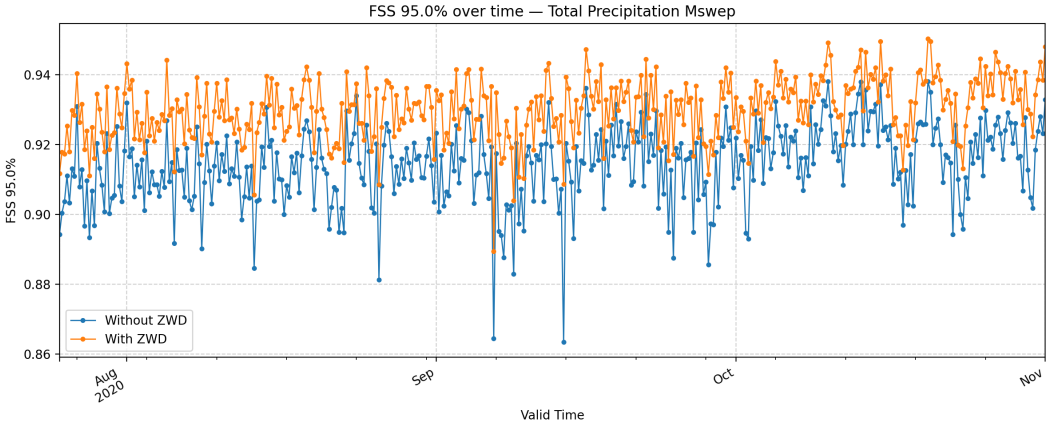


Figure S8: FSS at the 95th percentile over the test period for the 1.3 B Aurora, Baseline (without ZWD) versus With ZWD.

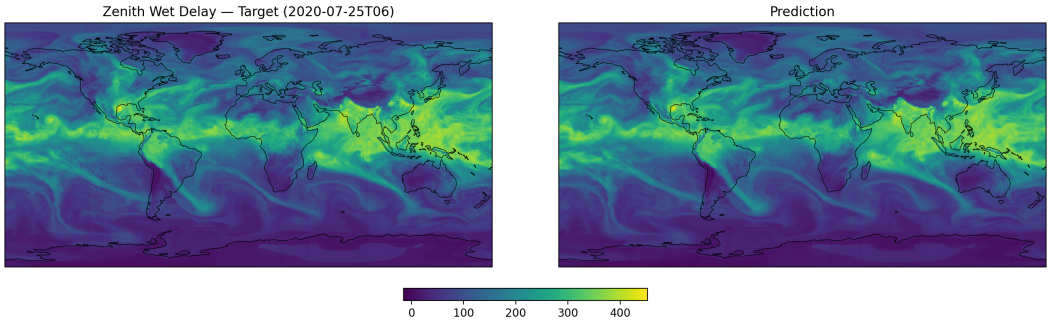


Figure S9: Step 1 ZWD target (left) and Aurora prediction (right) when initialized on 2020-07-25T06.

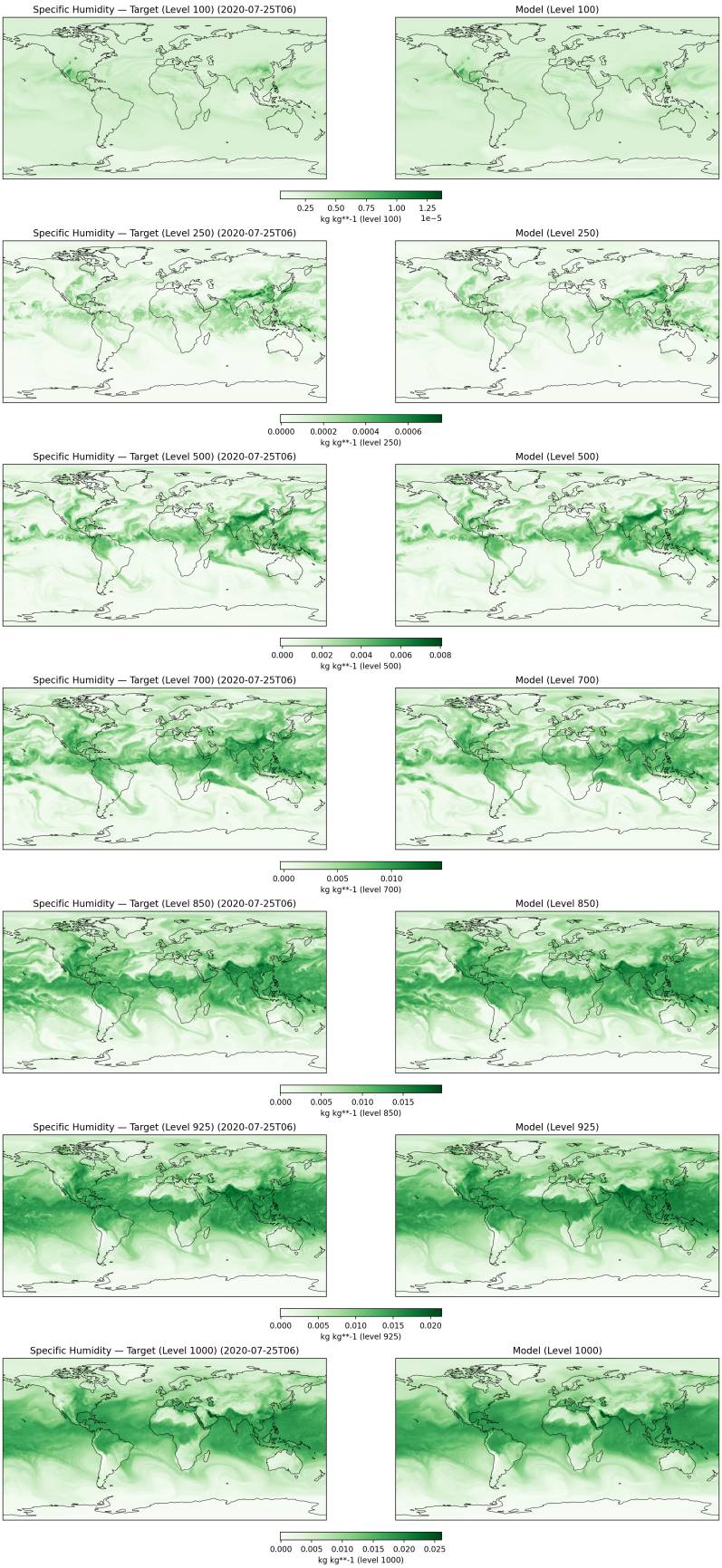


Figure S10: Specific humidity target (left) and Aurora prediction (right) across pressure levels when initialized on 2020-07-25T06.

better than that of u10, with the closest spectral match at planetary and synoptic scales. This complements the deterministic and threshold-based skill given in Table 1 of the main text.

Band	Scale range (km)	LSD (ZWD)	LSD range across pretrained variables
Planetary	5,000 - 20,000	0.006	0.005 (MSLP) - 0.008 (t2m)
Synoptic	1,000 - 5,000	0.022	0.015 (v10) - 0.021 (u10)
Upper mesoscale	250 - 1,000	0.125	0.045 (MSLP) - 0.130 (u10)
Lower mesoscale	10 - 250	0.279	0.126 (MSLP) - 0.379 (u10)
Total	10 - 20,000	0.252	0.113 (MSLP) - 0.339 (u10)

Table S14: Step 1 Log Spectral Distance (LSD) of ZWD across four wavenumber bands, compared to the range spanned by Aurora's standard pretrained variables; lower values indicate a more faithful power spectrum. The *Total* row is the LSD integrated over the full wavenumber range, not the sum of the band-wise values.

Text S12 Regional Extreme-Event Case Studies

The aggregate metrics in the main text and the preceding sections are computed globally over the held-out test period. To address whether the ZWD-induced improvement also manifests during individual high-impact precipitation events (and to provide the region-aggregated time series around heavy-rain episodes that complement the global diagnostics) we examine three extreme events drawn from distinct basins and meteorological regimes, all falling within the test set:

- **South Asia monsoon** (initialised 2020-08-04 12Z): an active-monsoon episode producing extreme accumulations over the Indian subcontinent and the Bay of Bengal.
- **East Asia Typhoon Bavi** (initialised 2020-08-24 06Z): a mature typhoon tracking towards the Korean peninsula.
- **Central America Hurricane Delta** (initialised 2020-10-04 06Z): a rapidly intensifying hurricane over the north-western Caribbean.

For each event we evaluate the 1.3 B Models A ("With ZWD") and B ("Without ZWD") over a regional domain enclosing the system. We report: the six-hour accumulated precipitation forecasts against the MSWEP target; the regional spatial RMSE and percentile-threshold ETS as time series spanning ±90 h around the event; the spatial *skill difference* for extreme cells only (the absolute "Without ZWD" error minus the absolute "With ZWD" error, so that green indicates the "With ZWD" model is closer to the observation); the threshold-exceedance hit/miss/false-alarm classification; and the *skill gain by intensity bin*, which stratifies the mean error reduction by the observed-precipitation percentile of each grid cell. The skill-by-intensity diagnostic shows the same signature across all three events: ZWD is neutral or marginally negative at low-to-moderate intensities and strongly positive in the extreme upper tail.

South Asia Monsoon (2020-08-04)

This is the clearest of the three cases. The regional RMSE time series (Figure S12a) shows the "With ZWD" model below the baseline "Without ZWD" at nearly every six-hourly step across the full ±90 h window, with the largest separation at the high-RMSE peaks; the threshold ETS series (Figure S12b) shows a consistent "With ZWD" advantage

that grows with threshold, most pronounced at the ≥10 and ≥20 mm levels. The skill-by-intensity decomposition (Figure 2c of the main text) rises monotonically with intensity to +1.52 mm of mean error reduction in the top percentile bin, while the spatial skill-difference map (Figure S13a) is overwhelmingly green over the regions of heaviest rainfall, and the exceedance panels (Figure S13b) show fewer missed events for the With-ZWD model at the higher thresholds.

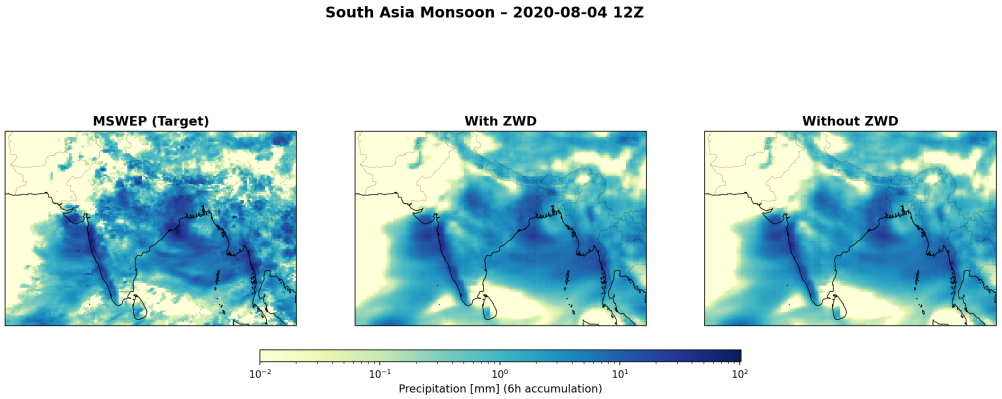


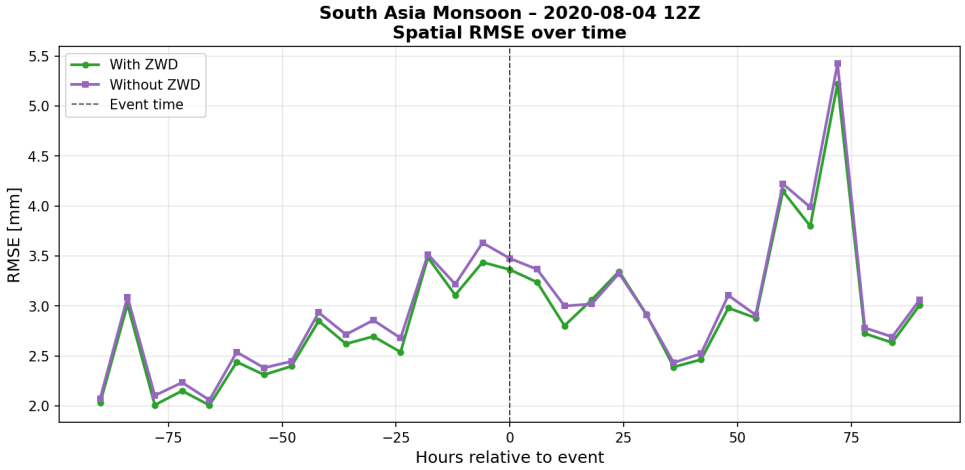
Figure S11: South Asia monsoon (2020-08-04 12Z): six-hour accumulated precipitation for the MSWEP target (left), model "With ZWD" (centre) and model "Without ZWD" (right).

East Asia Typhoon Bavi (2020-08-24)

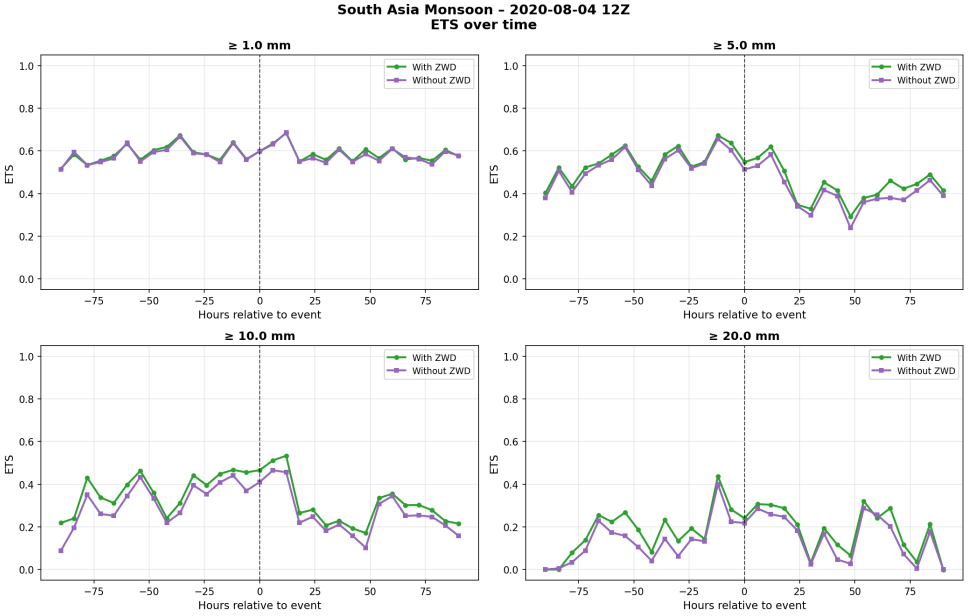
The typhoon case confirms the same picture with a stronger extreme-tail signal. The skill-by-intensity decomposition (Figure 2d of the main text) reaches +2.17 mm of mean error reduction in the top percentile bin (the largest of the three events) and the spatial skill-difference map (Figure S16a) shows a coherent green band of improvement along the typhoon's heavy-rain swath. The regional RMSE series (Figure S15a) is noisier than the monsoon case, with the two models crossing at several steps, but the "With ZWD" model holds a clear advantage through the pre-event intensification phase and at the RMSE peak around the event time; the ETS series (Figure S15b) again favours the With-ZWD model most at the heavier thresholds.

Central America Hurricane Delta (2020-10-04)

The hurricane case is included as a more neutral counterexample. Here the region-aggregated RMSE series (Figure S17b) shows the two models essentially overlapping, and the spatial skill-difference map (not shown) is mixed rather than predominantly green (adding ZWD neither helps nor harms the bulk forecast). Even so, the skill-by-intensity decomposition (Figure 2e of the main text) preserves the same qualitative signature as the other two events, with a clear positive gain (+0.71 mm) confined to the most extreme percentile bin. In all three cases the skill-by-intensity gain is thus confined to the most extreme bin (Figure 2c-e of the main text), and the hurricane shows this tail gain with essentially unchanged bulk RMSE.

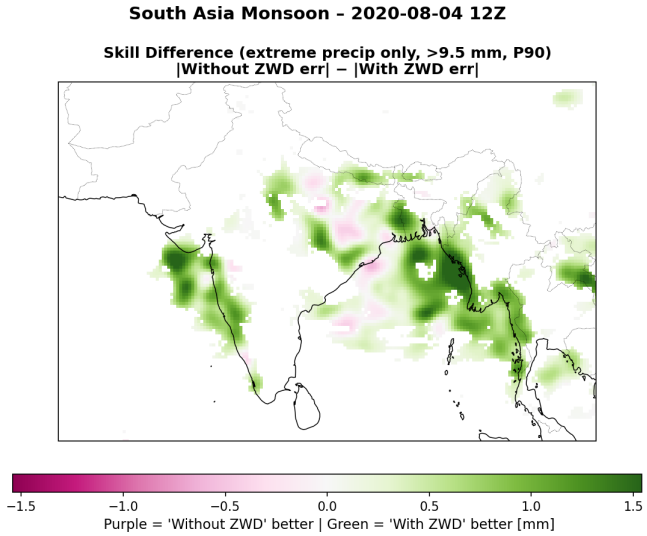


(a) Regional spatial RMSE.

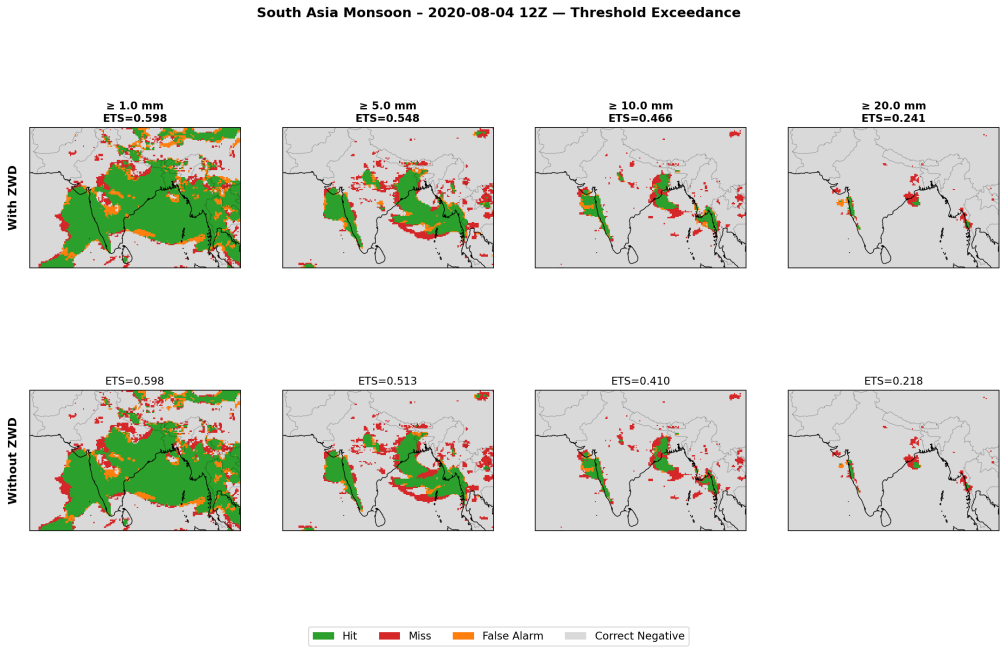


(b) Regional ETS at four climatological thresholds.

Figure S12: South Asia monsoon: region-aggregated skill as a function of time relative to the event ($t = 0$, dashed line), "With ZWD" versus "Without ZWD". The "With ZWD" model is consistently more skillful, most clearly at the heavier thresholds.



(a) Skill difference, extreme cells only.



(b) Threshold-exceedance classification (top: "With ZWD"; bottom: "Without ZWD").

Figure S13: South Asia monsoon: spatial verification. (a) Per-cell skill difference for extreme precipitation (>P90); green indicates the "With ZWD" forecast is closer to MSWEP. (b) Hit/miss/false-alarm/correct-negative classification at four absolute thresholds.

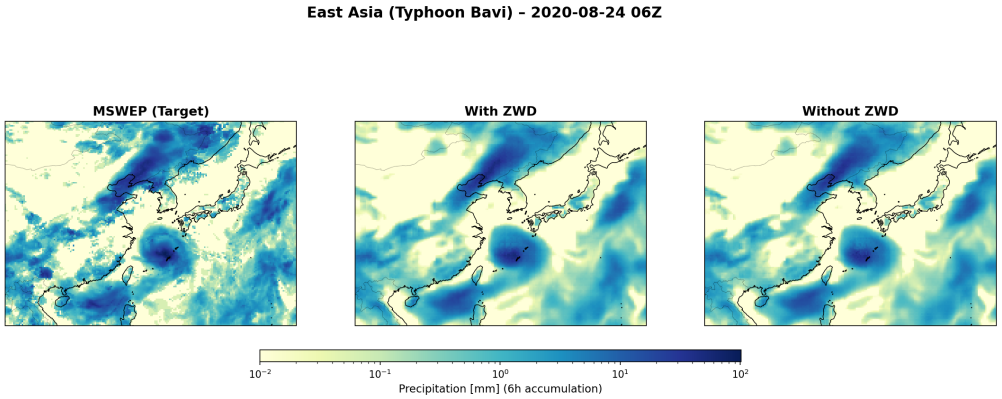
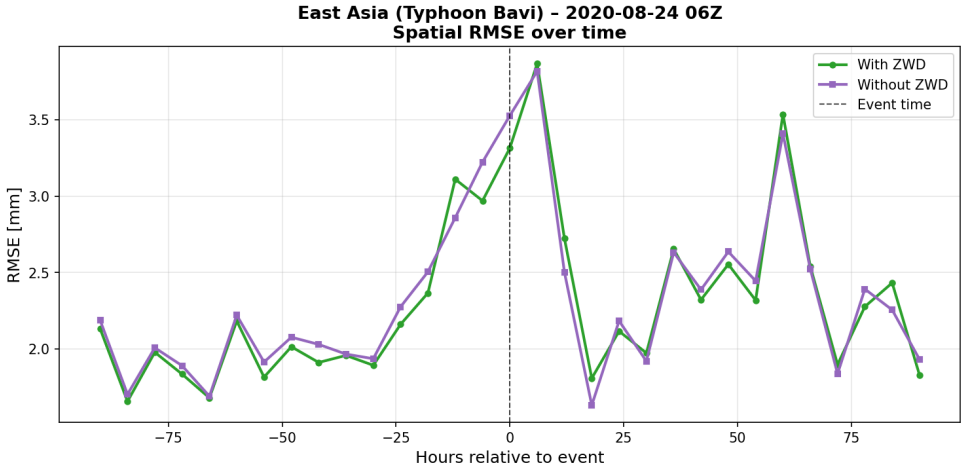
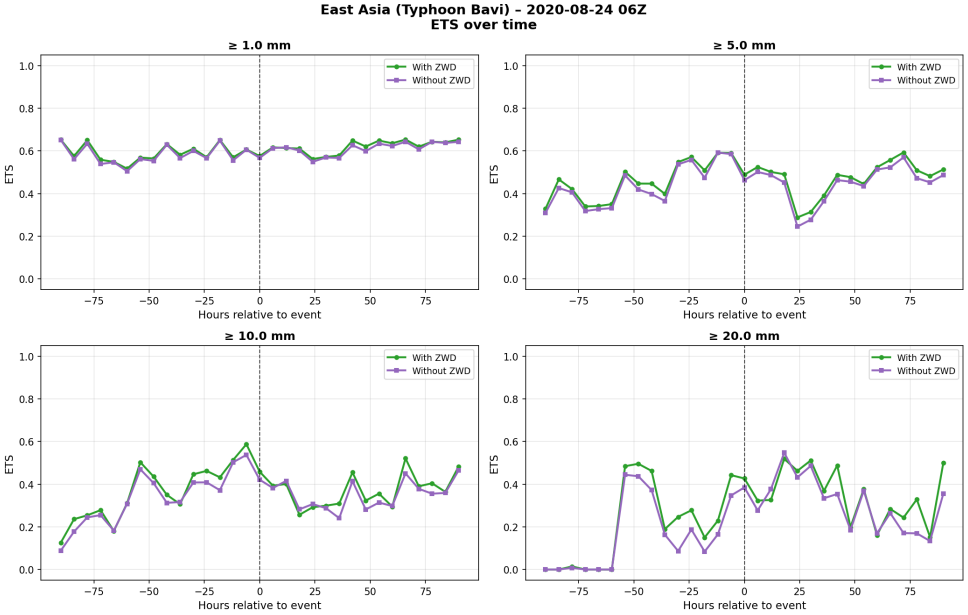


Figure S14: East Asia Typhoon Bavi (2020-08-24 06Z): six-hour accumulated precipitation for the MSWEP target (left), model "With ZWD" (centre) and model "Without ZWD" (right).

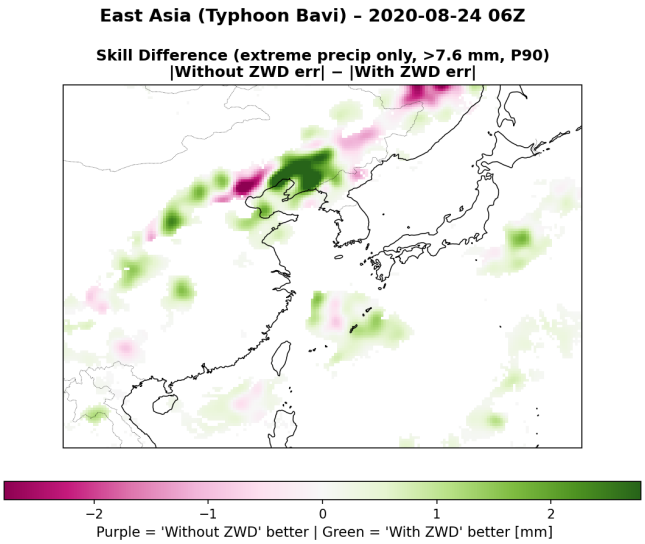


(a) Regional spatial RMSE.

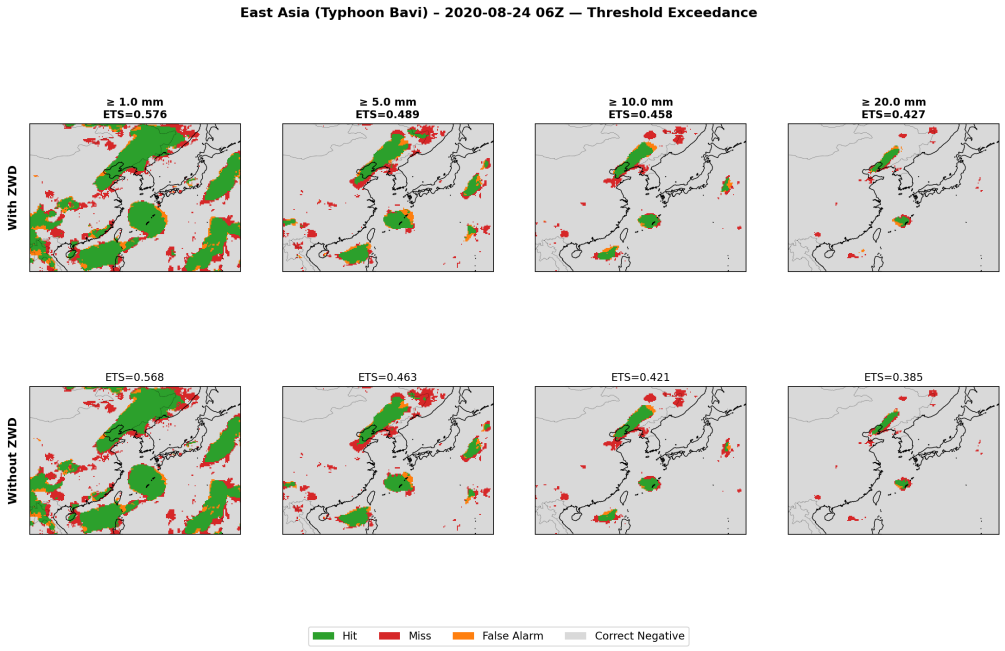


(b) Regional ETS at four climatological thresholds.

Figure S15: East Asia Typhoon Bavi: region-aggregated skill as a function of time relative to the event ($t = 0$, dashed line), "With ZWD" versus "Without ZWD".

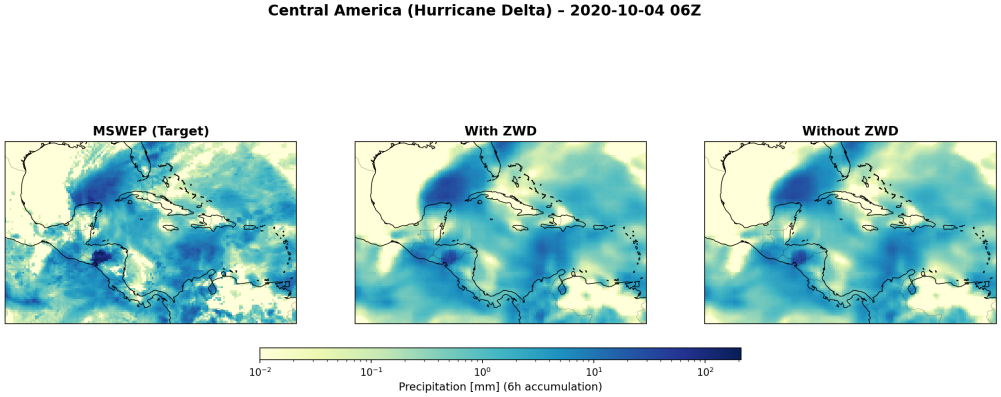


(a) Skill difference, extreme cells only.

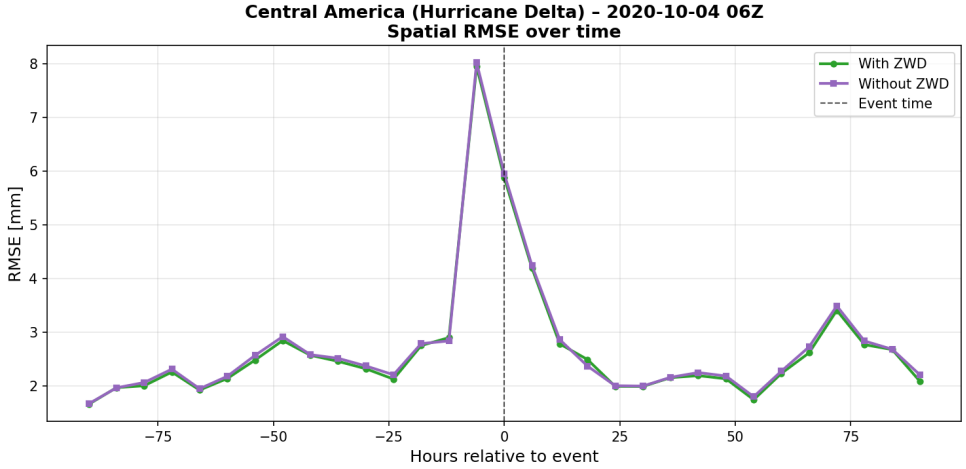


(b) Threshold-exceedance classification (top: With ZWD; bottom: Baseline).

Figure S16: East Asia Typhoon Bavi: spatial verification. (a) Per-cell skill difference for extreme precipitation (>P90); green indicates the "With ZWD" forecast is closer to MSWEP. (b) Hit/miss/false-alarm/correct-negative classification at four absolute thresholds.



(a) Six-hour accumulated precipitation: MSWEP target, "With ZWD", "Without ZWD".



(b) Regional spatial RMSE over time.

Figure S17: Central America Hurricane Delta (2020-10-04 06Z): a more neutral case. The bulk regional RMSE (b) is essentially unchanged between the two models, yet the extreme-tail skill gain persists (Figure 2e of the main text), mirroring the other two events.