npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-025-01265-9

A regional high resolution AI weather model for the prediction of atmospheric rivers and extreme precipitation

Jorge Baño-Medina[1], Agniv Sengupta[1] ✉, Daniel Steinhoff[1], Patrick Mulrooney[1], Thomas Nipen[2], Mario Santa-Cruz[3], Yanbo Nie[1] & Luca Delle Monache[1]

Accurate precipitation forecasting often relies on high-resolution numerical weather prediction (NWP) models, which are essential for capturing fine-scale and nonlinear atmospheric dynamics. However, the computational demands of these models can be substantial. Leveraging recent advancements in artificial intelligence (AI), we present a stretched-grid AI-driven weather model with 6-km horizontal grid increments over the Western United States and ~31 km in other regions globally. The model employs an autoregressive framework to generate forecasts in minutes and is evaluated against global and regional NWP systems, as well as a lower-resolution AI model. Our results show that the regional AI model reduces 24-h accumulated precipitation errors, performs competitively with the regional NWP model, and effectively captures extreme precipitation events, particularly those linked to atmospheric rivers, which global coarser models often underestimate. This work underscores the potential of regional, high-resolution AI models for precipitation forecasting at km-scales, and discusses some of the challenges for future development.

Precipitation forecasting has traditionally relied on numerical weather prediction (NWP) systems, which solve prognostic differential equations and use sub-grid parameterizations to model the atmosphere. Given initial conditions, these models generate multivariate, four-dimensional global forecasts that capture the evolution of weather patterns across space—defined by latitude, longitude, and altitude—and time[1]. NWP systems also incorporate physical parameterization schemes that simulate microphysical processes such as the formation, coalescence, and growth of cloud droplets, the conversion of cloud water into rain, as well as convective-scale processes, boundary layer effects, and the land surface system[2]. These components are crucial for improving the accuracy and representativeness of precipitation forecasts. To better resolve and simulate mesoscale phenomena, such as terrain-induced flows and precipitation processes, an increasingly common approach is to employ limited-area models with finer spatio-temporal resolution. These regional models, while still requiring sub-grid parameterizations, allow for improved representation of physical processes compared to coarser global models over targeted domains. In this approach, coarser global NWP model output is used as initial and lateral boundary conditions to force the simulation in the inner domain[3–6]. However, high-resolution simulations demand significant computational resources, and forecasts still exhibit systematic biases and errors[7,8].

Artificial intelligence (AI)-driven weather models present a unique opportunity to generate precipitation forecasts with very short inference times[9–12], using minimal computational resources (e.g., 1 GPU for inference). The bulk of the computational burden for these models is tackled offline during their training, which involves learning a function describing the relationship between two consecutive atmospheric states separated in time (e.g., 6 h), typically using global reanalysis datasets such as the European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis version 5 (ERA5[13]), which contains a long record of pressure-level and single-level variables at 31-km horizontal resolution. Once trained and given an initial condition, these AI-driven models are capable of generating long rollouts of global forecasts in a matter of seconds to minutes with limited computational resources (and a fraction of the energy needed by traditional NWP methods), each time using the prediction of the previous time-step as the initial condition (i.e., in an autoregressive "mode"). However, these models often tend to underestimate precipitation totals and produce overly smoothed fields[11,14–16], which limits their effectiveness in precipitation forecasting. One promising alternative is to train AI models on regional, high-resolution datasets, either in a downscaling setting[17,18] -where a learnable function is tasked to map low-resolution variables to high-resolution ones at the same instant of time- or in a weather-forecasting/autoregressive "mode", such as limited-area[19] or stretched-grid

[1]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. [2]Norwegian Meteorological Institute, Oslo, Norway. [3]European Center for Medium-Range Weather Forecasts, Reading, UK. ✉e-mail: agsengupta@ucsd.edu

configurations[20]. Limited-area AI models are trained to predict weather evolution within a confined, high-resolution domain, using boundary conditions that are extracted from coarser global data and typically span only a few hundred kilometers. In contrast, stretched-grid AI setups model the entire globe with spatial resolution refined over the region of interest. In addition, alternative loss functions to the classical mean square error can alleviate the characteristic smoothing from AI-based weather forecasts, approximating the effective resolution of the fields towards the dataset's one they were trained on[21].

Precipitation forecasting in the Western United States presents significant challenges due to the region's complex topography and diverse climate zones, ranging from coastal to desert environments (see Fig. S1). Atmospheric rivers (ARs)—narrow corridors of concentrated water vapor in the atmosphere—play a major role in regional precipitation and are crucial for water resource management[22]. Notably, 85% of California's interannual precipitation variability has been linked to precipitation on just the wettest 5% of days each year, most of which are associated with AR events[23]. Currently, operational forecasting tools are primarily physics-based, with machine learning (ML) methods largely used for post-processing dynamical model outputs[24]. While global autoregressive models with uniform grid spacing—such as FourCastNet[25]—have demonstrated promise in forecasting ARs[26,27], there remains a lack of high-resolution, autoregressive AI-driven models specifically tailored for this purpose. Furthermore, comprehensive comparisons between these emerging AI approaches and state-of-the-art physics-based models are still lacking for ARs and precipitation in the region.

In this study, we leverage a high-resolution (6 km) regional reanalysis dataset from the Center for Western Weather and Water Extremes (CW3E), which spans Western and Central North America and the North Pacific, in conjunction with the global ERA5 dataset at ~31 km resolution, to train an AI-driven stretched-grid model based on a graph-transformer architecture (see Methods section). The model predicts multiple single- and pressure-level weather fields (see Training datasets and variables section for a list of variables). Precipitation forecasts over the Western United States are evaluated and compared against three baselines for three winter periods (2020–2021, 2021–2022, and 2022–2023): a coarser version of the AI stretched-grid model (architecturally similar to the Artificial Intelligence Forecasting System (AIFS) from ECMWF[12]), and two state-of-the-art NWP models—the Integrated Forecasting System (IFS) from ECMWF at 0.25° resolution and a 9-km version of the Weather Research and Forecast limited-area model tailored for the prediction of ARs over the Western US (West-WRF[8]). The configuration of high-resolution AI models is also explored by analyzing sensitivity to key training hyperparameters, such as learning rate, loss function, and variable scaling. This work shows the potential to predict (extreme) precipitation events associated with ARs, and also offers a comprehensive and thorough evaluation of regionally focused AI models in km-scale precipitation forecasting, extending previous analysis[20].

Results

The results were computed over the test period, which includes forecasts initialized at 00 UTC during three winter seasons—2020–2021, 2021–2022, and 2022–2023—each defined as November 1 through March 31. With 75% of the total annual precipitation being received during the winter months of November through March, this extended winter season has been used for model verification following a similar definition used in prior studies[28,29]. Precipitation fields from the parameter-elevation regressions on independent slopes model (PRISM[30]) at daily temporal resolution are used as the observational reference (see Methods) for validating both the AI and NWP models. To enable a consistent comparison, all forecast fields were regridded to PRISM's 4-km horizontal resolution using budget-conserving interpolation[16,31,32], preserving domain-integrated precipitation totals. Regridding to a common high-resolution grid is essential for inter-comparison purposes and for assessing the added value of (AI-driven) regional models compared to global models, one of the central objectives of this study. AI models are initialized directly from reanalysis data, whereas dynamical models use analysis fields. This distinction may give AI models an advantage by providing more accurate initial conditions. In the future, we plan to fine-tune on analysis datasets for operational purposes (see Discussion section). Precipitation outputs from both the dynamical and AI models are available at 6-hour intervals. To compute daily precipitation, we use the −12 to +12 UTC window, consistent with PRISM's definition of a daily field[30]. Under this definition, a 1-day lead time corresponds to the +12-h to +36-h interval from the initial time, and subsequent lead times follow accordingly. The validation domain is located in the western United States, within the broader 6-km regional dataset domain, which was defined to capture the inland penetration of ARs (Fig. S1).

Metrics-based evaluation of precipitation fields

Figure 1 presents key precipitation statistics over the validation domain (in rows), including 24-h accumulated precipitation (hereafter referred to as daily precipitation) mean, the frequency of rainy days, and the 50th, 90th, and 98th percentiles of the wet-day distribution at each grid point. Here, we follow the convention used by the US National Weather Service, defining wet or rainy days as those with recorded precipitation exceeding 0.01 inches, which is approximately equivalent to 0.25 mm. Additionally, the AR precipitation ratio is shown, which represents the proportion of precipitation on days associated with ARs relative to total precipitation (see Table 1). These statistics were computed over the test period.

The first column shows the reference values from PRISM, while columns two through five display the values from the IFS, West-WRF, the global AI model (hereafter referred to as AI 31 km), and the stretched-grid AI model (hereafter referred to as AI 6 km), respectively, considering a 5-day lead time (i.e., accumulated precipitation in the 108–132 h lead time interval). PRISM's precipitation statistics reveal a clear spatial pattern characterized by wetter conditions along the coast and across the Sierra Nevada Mountains in northeastern California, and drier conditions across the mostly desertic central and eastern regions of the validation domain. These inland areas experience lower precipitation frequency and intensity, including less intense high-impact events, as shown by the 98th percentile values. Precipitation totals are highest at northern latitudes, with more than 50% of days classified as rainy in the Pacific Northwest states of Washington and Oregon, and gradually decreasing moving southward along the Pacific coast. On average, the validation domain receives 2.16 mm of precipitation per day, with 35.94% of days classified as rainy. The averaged 50th, 90th, and 98th percentiles of the wet-day distribution are about 3, 13, and 23 mm, respectively. Precipitation in this region arises from a range of atmospheric mechanisms. In the Sierra Nevada Mountains, both convective processes and orographic lifting contribute significantly[33]. Additionally, ARs are responsible for a substantial portion of the accumulated precipitation, contributing as much as half of the total precipitation in the Sierras (see column 1, last row), during the test period. While their impact is most pronounced along the western portion of the domain, many ARs penetrate deeply into the Intermountain West, highlighting their extensive reach across the continent.

Both the dynamical and AI-based models can capture the key spatial precipitation patterns described above, though with notable differences in performance. The global models tend to underestimate daily precipitation totals in areas of high precipitation and the upper percentiles of the wet-day distribution—particularly the 98th percentile—while overestimating the frequency of rainy days almost everywhere in the validation domain (see Fig. S2). This overestimation may partly reflect inherited wet biases from ERA5, which is used in model training or initialization[34]. In contrast, the regional models—West-WRF and the stretched-grid AI 6 km model—show much better agreement with PRISM in capturing the magnitude of high-impact precipitation events, especially the 98th percentile, and frequency of rainy days (see Fig. S2). For the percentile 50th, both West-WRF and the AI 6-km underestimate the values in the southern part of the validation domain, presenting similar bias patterns across the area of study among them. Additionally, they provide much finer spatial detail in the fields than IFS and

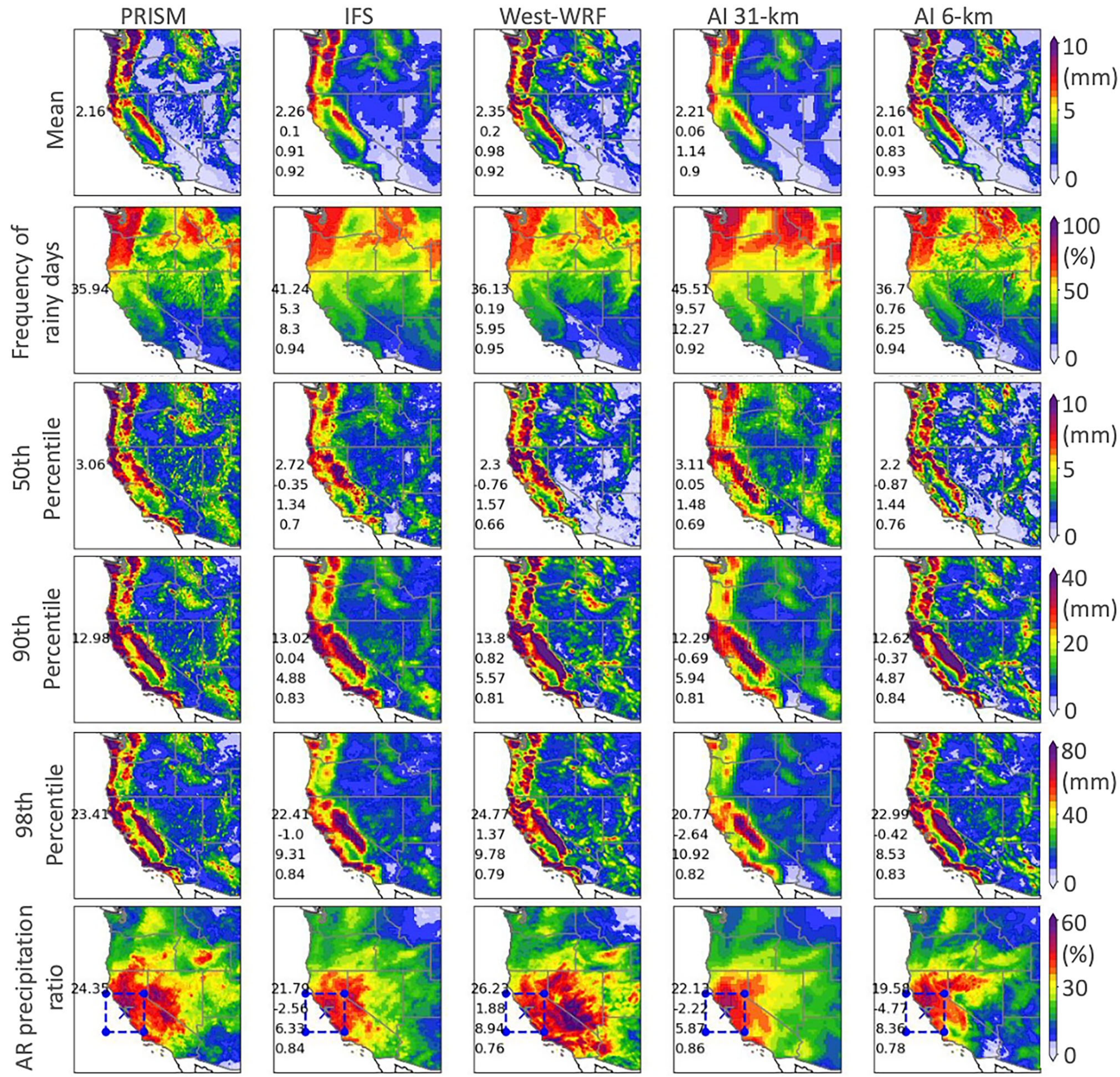


Fig. 1 | Observed and predicted key daily precipitation statistics from dynamical and AI models (both global and regional) over the validation domain at a 120-h forecast lead time (i.e., accumulated precipitation in the 96–120-h lead time interval) during the test period. The blue box in the bottom row highlights a 200 × 200-km region centered at 37.5°N, 122.5°W, which corresponds to the reference grid point used to identify days associated with ARs. Values in rows 2 to 5 within each sub-panel show, in descending order, the mean of the field (mm), the bias (mm), the root mean square error (RMSE; mm), and the correlation coefficient. The last three values are computed relative to the observed field, shown in column 1.

the AI 31-km, specifically the centers of precipitation maxima in the Sierra Nevada and Cascade ranges. This result is consistent with the overly smooth spatial fields typically produced by AI models[35], and underscores the added value of high-resolution regional modeling for reproducing precipitation.

Figure 2a–g illustrates the performance of the models as a function of forecast lead time across various metrics: root mean square error (RMSE), RMSE over the blue bounding box shown in Fig. 1 where ARs contribute greatly to precipitation for days associated with ARs (see Table 1), Spearman correlation, and the empirical threshold score (ETS) for different thresholds: 0.2 mm, and the 50th, 90th, and 98th percentiles, computed at each grid point. The fields are vectorized before the computation of the validation metric, yielding a single value per model and forecast lead time. Forecast errors increase with lead time, rising from about 3 to 4.5 mm for the AI 31-km model. At lead times beyond 3 days, both the IFS and AI 6-km models exhibit higher errors than AI 31-km. In contrast, West-WRF shows errors ranging from 4 mm at a 1-day lead to 6 mm at 6 days. This pattern holds across other metrics: correlation and ETS values are generally highest for AI 31-km, with IFS and AI 6-km performing comparably but slightly worse, and West-WRF performing the lowest. As expected, the empirical threat score (ETS) decreases with increasing thresholds (see the second row of Fig. 2). Errors are especially pronounced for AR-related precipitation because of the larger rainfall amounts involved.

The AI models, particularly the AI 31-km, tend to produce overly smoothed prediction fields, likely due to the use of mean squared error as the loss function during training, which may partly explain the lower error values observed in Fig. 2. To further explore this, the distribution of precipitation on rainy days over the validation domain—defined as days with observed precipitation exceeding 0.25 mm at each grid point—is shown in

Table 1 | List of AR events during the test period with a rank of 2 or higher that impacted the San Francisco area (37.5°N, 122.5°W)

ID	Start date	Max date	End date	AR scale	Max IVT ($kg\ m^{-1}\ s^{-1}$)
1	2020-11-17T00	2020-11-18T00	2020-11-19T00	2	590
2	2020-12-11T00	2020-12-12T12	2020-12-13T00	2	655
3	2021-02-14T00	2021-02-15T00	2021-02-17T00	2	498
4	2021-12-12T00	2021-12-13T00	2021-12-15T00	3	500
5	2022-12-26T00	2022-12-27T12	2022-12-28T00	4	1095
6	2022-12-29T00	2022-12-31T00	2023-01-01T00	3	606
7	2023-01-04T00	2023-01-05T00	2023-01-06T00	3	832
8	2023-01-07T00	2023-01-08T12	2023-01-09T00	2	664
9	2023-01-09T00	2023-01-09T12	2023-01-10T00	2	817
10	2023-01-11T00	2023-01-13T12	2023-01-14T00	3	598
11	2023-03-09T00	2023-03-10T12	2023-03-11T00	3	963
12	2023-03-14T01	2023-03-14T15	2023-03-16T00	2	742

The ranks are assigned following the AR scale developed by Ralph et al. (2019[56]).

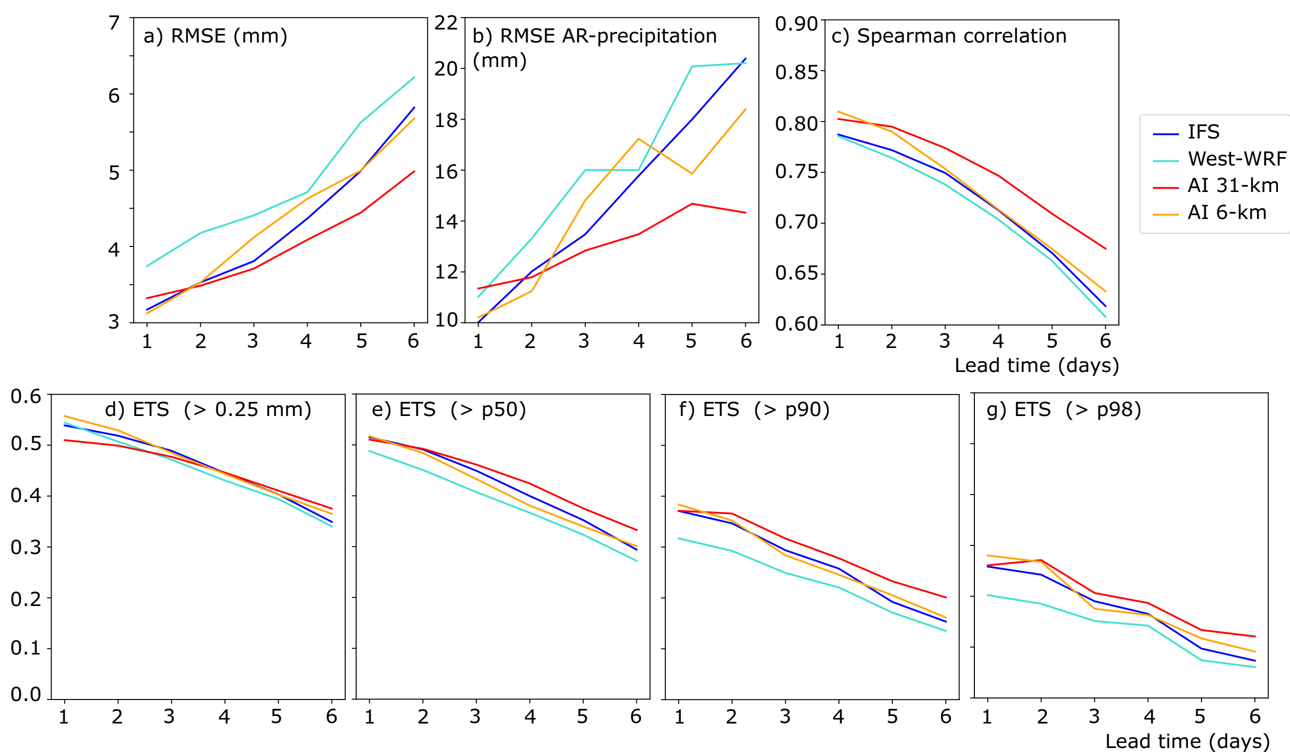


Fig. 2 | Validation of daily precipitation for dynamical and AI models as a function of forecast lead time. **a** RMSE (mm), **b** RMSE (mm) over the blue bounding box shown in Fig. 1 for days associated with ARs, **c** Spearman correlation, and the empirical threshold score (ETS) for different thresholds: **d** 0.25 mm, and the **e** 50th, **f** 90th, and **g** 98th percentiles, computed at each grid point. Results are shown up to 7 days of forecast lead time. Models displayed here include dynamical global (IFS, in blue) and regional (West-WRF, in turquoise), as well as AI global (AI 31-km, in red) and regional (AI 6-km, in orange) models.

the quantile-quantile maps in Fig. 3 (top row) for (a) 2-day, (b) 4-day, and (c) 6-day forecast lead times. The main panels display percentiles from the 1st to the 99th at 2-percentile intervals, while the inset panels focus on extreme values, showing the 99th to 100th percentiles in increments of 0.0001. The 1:1 diagonal line represents perfect alignment between the observed and predicted distributions. All models capture the lower percentiles of precipitation well. However, the global models (IFS and AI 31-km) tend to underestimate precipitation above 20 mm, with underestimation becoming more pronounced at higher percentiles (above 40 mm). In contrast, the high-resolution models (AI 6-km and West-WRF) show little to no underestimation up to the 99th percentile, closely aligning with the 1:1 line. For the most extreme precipitation values (above the 99th percentile; see inset panels), the AI 6-km model begins to underestimate precipitation above 150 mm, while West-WRF remains consistent with observations for a 2-day lead time. The IFS and AI 31-km models show even stronger underestimation than in the 25–50 mm range. As lead time increases, all models gradually underestimate the highest precipitation values more strongly, indicating a loss of predictability for extreme events. Notably, the AI 31-km model exhibits nearly the same underestimation of the highest percentiles across lead times, suggesting that its forecasts are inherently very smooth. These results, along with the validation metrics in Fig. 2, suggest that the low error values of the AI 31-km model may arise from its smooth, low-intensity fields, while the AI 6-km model achieves lower error values than West-WRF while showing accurate estimates of

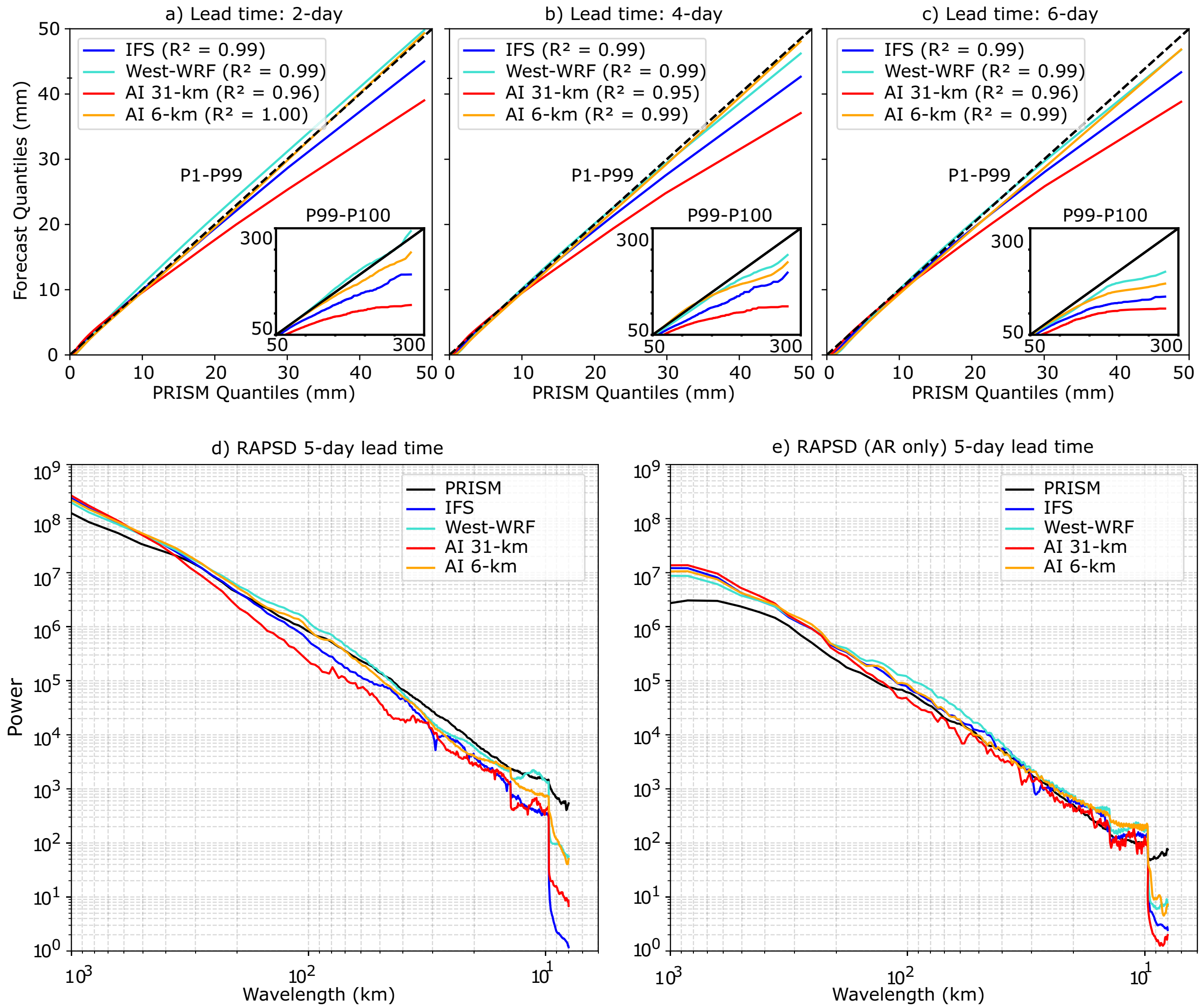


Fig. 3 | Validation of intensity and spectral frequency of the distribution of precipitation from dynamical and AI models. Quantile-quantile maps showing the distribution of precipitation on rainy days over the validation domain-defined as days with precipitation exceeding 0.25 mm at each grid point-for **a** 2-day, **b** 4-day, and **c** 6-day forecast lead times. The main panels display percentiles from the 1st to the 99th at two-percentile intervals, while the inset panels focus on extreme values, showing the 99th to 100th percentiles in increments of 0.0001. The 1:1 diagonal line represents perfect alignment between the observed and predicted distributions. Four models are shown here: IFS (in blue), West-WRF (in turquoise), AI 31-km (in red), and AI 6-km (in orange). Panel (**d**) shows the radially averaged power spectral density (RAPSD) at 5-day forecast lead time, and Panel (**e**) does the same but considering only the samples related to AR precipitation. PRISM, shown in black, serves as the observational target.

precipitation intensity, particularly for the tails of the distribution associated with extreme precipitation events.

To further assess the realism of the precipitation fields' intensity and distribution, the radially averaged power spectral density is shown in the bottom row of Fig. 3 for all days in the validation period (Fig. 3d), as well as only for the days associated with ARs (Fig. 3e), at a 5-day lead time. PRISM is shown in black as a reference. The smoothness and underestimation of the precipitation fields for both the IFS and AI 31-km models are evident in these panels, with low values at higher frequencies. In contrast, the regional models closely follow the power spectral density of the PRISM observational target, showing only a slight underestimation at the highest frequencies, much less than the coarser-resolution models. This behavior is practically identical when considering only the days linked to AR events.

The results from analyzing location error (Fig. 2) and the distribution of precipitation intensity (Fig. 3) demonstrate significant potential for regional models, such as West-WRF, to improve when considering neighborhood metrics like the fraction skill score (FSS). While West-WRF exhibited higher error values compared to models like AI 31-km, this was partly due to AI 31-km's tendency to significantly underestimate the tail of the precipitation distribution. The FSS addresses this by sacrificing location error in favor of intensity error, computing accuracy over a surrounding window of a specified size around each grid point. The larger the window, the more intensity errors are penalized in the FSS. Figure 4 shows the FSS at 2-day, 4-day, and 6-day forecast lead time for the IFS (blue), West-WRF (turquoise), AI 31-km (red), and AI 6-km (orange), as a function of window size, considering a threshold of 20 mm (see Methods section for details on the computation). For all models, the FSS increases with window size, meaning that larger windows place more emphasis on intensity accuracy rather than location accuracy, which leads to higher FSS values. For small window sizes, West-WRF shows FSS values similar to the IFS, which is a coarser model. However, as the window size increases, the FSS for West-WRF rises more rapidly than for the coarser models, even surpassing the AI 6-km, for window sizes beyond approximately 280 by 280 km (70 by 70 grid points) and 160 by 160 km (40 by 40 grid points) for 2-day and 4-day forecast lead time, respectively. This suggests that when location error is sufficiently penalized, West-WRF outperforms the coarser models. The AI

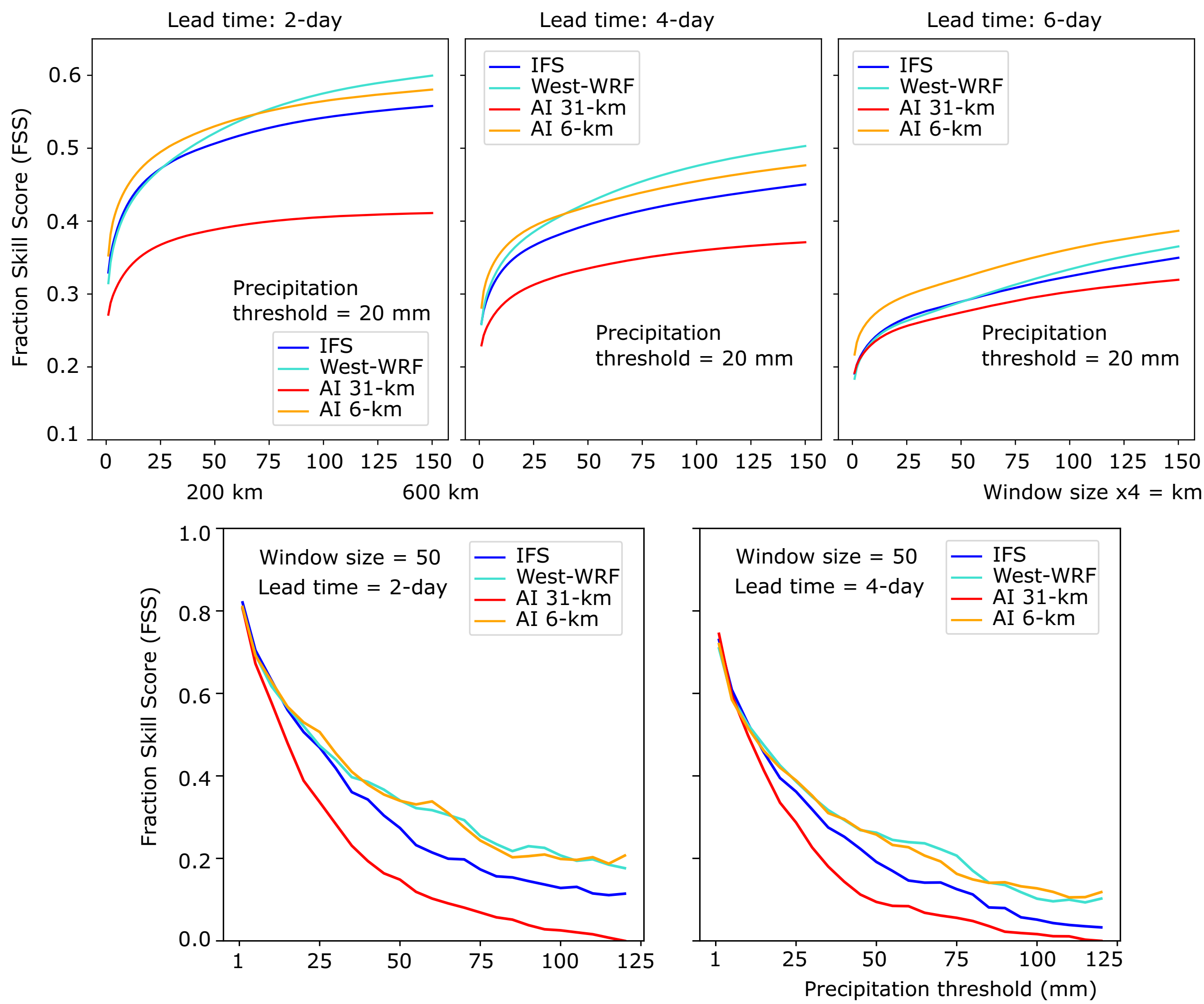


Fig. 4 | Fraction skill score (FSS) for dynamical (IFS and West-WRF) and AI (AI 31-km and AI 6-km) models at 2-day, 4-day, and 6-day forecast lead times. In the top row, results are presented for the IFS (blue), West-WRF (turquoise), AI 31-km (red), and AI 6-km (orange) as a function of window size, with a threshold of 20 mm (see section Methods for computation details). For FSS, higher values indicate better performance. In the bottom row, the FSS is plotted against different precipitation thresholds for 2-day and 4-day lead times, using a fixed window size of 50 by 50 grid points (i.e., 200 km by 200 km).

6-km performs better than both the IFS and AI 31-km models, and remains competitive with the regional West-WRF, even showing higher FSS values at 6-day lead time, further emphasizing its ability to deliver accurate precipitation forecasts in terms of both location and intensity. Additionally, the AI 6-km is also competitive with or even outperforms West-WRF when considering a wider range of precipitation thresholds in the computation of FSS (bottom row in Fig. 3), and clearly surpasses the IFS and the AI 31-km.

Case study: atmospheric river event of January 5th, 2023

Forecasts were initialized on January 2 at 00 UTC, three days before the event's peak (see Table 1). The AI 31-km model fails to resolve fine-scale features, producing broad spatial patterns that diverge from the PRISM reference. The IFS captures some aspects of the precipitation intensity and distribution, though its representation remains coarser than PRISM. By contrast, the high-resolution models—West-WRF and AI 6-km—generate more detailed structures that closely resemble PRISM and are consistent with the RAPSD results in Fig. 3. Notably, the AI 6-km model reproduces the spatial distribution of precipitation with elevated rainfall over both the San Francisco Bay Area and the Sierra Nevada, including the observed maximum in the northern Sierra relative to the central and southern regions. The AI 31-km model, however, substantially underestimates precipitation in the most extreme rainfall areas, consistent with Fig. 3. For the dynamical models, precipitation tends to be overpredicted in the Sierra Nevada and underestimated in the Coastal Ranges, a persistent bias in NWP models[8]. This bias is less pronounced in the AI-based simulations, with the AI 6-km model even showing some underestimation in the southern Sierra.

Discussion

This study evaluates the added value of a stretched-grid artificial intelligence (AI)-driven weather model, with a refined 6-km grid spacing over the North Pacific and a large portion of North America, and 31-km elsewhere, compared to its coarser version, based on a uniform 31-km resolution. The focus is on precipitation forecasting performance, benchmarked against both global and regional state-of-the-art dynamical models. The analysis centers on the Western United States, a region characterized by high spatial variability, including mountainous terrain, desert landscapes, coastal zones, and high-impact weather events often associated with ARs, which are key to water management in the region[22]. This allows us to test and compare the models under different climate conditions with very distinct precipitation statistics.

The AI 6-km model effectively reproduces key precipitation statistics across the domain, delivering more realistic forecasts than its global counterpart (Fig. 1). While the AI 31-km achieves lower root mean square error

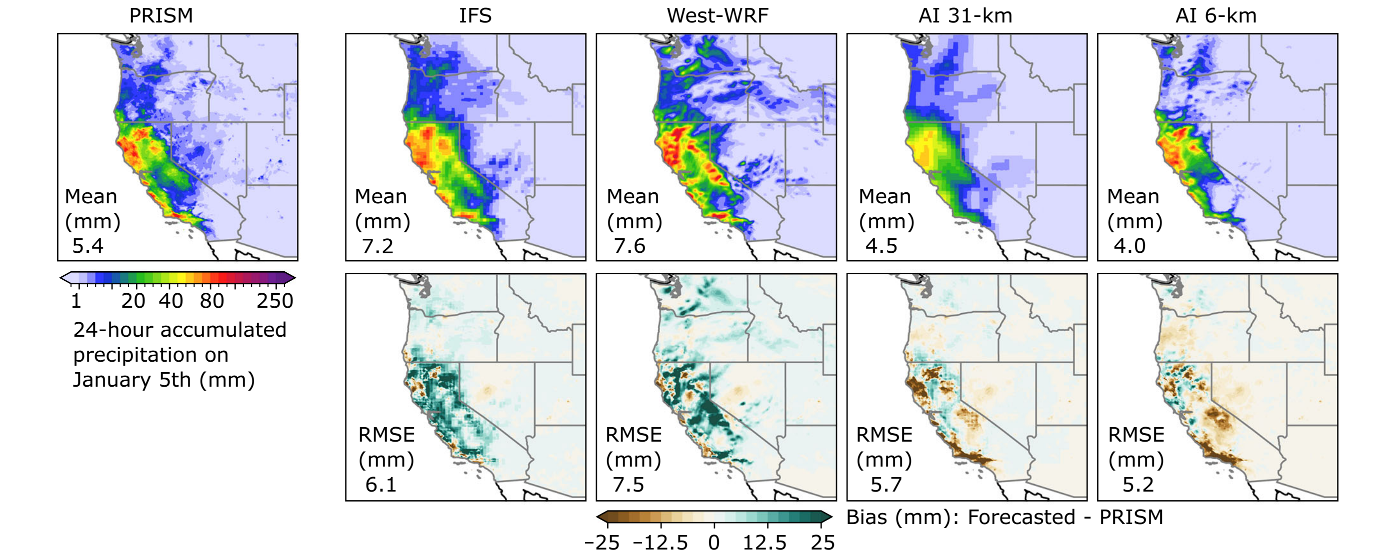


Fig. 5 | Case study for an atmospheric river of rank 3 from January 5, 2023. The top row shows 24-h accumulated precipitation over the validation domain from the observational reference PRISM, dynamical models: IFS and West-WRF, and the AI 31-km and AI 6-km models. The bottom row displays the difference between their forecasted 24-h accumulated precipitation and PRISM. All forecasts were initialized on January 2 at 00 UTC, 3 days prior to the maximum date of the event (see Table 1).

(Fig. 2), the AI 6-km offers greater spatial detail, largely due to training on a finer-scale dataset. It also exhibits little-to-none underestimation of extreme precipitation percentiles and a power spectral density more consistent with observations (Fig. 3). Neighborhood-based verification metrics—which account for spatial and temporal phase errors by relaxing location accuracy in favor of intensity accuracy—reveal that the global AI 31-km performs poorly as compared to the other methods tested, whereas the AI 6-km is competitive with a 9-km version of the Weather Research and Forecast limited-area model tailored for the prediction of ARs over the Western US (West-WRF), and significantly outperforms the European Center for Medium-Range Weather Forecasts (ECMWF)'s global dynamical model, the integrated forecasting system (IFS; Fig. 4)). These results suggest that regionally focused AI models not only surpass their coarser global counterparts but also rival limited-area dynamical models in performance, all while operating at a fraction of the computational cost (e.g., 6–7 min to generate a 7-day forecast with the Stage B2 AI 6-km model on a single H100 GPU). Notably, the AI 6-km model demonstrates strong skill in capturing both the location and intensity of precipitation, including during high-impact events (Fig. 5).

Here, we initialized the AI models with reanalysis data, which might be an advantage as compared to the dynamical models, which were initialized with global operational analysis and require some iterations to spin-up. Thus, additional model fine-tuning on operational datasets, such as initialization fields from the IFS or the global forecast system (GFS) run by the United States National Weather Service, may be essential for deploying these models in real-time operations. Without this adaptation, forecast errors can arise due to distributional shifts between the operational input data and the data used during training[14,36]. Another important aspect toward the operationalization of AI-based weather modeling is the ability to interpret and understand the spatio-temporal dynamics the models have learned[37]. This is crucial for two main reasons. First, it fosters confidence and trust within the scientific community[38,39]. Second, it enables broader applications beyond weather forecasting, such as the analysis of atmospheric dynamics, the design of observation networks, predictability studies[40], and sensitivity analyses that help identify key upstream regions influencing forecasts at specific locations[37]. A notable example of such broader applicability is the potential use of AI weather models within the framework of the Atmospheric River Reconnaissance program[41]. This initiative aims to improve initial condition observations by coordinating targeted flight campaigns in the Pacific Ocean, guided by adjoint-based sensitivity analyses. AI-based sensitivity approaches could complement these efforts by offering alternative or enhanced identification of critical observation regions.

The stretched-grid AI model was able to capture the important characteristics of precipitation associated with ARs, showing accurate values in both intensity and location as exemplified by the case study presented here. This is particularly relevant given the fact that climate change simulations indicate that ARs are projected to become more frequent and to produce more intense precipitation[42,43]. The performance of AI models under out-of-distribution climate change conditions—regimes not encountered during training—has been evaluated in the context of a climate attribution experiment involving an AR event[27], as well as through testing on general circulation model (GCM) outputs[44], both of which reveal limitations in the models' extrapolation capabilities. In this study, we analyzed a suite of ARs across five ranks defined by the AR scale and tested the model on the anomalous 2022–2023 winter season, during which cumulative water totals exceeded historical norms due to a sequence of high-impact AR events. This serves as a preliminary assessment of the model's extrapolation ability under extreme conditions. As previously described, the results are promising, with AI models being able to reproduce the main features of precipitation associated with ARs. Future investigations could focus on tailored cross-validation strategies—for instance, withholding high-impact AR events from the training set, as demonstrated in recent studies on cyclones[45].

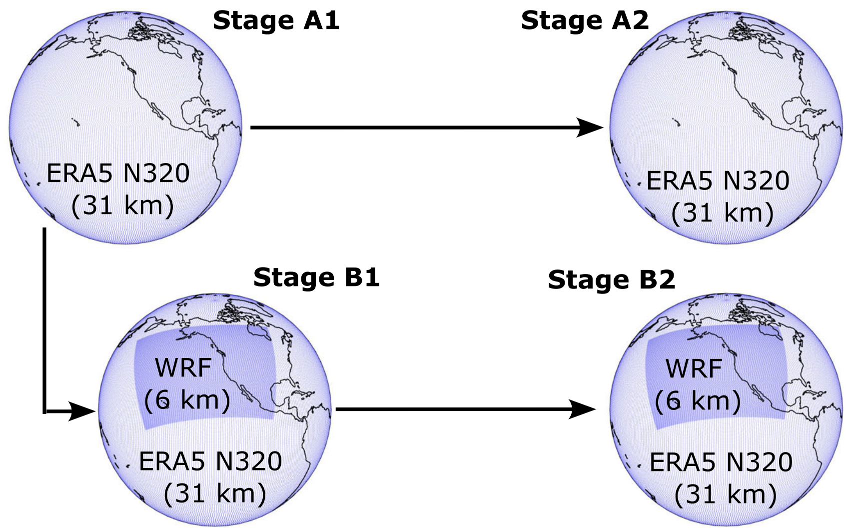
Probabilistic forecasting represents a natural extension of this work, particularly given the relatively low computational cost of generating large ensembles with AI-based models -potentially enabling a better representation of the tails of the precipitation distribution[26,35,46]. Strategies for ensemble generation include training with alternative loss functions[21], employing novel deep-learning architectures such as diffusion models[18,46–49], and propagating uncertainty in initial conditions and model parameters through, e.g., multi-model ensembles[26,50]. These approaches have yielded well-calibrated ensembles, with CRPS- and diffusion-based methods in particular producing sharper fields at the level of individual members than deterministic fields.

Methods

AI models: stretched-grid

Stretched-grid AI-driven weather models were first introduced by Nipen et al. (2024)[20] to enhance spatial resolution over targeted regions by training on a combination of high-resolution regional datasets and lower-resolution

Fig. 6 | Schematic to train the global and stretched-grid AI-driven weather models. The figure shows the four training stages, i.e., A1, A2, B1, B2, and the table contains the parameters for each of them.



	Stage A1	Stage A2	Stage B1	Stage B2
Global dataset	ERA5 N320	ERA5 N320	ERA5 N320	ERA5 N320
Regional dataset	-	-	WRF 6-km	WRF 6-km
Global refinement	7	7	7	7
Regional refinement	-	-	9	9
Input to hidden	12	12	12	12
Hidden to output	3	3	3	3
# of grid points	540 K	540 K	1.8 M	1.8 M
# of mesh nodes	164 K	164 K	1.1 M	1.1 M
# of edges	4.9 M	4.9 M	27.5 M	27.5 M
Local learning rate	3.125 e-5	8 e-7	4e-6	3e-6
# of GPUs (nodes)	64 (16)	64 (16)	64 (16)	64 (16)
Training period	1979-2020	2012-2020	2012-2020	2012-2020
# of steps	260 K	8.5 K	24.7 K	5.6 K
# of GPU hours	13 K	1 K	2.7 K	2.6 K
Rollout	1	1-12 (10)	1	1-8 (3)

global datasets. In our case, we leverage CW3E's 6-km high-resolution model (see Section Training datasets in the Methods for details), which covers the North Pacific and a significant portion of North America (see stages B1 and B2 in Fig. 6 and Fig. S1 in Supplementary material). For the global forcing, we use ERA5 data provided on its N320 reduced Gaussian grid (i.e., 320 latitudes), which contains ~500,000 grid points—about half as many as the standard 0.25-degree latitude-longitude version. This reduction is crucial for optimizing memory usage and computational efficiency. The combination of the high-resolution regional model and the N320 global forcing results in ~1.8 million grid points for both input and output graphs. The model architecture is a graph-transformer, composed of an encoder, a processor, and a decoder. It builds on the principles of graph neural networks (GNNs[11,51]), where nodes represent spatial locations and edges encode the relationships or interactions between them, allowing the model to capture both local and long-range spatial dependencies effectively. In the context of weather modeling, notable examples of GNN-based architectures include GraphCast by DeepMind[11] and AIFS by ECMWF[12], which employ GNNs and graph-transformers, respectively. The stretched-grid model proposed here, along with that of Nipen et al. (2024)[20], can be viewed as an extension of the AIFS framework toward higher-resolution, regionally focused domains. An encoder-processor-decoder architecture is used to learn changes in weather patterns relative to the previous state. These predicted changes are then added to the input fields through a residual connection to produce the final model outputs. The architecture operates on a graph structure that includes nodes at the model's native resolution as well as in a latent representation, obtained by recursively refining an icosahedron multiple times. Each component of the architecture, i.e., the encoder, processor, and decoder, leverages these nodes to develop its own subgraphs, to perform its respective computations.

1. Encoder: the encoder subgraph maps the input weather data into a latent representation with reduced spatial resolution, effectively compressing the input information. The resolution of the latent space is determined by refining an icosahedron seven times for the global domain and nine times for the regional domain. This approach mirrors the AIFS model's refinement strategy for global coverage and follows Nipen et al. (2024)[20], who applied increased refinement over the regional domain. As in Nipen et al. (2024)[20], each node in the latent grid aggregates information from its 12 nearest neighbors in the native grid, allowing the model to incorporate spatial context during encoding. The number of channels (i.e., features per node) is 172 (i.e., 86 variables per time frame) in the input space—derived from the prognostic and forcing variables (see Training datasets section)—and is increased to 512 in the latent space via a feed-forward neural network, consistent with the configuration used in GraphCast[11].
2. Processor: the processor consists of a sequence of 16 layers (or blocks) that perform message passing over the latent space. These layers exploit the multi-scale spatial relationships—encoded as edges in the processor subgraph—of the refined icosahedral grid to iteratively update node representations. Information is aggregated from both local and global contexts, allowing the model to capture a wide range of spatial dependencies. Each layer employs self-attention mechanisms

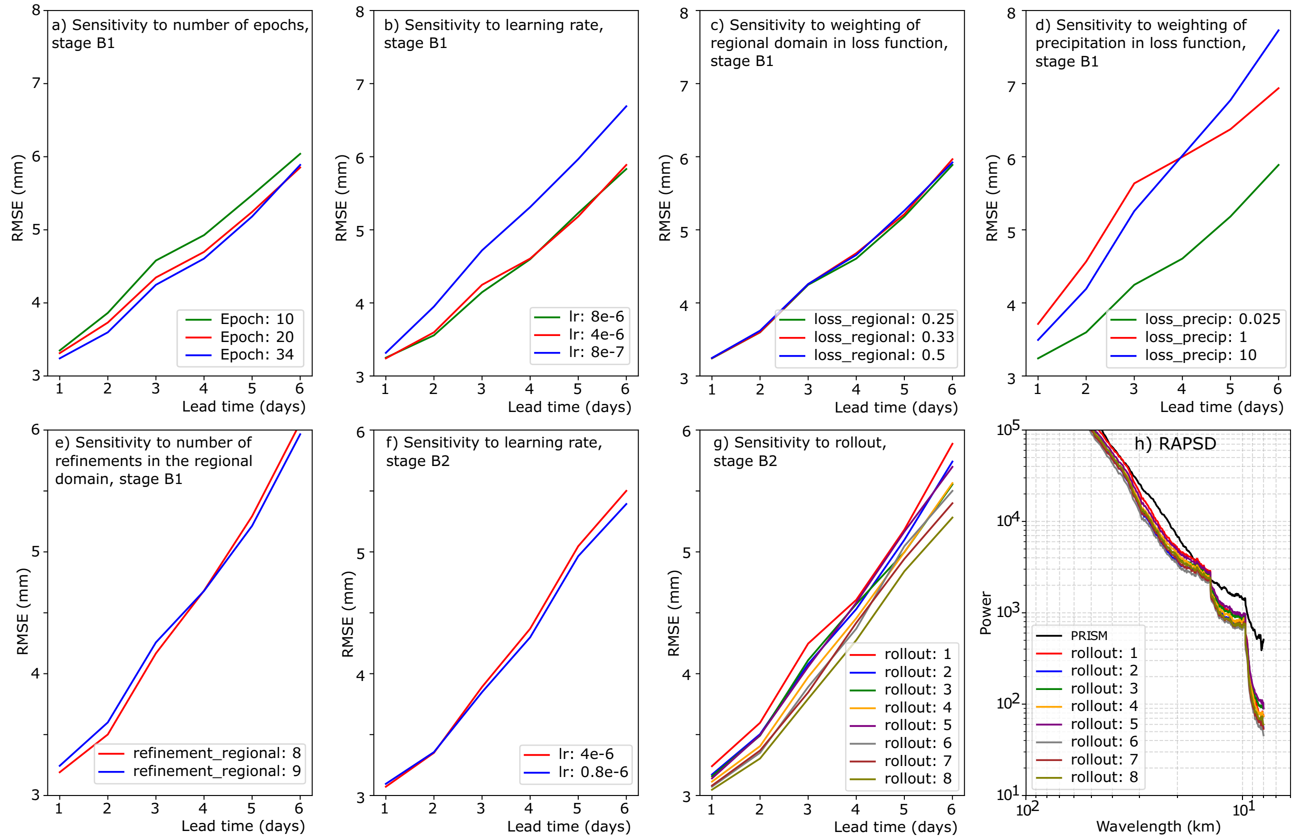


Fig. 7 | Sensitivity of the AI 6-km forecast error to certain hyperparameters. **a** number of epochs, **b** learning rate in stage B1, **c** scaling of the regional domain in the loss function, **d** scaling of precipitation in the loss function, **e** the number of refinements of the icosahedron over the regional domain, **f** learning rate in stage B2, **g** rollout, and **h** RAPSD across different rollouts.

alongside multi-layer feed-forward networks to transform the latent node features, progressively enhancing the expressiveness and predictive capacity of the latent space.

3. Decoder: the decoder subgraph maps the latent space back to the native space. The number of channels matches the number of prognostic and diagnostic variables 6 h into the future, encompassing a total of 76 variables. In this step, each node in the native grid aggregates information from its three nearest neighbors in the latent grid[11,12,20].

The training framework is illustrated in Fig. 6 and consists of four stages: A1, A2, B1, and B2. The specific characteristics of each stage—such as the number of epochs, learning rates, and other hyperparameters—are detailed in the accompanying table within the same figure. Stage A1 uses a configuration similar to that of AIFS and serves as a global pre-training phase. This stage is primarily necessary due to the limited temporal coverage of the high-resolution dataset (only 8 years for training). If a longer dataset were available (e.g., 30 years), this step could potentially be omitted. Stage B1 builds upon Stage A1 via transfer learning, initializing training from the pre-trained weights and incorporating the 6-km regional dataset into both the input and output graphs. This stage fine-tunes the encoder, processor, and decoder to adapt the model to the higher-resolution regional domain. Both A1 and B1 minimize the weighted root mean squared error (RMSE) for a 6-h lead time forecast (i.e., a single autoregressive step). The loss function weights each node based on the area it represents on the Earth's surface, which is estimated by using Voronoi cells. It also includes variable- and level-specific weights to balance the contribution of different atmospheric fields. In Stage B1, the loss over the regional domain is rescaled to represent 25% of the total loss, thereby focusing training more heavily on the region of interest. Stages A2 and B2 extend the training to multi-step autoregressive forecasting, minimizing error across forecast lead times ranging from 6 to 72 h (i.e., 1 to 12 steps) and 6 to 48 h (i.e., 1 to 8 steps), respectively. The rollout is increased by one step after each epoch (706 training iterations per epoch), starting from 1 step and progressing up to 8/12 steps. This stage is crucial for improving the model's stability through forecast length (Fig. 7). Notably, Stage A2 is used solely for benchmarking the global model against the stretched-grid version. Key hyperparameters, such as the number of epochs, the learning rate, the re-scaling of the regional domain and precipitation in the loss function, the refinements of the icosahedron over the regional domain, and the rollout were optimized through an empirical search, which involved training multiple versions of the models (Fig. 7). For Fig. 1 through Fig. 5, we used the model configuration shown in Fig. 6. Although no single AI 6-km model consistently outperformed other models trained with different configurations of rollout across all metrics, we chose this version because it offers a strong balance between sharpness and forecast error.

The stretched-grid model (trained across stages A1, B1, and B2) requires ~12 days to complete training on 64 H100 GPUs, each with 96 GB of memory. The model is sharded across 4 GPUs per node, enabling a batch size of 16 through data parallelism—each of the 16 samples in a batch is processed simultaneously across the 16 nodes. The models were developed and trained within Anemoi, an open-source AI-modeling framework (http://anemoi.readthedocs.io)[12]. The reader is referred to Nipen et al. (2024)[20] for more details on stretched-grid AI models.

Dynamical benchmarks: IFS and West-WRF

The integrated forecasting system (IFS) from ECMWF is used as a benchmark for evaluating the AI-based models. IFS is a state-of-the-art global NWP system and has been widely used to intercompare global AI models in several recent studies[9,11,12,25,26]. In this work, we use the 0.25-degree version of IFS, which enables a direct comparison with the global AI model and serves as a reference for assessing the added value of the stretched-grid approach in the Western United States. IFS is initialized daily at 00 and 12 UTC and generates 16-day length simulations. Here, the forecasts initialized at 00 UTC for the test period were downloaded from the TIGGE data repository.

CW3E's West-WRF near-real-time (NRT) dynamical model is used as a regional benchmark[7,8]. West-WRF is based on the Weather Research and Forecasting (WRF) model[52,53]. It operates at a 9-km horizontal grid spacing (covering a domain nearly identical to that of the 6-km CW3E reanalysis), and has 100 vertical levels to a model top of 10 hPa. Full model output is available hourly. West-WRF is initialized and forced at its lateral boundaries using the ECMWF HRES model (0.1 degree grid spacing). West-WRF has been extensively fine-tuned to enhance the forecasting of ARs and associated precipitation, making it one of the most advanced numerical weather prediction systems in the region at these spatial scales.

Both IFS and West-WRF are initialized using ECMWF's operational analysis.

Training datasets and variables

Two datasets are used to train the stretched-grid AI-driven weather model: a global lower-resolution dataset and a regional higher-resolution one. The global dataset is ERA5[13], a state-of-the-art reanalysis product that integrates a wide range of weather observations with numerical forecasts through data assimilation. ERA5 currently provides the most comprehensive and accurate reconstruction of the atmospheric state at any given time. It offers hourly data at 31 km spatial resolution on the N320 projection for a broad range of atmospheric and oceanic variables across multiple pressure levels and single-level fields.

Specifically, 76 variables are used: 75 prognostic and one diagnostic—6-hourly accumulated precipitation. The prognostic variables include geopotential, temperature, zonal and meridional winds, and specific humidity at 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa), along with the following single-level variables: mean sea level pressure, surface pressure, 2-m temperature, 2-m dewpoint temperature, skin/sea surface temperature, integrated water vapor, horizontal components of integrated vapor transport, and horizontal components of surface winds. An additional 11 variables are used as forcing fields, including land-sea mask, surface geopotential, incoming solar radiation, the cosine and sine of Julian day, cosine and sine of longitude and latitude, and cosine and sine of local time.

The same set of variables is extracted from CW3E's 6-km reanalysis, which serves as the high-resolution regional dataset. ERA5 model-level output, the ERA5-LAND land surface model, and UK Met Office OSTIA SST are used as forcing. ERA5 is also used to "nudge" (four-dimensional data assimilation) the WRF solution above the boundary layer. This reanalysis was generated in a similar manner to the West-WRF NRT dynamical model described above, being based on the WRF dynamical model, but (1) using ERA5 as initial and lateral boundary conditions, (2) having a different spatial domain, and (3) using different options for parameterization schemes compared to the West-WRF NRT model. The reanalysis is initialized annually on September 1 (to minimize initialization errors in snow cover), with the first 15 days discarded for model spin-up. The reanalysis has 100 vertical levels to a model top at 10 hPa.

Observational dataset

The Parameter-elevation Regressions on Independent Slopes Model (PRISM[30]) is used as the ground truth in this study. PRISM provides daily precipitation data at a 4-km spatial resolution, making it well-suited for evaluating the models' ability to capture fine-scale precipitation processes. PRISM spans the entire continental US, but only data within the Western US validation domain were utilized. Moreover, PRISM has been widely used to validate machine learning models across the United States[24,32,54,55].

Atmospheric river catalog

During the test period—covering the winters of 2020–2021, 2021–2022, and 2022–2023—a total of 12 ARs with a scale rating of 2 or higher (see Ralph et al., 2019[56] for details on the AR scale) impacted the area surrounding San Francisco (approximate coordinates: 37.5°N, 122.5°W). Table 1 presents the start, peak, and end dates for each AR event, along with the maximum integrated vapor transport (IVT) values recorded by ERA5. Data were obtained from the Center for Western Weather and Water Extremes (CW3E) AR catalog (https://cw3e.ucsd.edu/Projects/ARCatalog/catalog.html).

Metrics

Given a prediction field $P \in {}^{D \times N}$ and an observation field $O \in {}^{D \times N}$, where D is the number of initial conditions and N is the number of grid points in the validation domain, the following list of metrics is used to validate the models:

1. Root mean square error (RMSE): The RMSE is computed as the root mean square difference between the prediction (P) and observation fields (O) (Equation (1)).

$$RMSE = \sqrt{\Sigma \frac{(P - O)^2}{ND}} \tag{1}$$

2. Empirical threshold score (ETS): The ETS measures the accuracy of a prediction model in forecasting events that exceed a specified threshold (see Equation (2)). It accounts for the number of hits (H; correct predictions of events above the threshold), false alarms (FA; predictions of events that did not occur), misses (M; events that occurred but were not predicted), and correct negatives (CN; correct predictions of non-events). The parameter E represents the number of hits expected by random chance and is computed based on the assumption that forecasts and observations are statistically independent (see Equation (3)). The ETS is defined as:

$$ETS = \frac{H - E}{H + FA + M - E} \tag{2}$$

$$E = \frac{(H + FA)(H + M)}{H + FA + M + CN} \tag{3}$$

This metric has been widely used to validate precipitation forecasts in the literature, including those produced by AI-based models[20]. In Fig. 2, both the prediction and observation fields are first reshaped into vectors before calculating the error. Furthermore, we employ percentile-based thresholds, meaning that the threshold for defining an event varies across grid points and corresponds to a specific percentile of the local (wet-day) precipitation distribution. This allows for a more regionally adaptive evaluation of the model's performance.

3. The FSS measures the spatial agreement between forecasted and observed fields by comparing, at each grid point i, the fraction of neighboring points within a radius r that exceed a predefined threshold. Let f_i and o_i denote the fraction of points within the neighborhood around i exceeding the threshold in the forecast and observation, respectively. The FSS is computed as:

$$FSS = \frac{1}{D} \Sigma_{d=1}^{D} \left(1 - \frac{\Sigma(f_i - o_i)^2}{\Sigma(f_i^2 + o_i^2)} \right) \tag{4}$$

where N is the total number of grid points. FSS values range from 0 (no skill) to (perfect agreement), with higher values indicating better spatial consistency between forecast and observation. An FSS of 0.5 is typically considered to be the threshold of a useful forecast[57].

4. Radially averaged power spectral density (RAPSD): This metric is used to quantify the distribution of variance (or "power") and spatial complexity in a two-dimensional spatial field across different spatial scales. It provides a compact, one-dimensional summary of how much

structural complexity or variability exists at different spatial frequencies. To compute the RAPSD:

(a) The 2D Fourier transform of the spatial field is calculated, converting it from physical space (e.g., a precipitation map) to frequency space.
(b) The power spectrum is then computed by taking the squared magnitude of the complex Fourier coefficients, which corresponds to multiplying each coefficient by its complex conjugate.
(c) This 2D power spectrum is radially averaged over concentric circular frequency bands (annuli) centered on the origin of the frequency domain. This step yields a 1D vector representing the mean power as a function of radial spatial frequency.
(d) Finally, the RAPSD is averaged across multiple samples (i.e., different simulations).

Data availability

The datasets generated during the current study are available in a companion repository published by the University of California, San Diego Library (https://doi.org/10.6075/J0H70G60)[58].

Code availability

The underlying code for this study is available in a GitHub repository and can be accessed via this link: https://github.com/CW3E/regional-ai-6km.

Received: 9 July 2025; Accepted: 29 October 2025;
Published online: 12 December 2025

References

1. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
2. Stensrud, D. J. *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models* (Cambridge Univ. Press, 2007).
3. Doyle, J. D. et al. An intercomparison of model-predicted wave breaking for the 11 january 1972 boulder windstorm. *Monthly Weather Rev.* **128**, 901–914 (2000).
4. Mass, C. F., Ovens, D., Westrick, K. & Colle, B. A. Does increasing horizontal resolution produce more skillful forecasts? the results of two years of real-time numerical weather prediction over the pacific northwest. *Bull. Am. Meteorol. Soc.* **83**, 407–430 (2002).
5. Bukovsky, M. S. & Karoly, D. J. Precipitation simulations using wrf as a nested regional climate model. *J. Appl. Meteorol. Climatol.* **48**, 2152–2159 (2009).
6. Bauer, H.-S. et al. Quantitative precipitation estimation based on high-resolution numerical weather prediction and data assimilation with wrf–a performance test. *Tellus A Dyn. Meteorol. Oceanogr.* **67**, 25047 (2015).
7. Martin, A. et al. Evaluation of atmospheric river predictions by the wrf model using aircraft and regional mesonet observations of orographic precipitation and its forcing. *J. Hydrometeorol.* **19**, 1097–1113 (2018).
8. Delle Monache, L. et al. Cw3e's west-wrf 200-member ensemble. *Monthly Weather Rev.* **153**, 1401–1429 (2025).
9. Bonev, B. et al. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International Conference on Machine Learning* 2806–2823 (PMLR, 2023).
10. Bi, K. et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**, 533–538 (2023).
11. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
12. Lang, S. et al. Aifs–ecmwf's data-driven forecasting system. Preprint at arXiv:2406.01465 (2024).
13. Hersbach, H. et al. The era5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
14. Ben Bouallegue, Z. et al. The rise of data-driven weather forecasting: a first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bull. Am. Meteorol. Soc.* **105**, E864–E883 (2024).
15. Liu, C.-C. et al. Evaluation of five global ai models for predicting weather in eastern Asia and western Pacific. *npj Clim. Atmos. Sci.* **7**, 221 (2024).
16. Radford, J. T., Ebert-Uphoff, I. & Stewart, J. Q. A comparison of ai weather prediction and numerical weather prediction models for 1–7-day precipitation forecasts. *Weather Forecast.* **40**, 561–575 (2025).
17. Baño-Medina, J., Manzanas, R. & Gutiérrez, J. M. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geosci. Model Dev.* **13**, 2109–2124 (2020).
18. Mardani, M. et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Commun. Earth Environ.* **6**, 124 (2025).
19. Adamov, S. et al. Building machine learning limited area models: kilometer-scale weather forecasting in realistic settings. Preprint at arXiv:2504.09340 (2025).
20. Nipen, T. N. et al. Regional data-driven weather modeling with a global stretched-grid. Preprint at arXiv:2409.02891 (2024).
21. Lang, S. et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. Preprint at arXiv:2412.15832 (2024).
22. Ralph, F. M., Dettinger, M. D., Schick, L. J. & Anderson, M. L. in *Atmospheric Rivers* (eds Ralph, F. M, Dettinger, M. D, Schick, L. J. & Anderson, M. L.). Ch. 1, pp. 1–13 (Springer, 2020).
23. Dettinger, M. & Cayan, D. R. Drought and the California delta—a matter of extremes. *San Franc. Estuary and Watershed Sci.* **12,** 1–6 (2014).
24. Hu, W. et al. Deep learning forecast uncertainty for precipitation over the western united states. *Monthly Weather Rev.* **151**, 1367–1385 (2023).
25. Pathak, J. et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. Preprint at arXiv:2202.11214 (2022).
26. Baño-Medina, J., Sengupta, A., Watson-Parris, D., Hu, W. & Delle Monache, L. Toward calibrated ensembles of neural weather model forecasts. *J. Adv. Model. Earth Syst.* **17**, e2024MS004734 (2025).
27. Baño-Medina, J. et al. Harnessing ai data-driven global weather models for climate attribution: an analysis of the 2017 oroville dam extreme atmospheric river. *Artif. Intell. Earth Syst.* **4**, 240090 (2025).
28. Sengupta, A. et al. Role of evolving sea surface temperature modes of variability in improving seasonal precipitation forecasts. *Commun. Earth Environ.* **6**, 256 (2025).
29. Gibson, P. B. et al. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun. Earth Environ.* **2**, 159 (2021).
30. Daly, C., Neilson, R. P. & Phillips, D. L. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol. Climatol.* **33**, 140–158 (1994).
31. Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A. & Speranza, A. Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather Forecast.* **18**, 918–932 (2003).
32. Ghazvinian, M. et al. Deep learning of a 200-member ensemble with a limited historical training to improve the prediction of extreme precipitation events. *Monthly Weather Rev.* **152**, 1587–1605 (2024).
33. Cannon, F., Hecht, C. W., Cordeira, J. M. & Ralph, F. M. Synoptic and mesoscale forcing of southern california extreme precipitation. *J. Geophys. Res. Atmos.* **123**, 13–714 (2018).
34. Lavers, D. A., Simmons, A., Vamborg, F. & Rodwell, M. J. An evaluation of era5 precipitation for climate monitoring. *Q. J. R. Meteorol. Soc.* **148**, 3152–3165 (2022).
35. Brenowitz, N. D. et al. A practical probabilistic benchmark for ai weather models. *Geophys. Res. Lett.* **52**, e2024GL113656 (2025).
36. Xu, H., Duan, Y. & Xu, X. Exploring the integration of a global ai model with traditional data assimilation in weather forecasting. *Environ. Res. Lett.* **19**, 124079 (2024).

37. Baño-Medina, J. et al. Are AI weather models learning atmospheric physics? a sensitivity analysis of cyclone xynthia. *npj Clim. Atmos. Sci.* **8**, 92 (2025).
38. Reichstein, M. et al. Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019).
39. Camps-Valls, G. et al. Artificial intelligence for modeling and understanding extreme weather and climate events. *Nat. Commun.* **16**, 1919 (2025).
40. Vonich, P. T. & Hakim, G. J. Predictability limit of the 2021 pacific northwest heatwave from deep-learning sensitivity analysis. *Geophys. Res. Lett.* **51**, e2024GL110651 (2024).
41. Ralph, F. M. et al. West coast forecast challenges and development of atmospheric river reconnaissance. *Bull. Am. Meteorol. Soc.* **101**, E1357–E1377 (2020).
42. Espinoza, V., Waliser, D. E., Guan, B., Lavers, D. A. & Ralph, F. M. Global analysis of climate change projection effects on atmospheric rivers. *Geophys. Res. Lett.* **45**, 4299–4308 (2018).
43. Baek, S. H. & Lora, J. M. Counterbalancing influences of aerosols and greenhouse gases on atmospheric rivers. *Nat. Clim. Change* **11**, 958–965 (2021).
44. Rackow, T. et al. Robustness of AI-based weather forecasts in a changing climate. Preprint at arXiv:2409.18529 (2024).
45. Sun, Y. Q. et al. Can ai weather models predict out-of-distribution gray swan tropical cyclones? *Proc. Natl Acad. Sci. USA* **122**, e2420914122 (2025).
46. Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
47. Pathak, J. et al. Kilometer-scale convection allowing model emulation using generative diffusion modeling. Preprint at arXiv:2408.10958 (2024).
48. Addison, H. et al. Machine learning emulation of precipitation from km-scale regional climate simulations using a diffusion model. Preprint at arXiv:2407.14158 (2024).
49. Larsson, E. et al. Diffusion-lam: probabilistic limited area weather forecasting with diffusion. Preprint at arXiv:2502.07532 (2025).
50. Scher, S. & Messori, G. Ensemble methods for neural network-based weather forecasts. *J. Adv. Model. Earth Syst.* **13** (2021).
51. Keisler, R. Forecasting global weather with graph neural networks. Preprint at arXiv:2202.07575 (2022).
52. Powers, J. G. et al. The weather research and forecasting model: Overview, system efforts, and future directions. *Bull. Am. Meteorol. Soc.* **98**, 1717–1737 (2017).
53. Skamarock, W. C. et al. A description of the advanced research wrf model version 4.3. Tech. Rep. NCAR/TN-556+STR, NCAR (2019).
54. Vandal, T. et al. Deepsd: generating high resolution climate change projections through single image super-resolution. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2017).
55. Sha, Y. et al. Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. part ii: daily precipitation. *J. Appl. Meteorol. Climatol.* **59**, 2075–2092 (2020).
56. Ralph, F. M. et al. A scale to characterize the strength and impacts of atmospheric rivers. *Bull. Am. Meteorol. Soc.* **100**, 269–289 (2019).
57. Roberts, N. M. & Lean, H. W. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Rev.* **136**, 78–97 (2008).
58. Baño-Medina, J. et al. A regional high resolution AI weather model for the prediction of atmospheric rivers and extreme precipitation. *UC. San. Diego. Lib. Digit. Coll.* https://doi.org/10.21203/rs.3.rs-7087242/v1 (2025).

Acknowledgements
This research is supported by the Office of Naval Research (ONR) (Award Number N000142412731), the California Department of Water Resources Atmospheric River Program Phase V (Grant 4600015671), and the US Army Corps of Engineers (USACE) Forecast Informed Reservoir Operations Phase 3 (USACE W912HZ-24-2-0001). The authors gratefully acknowledge the computational resources provided by the National Artificial Intelligence Research Resource Pilot (NAIRR Award 240367) and the DeltaAI advanced computing and data resource, which is supported by the National Science Foundation (award NSF-OAC 2320345). The authors thank Matthew Chantry and Baudouin Raoult from ECMWF for their helpful suggestions on implementing the models within the Anemoi framework.

Author contributions
J.B.-M., A.S., and L.D.M. conceived the idea of the study. D.S. and P.M. generated the high-resolution reanalysis data and the West-WRF NRT model output. J.B.-M. trained the AI models, performed inference, and validated the results. T.N. and M.S.-C. provided feedback on the design of the AI-driven stretched-grid model and contributed to the Anemoi framework. Y.N. curated the AI-driven forecasts for publication. J.B.-M. and Y.N. were responsible for developing the GitHub repository to ensure code reproducibility. All authors reviewed the manuscript and contributed to the writing.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-025-01265-9.

Correspondence and requests for materials should be addressed to Agniv Sengupta.