https://doi.org/10.5194/egusphere-2026-35
Preprint. Discussion started: 14 January 2026

Relaxation experiments in ML-based weather prediction models to study subseasonal predictability

Siyu Li[1], Juliana Dias[2], Benjamin Moore[3], and Julian Quinting[1,4]

[1]Institute of Meteorology and Climate Research Troposphere Research (IMKTRO), Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Physical Sciences Laboratory, National Oceanic and Atmospheric Administration, Boulder, United States
[3]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, United States
[4]Institute of Geophysics and Meteorology, University of Cologne, Cologne, Germany

Correspondence: Siyu Li (siyu.li@kit.edu)

Abstract. This study explores the use of relaxation experiments in machine learning-based weather prediction (MLWP) models to identify sources of subseasonal predictability in comparison to a traditional numerical weather prediction (NWP) system. Relaxation involves nudging specific regions of a model toward reanalysis data to isolate their influence on forecast skill. We apply this technique to two MLWP models, Pangu-Weather (fully data-driven) and NeuralGCM (hybrid) and compare the experiments to the Unified Forecast System (UFS). The focus is on week 3–4 forecast of two major precipitation events in western North America in winter 2022/2023, both linked to Madden-Julian Oscillation (MJO) activity. For the two cases, MLWP models exhibit higher forecast skill than the UFS at subseasonal lead times. Though tropical relaxation improves the skill in all forecast systems, gains are greater for UFS, reflecting the MLWP models' stronger baseline performance. A Rossby wave source (RWS) analysis shows that tropical relaxation consistently improves the large-scale dynamic processes associated with the tropical-extratropical teleconnections leading to both events. These results highlight the potential of relaxation experiments as a low-cost, effective diagnostic for understanding and improving subseasonal forecasts, especially in emerging MLWP systems.

1 Introduction

Subseasonal-to-seasonal (S2S) forecasts targeting lead times of 2 weeks to 2 months are critical for anticipating extreme weather events, supporting agriculture, and enhancing community resilience (White et al., 2022). Despite steady advances in numerical weather prediction (NWP), reliable predictions beyond two weeks remain limited by intrinsic predictability constraints and systematic model biases (Vitart et al., 2012). Leveraging known sources of subseasonal predictability, including those originating at low latitudes such as the Madden–Julian Oscillation (MJO; Madden and Julian (1972)), is essential for improving forecast skill (Cassou, 2008; Lin et al., 2009; Merryfield et al., 2020). As machine learning-based weather prediction (MLWP) progressively emerges as a powerful tool for predictions, this study aims to investigate tropical sources of subseasonal predictability in MLWP systems.

The growing number of studies demonstrating skillful medium-range MLWP performance suggests that these systems also offer a promising path for advancing S2S forecasting, even as several important challenges remain. For instance, data-driven models like Pangu-Weather (Bi et al., 2022) and hybrid approaches such as NeuralGCM (Kochkov et al., 2024) have demonstrated skill comparable to state-of-the-art NWP models in the medium range. However, MLWP models can be highly sensitive to initial condition perturbations (Vonich and Hakim, 2024; Tian et al., 2024), raising questions about their robustness for longer lead times. While MLWP performance on subseasonal timescales including their sensitivity to teleconnection patterns remain less well understood, recent promising results (e.g. Diao and Barnes, 2025) highlight their potential. Moreover, their relatively low computational costs enable systematic analyses of sources of predictability.

A common diagnostic in NWP for assessing sources of predictability is the relaxation technique, which nudges forecasts toward a reference dataset over specific regions (Jung et al., 2010; Magnusson, 2017). This method has successfully illuminated the role of tropical forecast errors on the extratropical forecast skill and the potential to improve the representation of MJO-related teleconnections (Dias et al., 2021; Vitart and Balmaseda, 2024). Though such experiments can be computationally demanding, they provide valuable insights into error propagation and regional influence on forecast skill. The application of relaxation techniques to MLWP models has not been widely tested. Evaluating whether tropical relaxation improves mid-latitude subseasonal forecasts in MLWP models could inform both efforts related to their physical consistency and future model development (Perkan and Zaplotnik, 2025). Here, we investigate relaxation in MLWP forecasts for two high-impact precipitation events in western North America during winter 2022–2023 that were influenced by MJO activity and La Niña conditions. These events occurred from late December to mid January and late February to early March, respectively, and involved contrasting large-scale circulation patterns. We compare the prediction skill of ensemble forecasts from Pangu-Weather, NeuralGCM, and an experimental version of the Unified Forecast System (UFS) under different relaxation configurations, complementing analyses by Moore et al. (2025).

The primary objectives of this study are threefold. First, we test the general feasibility of applying the relaxation technique to both fully machine learning-based and hybrid weather prediction models. Second, we evaluate the impact of relaxation in these models in comparison to an NWP model, specifically in the context of subseasonal forecasts. Finally, we assess whether correcting tropical forecast errors through relaxation leads to improved mid-latitude forecast skill in MLWP models. The data, models, nudging technique and Rossby wave source diagnostic are introduced in Section 2. Results are presented in Section 3. The study ends with a concluding discussion in Section 4.

2 Data and methodology

2.1 Ensemble design and initialization approach

Pangu-Weather, NeuralGCM, and experimental UFS forecasts are all initialized from the same ensemble of data assimilations (EDA) from the ERA5 reanalysis data set (Hersbach et al., 2020). The EDA includes 10 ensemble members that account for uncertainties in the observations and the underlying model by perturbing model physical tendencies in the short forecasts that link subsequent analysis windows. It contains all atmospheric variables to initialize the models of this study and is available on a regular latitude-longitude grid with 0.5×0.5 °grid spacing. The EDA data is regridded with bilinear interpolation to match each model's grid spacing.

The subseasonal forecasts for the two cases are initialized on 15 December 2022 and 2 February 2023, respectively. The initialization of the subseasonal forecasts is achieved through a time-lagged combination of ensemble members from EDA following the approach of Moore et al. (2025). For example, a forecast initialized on 15 December 2022, 00 UTC incorporates ensemble members from forecasts issued on 14 December 2022, 12 UTC, 15 December 2022, 00 UTC and 15 December 2022, 12 UTC. With 10 ensemble members at each time, this yields a 30 member time-lagged ensemble. This methodology is applied consistently for the February case study.

2.2 MLWP and NWP forecast models

This study evaluates subseasonal prediction skill using three distinct modeling approaches: (1) an experimental version of the National Oceanic and Atmospheric Administration (NOAA) UFS, a state-of-the-art NWP model (Jacobs, 2021), (2) NeuralGCM, a hybrid neural network-based general circulation model (Kochkov et al., 2024), and (3) Pangu-Weather, a purely data-driven machine learning model (Bi et al., 2022).

2.2.1 UFS

The UFS is the Earth system modeling framework for current operational NOAA prediction systems, including the Global Forecast System (GFS) and Global Ensemble Forecast System. Experiments were performed with a prototype UFS version (labeled "HR1"). This coupled ocean–atmosphere dynamical model employs the Finite-Volume Cubed-Sphere Dynamical Core (Zhou et al., 2019) and was run globally at C96 resolution (approximately 1° latitude/longitude) with 6-hour forecast increments. The HR1 prototype includes updated GFDL microphysics and other physics packages, and its performance has been shown to be comparable to the operational GFS for large-scale forecasts while improving the representation of precipitation and mesoscale processes. In this study, the UFS model is used as a benchmark to compare the prediction skill of Pangu-Weather and NeuralGCM for the two cases.

2.2.2 NeuralGCM

NeuralGCM is a hybrid machine learning-enhanced general circulation model (Kochkov et al., 2024). The model leverages a differentiable dynamical core for solving the discretized governing dynamical equations as in NWP models and an ML-based physics module that parameterizes per vertical column the effect of unresolved physical processes with a neural network. In this study, subseasonal forecasts and relaxation experiments utilize the 1.4° resolution auto-agressive model with 12-hour interval, which provides output on 37 vertical sigma levels. On seasonal to climate timescales, the NeuralGCM models exhibit robust and stable performance when integrating the 1.4° deterministic configuration for periods of up to approximately two years. The deterministic model setting is used such that ensemble members only diverge because of different initial conditions taken from the EDA of ERA5. Sea surface temperature are prescribed from ERA5.

2.2.3 Pangu-Weather

Pangu-Weather is a fully data-driven deep learning model trained on 39 years (1979–2017) of ERA5 reanalysis data. It operates at a grid spacing of 0.25° with 13 pressure levels (Bi et al., 2022). Pangu-Weather ensemble members are generated using perturbations from the EDA described in section 2.1. Unlike NeuralGCM and UFS which incorporate a dynamical core, Pangu-Weather is fully data-driven and is trained separately for 1 h, 3 h, 6 h, 24 h lead times. Overall the model is autoregressive during inference, but not during training. This leads to its significantly lower computational resource requirements compared to the other two models of this study.

2.3 Verification and climatology data

All forecasts are relaxed and evaluated against ERA5 reanalysis data (Hersbach et al., 2020). This dataset provides reanalysis of atmospheric conditions at 0.25° grid spacing (Hersbach et al., 2020). ERA5 data are remapped using bilinear interpolation to each model's native resolution. Daily climatological means from the ERA5 for 1970—2019 were used to compute anomalies of all atmospheric variables. A 61-day sliding window is applied around each day-of-year and time-of-day combination, with weights that decrease linearly from the center to zero. This approach smooths the climatology by reducing sample noise, though it slightly diminishes the seasonal amplitude. These means, obtained from the WeatherBench2 dataset, were calculated following the method of Rasp et al. (2020). All models generate 30-day ensemble forecast using the method mentioned in the Section 2. For each case, we examine the subseasonal prediction of the large-scale circulation over the North Pacific and western North America at week 3–4 lead time. Forecasts on week 3–4 lead time are daily averaged during the validation periods during 30 December 2022–13 January and 17 February–3 March for the two cases, separately.

2.4 Setup of relaxation experiments

2.4.1 Setup in MLWP models

Relaxation, also referred to as nudging, is an established method in NWP models and has been used in many contexts including the assessment of the role of specific regions for subseasonal predictability (Jung et al., 2010). This approach incorporates an additional term into the model's prognostic equations to gently steer the model state toward reference data thereby constraining the model's evolution within the relaxation domain. In this study, we apply the relaxation to the three-dimensional model state vector $\mathbf{x}_t$ at leadtime t using the following equation

$$\hat{\mathbf{x}}_t = \mathbf{x}_t + \lambda(\mathbf{x}_{\text{ref,t}} - \mathbf{x}_t) \tag{1}$$

where $\mathbf{x}_{\text{ref,t}}$ is the reference data (ERA5 reanalysis) and λ is the relaxation coefficient that determines the strength of the relaxation. After applying the relaxation function, we provide the corrected state vector $\hat{\mathbf{x}}_\mathbf{t}$ to the model and continue the forecast to the next lead time.

Experiment	**Description**	**Abbrev.**
Control	*Model is run freely without relaxing.*	CRL
Wide tropical relaxing	*From 10°S and 10°N are fully nudged to the ERA5 reanalysis, with the degree of relaxing reduced to zero between 10°S/N and 30°S/N.*	WTR
Narrow tropical relaxing	*Full relaxing is restricted to 5°S–5°N, tapering to zero 20°S/N.*	NTR
Replay to ERA5	*Model is nudged to ERA5 globally; serves as the verification dataset as ERA5 reanalysis.*	Replay

Table 1. Description of experimental setups and their abbreviations.

To prevent discontinuities at the boundaries of the relaxed region, a hyperbolic tangent function is applied to create a tapering region (Figure A1). This function modulates the relaxation coefficient λ near the edges, ensuring a gradual change from the nudged to the free-running model areas. The transition function can be formulated as

$$\lambda(\phi) = \lambda_0 \left[0.5 - 0.5 \tanh\left(\frac{\phi - a}{b} \right) \right] \tag{2}$$

where λ_0 is the maximum relaxation coefficient. We take $\lambda_0 = 1.0$, which means that each forecast in the relaxed region is corrected by 100% at each time step. Parallel experiments with $\lambda_0 = 0.33$ (Magnusson, 2017) yield qualitatively similar results (not shown). ϕ denotes the latitude, a is the central point of the transition, and b controls the latitudinal width of the tapering region. This formulation ensures a gradual transition of $\lambda(\phi)$ from λ_0 to 0 at the boundaries.

The four types of relaxation experiments are Control (CRL), narrow tropical relaxation (NTR, relaxing from 20°S to 20°N including the tapering region), wide tropical relaxation (WTR, relaxing from 30°S to 30°N including the tapering region) and model replay (relaxation applied globally). More detailed information are available in Table 1. WTR is designed to assess the overall impact of the entire tropics, whereas NTR focuses more strictly on the deep tropics. In UFS, horizontal wind components, geopotential, specific humidity and temperature are nudged (Table 2). In Pangu-Weather, variables at all 13 pressure levels are nudged during model integration. All surface level variables, such as 2-m temperature, are excluded from relaxation in Pangu-Weather. In NeuralGCM, all variables except geopotential are relaxed along vertical layers between boundary layer and tropopause, including specific cloud ice and liquid water content. The relaxation is applied every 24 hours in both MLWP models.

Table 2. Variables used for relaxation in UFS, NeuralGCM, and Pangu models. In addition, the UFS model applies relaxation to pressure, which is not available in MLWP models.

Variable	Unit	UFS	NeuralGCM	Pangu-Weather
Temperature (T)	K	✓	✓	✓
Zonal wind (U)	$m\ s^{-1}$	✓	✓	✓
Meridional wind (V)	$m\ s^{-1}$	✓	✓	✓
Specific humidity (Q)	$kg\ kg^{-1}$	✓	✓	✓
Geopotential height(Z)	$m^2\ s^{-2}$	✓		✓
Specific cloud ice water content	$kg\ kg^{-1}$		✓	
Specific cloud liquid water content	$kg\ kg^{-1}$		✓	
Vertical levels		127	37	13

Note: "✓" indicates the variable is used for relaxation in the given model.

2.4.2 Setup in UFS

Relaxation experiments in UFS follow the approach of Dias et al. (2021). An Incremental Analysis Update (IAU) is used to reduce shocks by nudging the model toward ERA5 reanalysis. Increments are calculated as differences between 3-hour forecasts and reanalysis data, then applied over a 6-hour forecast window in a repeated "replay" cycle. In the UFS experiments, the λ_0 relaxation coefficient of 1 is used in the specified latitude bands for WTR and NTR experiments.

2.4.3 Rossbywave source analysis

Following Sardeshmukh and Hoskins (1988), the Rossby wave source (RWS) represents the vorticity tendency through divergent outflow in the upper troposphere, primarily driven by tropical convection. The full RWS is defined as the negative divergence of the product of the divergent wind vector and the absolute vorticity, i.e.,

$$RWS = -\nabla \cdot (\mathbf{V}_\chi \zeta) \tag{3}$$

where $\mathbf{V}_\chi = (u_\chi, v_\chi)$ is the divergent wind and ζ is the absolute vorticity. Expanding Eq. (2.4.3) gives

$$RWS = -\left(u_\chi \frac{\partial \zeta}{\partial x} + v_\chi \frac{\partial \zeta}{\partial y} + \zeta \nabla \cdot \mathbf{V}_\chi \right). \tag{4}$$

This diagnostic has been widely used to identify tropical sources of Rossby waves and their downstream propagation patterns (e.g., Hoskins and Karoly, 1996; Seo and Son, 2016; Moore et al., 2025).

https://doi.org/10.5194/egusphere-2026-35
Preprint. Discussion started: 14 January 2026

3 Tropical relaxation experiment results: two case studies

Building on the findings of Moore et al. (2025), we evaluate the MLWP forecast using the latitude-weighted centered anomaly correlation coefficient (ACC; Wilks, 2011) and mean absolute error (MAE) over the eastern North Pacific and western North America (30°–60°N, 170°E–90°W). The two precipitation events occurred outside the training period of both MLWP models.

3.1 Case study 1: December 2022 to January 2023

During the first event, lasting from 26 December 2022 to mid-January 2023, the total rainfall accumulation of ERA5 exceeds 450 mm in California (Fig. 1a). In some regions, the observed accumulated precipitation even reached values up to 1000 mm (DeFlorio et al., 2024). This extreme precipitation event led to at least 21 fatalities and caused property damage estimated between $5-7$ billion dollars (Schubert et al., 2024). The synoptic situation during this two-week period was characterized by a Rossby wave pattern featuring a positive geopotential height anomaly over the subtropical North Pacific, a negative height anomaly over the eastern North Pacific and a positive height anomaly over eastern North America. The anomalous, quasi-stationary upper-level trough over the northeastern Pacific (Fig. 1b) created a prolonged southwesterly flow along the U.S. West Coast. It was associated with enhanced cyclone and atmospheric river (AR) activity that impacted an area extending from California to British Columbia (not shown). During the two-week period, the mean water vapour flux at 850 hPa reached values of 40 $g\,kg^{-1}\,m\,s^{-1}$ at the coastline (green contours in Fig. 1b) favouring the enormous rainfall amounts in California and Oregon. From December 21 to 28, 2022, the MJO progressed through phases 4–5 in the real-time multivariate MJO (RMM; Wheeler and Hendon, 2004) phase space. This earlier MJO activity may have influenced the midlatitude circulation pattern linked to the precipitation event. The two-week period of the event itself co-occurred with an active MJO phase 6–7 from 29 December 2022 to 9 January 2023. Though MJO phases 6–7 are on average followed by a positive geopotential height anomaly over western North America, this event featured a negative geopotential height anomaly illustrating the enormous variability in the extratropical response to the MJO as also documented by Quinting et al. (2024). Recent studies suggest that the Rossby wave pattern was enhanced by the active MJO with convection over the western Pacific, promoting the ridge-trough-ridge tripole extending from the subtropical North Pacific to eastern North America (DeFlorio et al., 2024; Moore et al., 2025).

Forecasts initialized on 15 December 2022 and valid during weeks 3–4 (30 December–13 January) are shown in Figure 2. The CRL experiments across all models fail to adequately capture the dipole pattern of negative and positive geopotential height anomalies over the northern Pacific and eastern North America (Figure 2a–c). This was a forecast bust for operational forecasts from both NOAA and ECMWF (Moore et al., 2025). Notably, the UFS model exhibits the lowest prediction skill for 500-hPa geopotential height, with a regional mean ACC of 0.07 and the highest MAE of 93 m (Figure 2a). Pangu-Weather shows a similar prediction skill with an ACC of 0.07 and MAE of 92 m in the target region (Fig. 2b). The low forecast skill from CRL in Pangu-Weather might be associated with a systematic negative temperature bias (Ben Bouallègue et al., 2024), likely stemming from limitations in its model architecture and training procedure (Ennis et al., 2025). In contrast, NeuralGCM demonstrates a comparative better representation of the large-scale circulation (Figure 2c). The 500-hPa geopotential height trough extends further east and a weak positive gepotential height anomaly exists over eastern North America. This contributes to a higher

https://doi.org/10.5194/egusphere-2026-35
Preprint. Discussion started: 14 January 2026

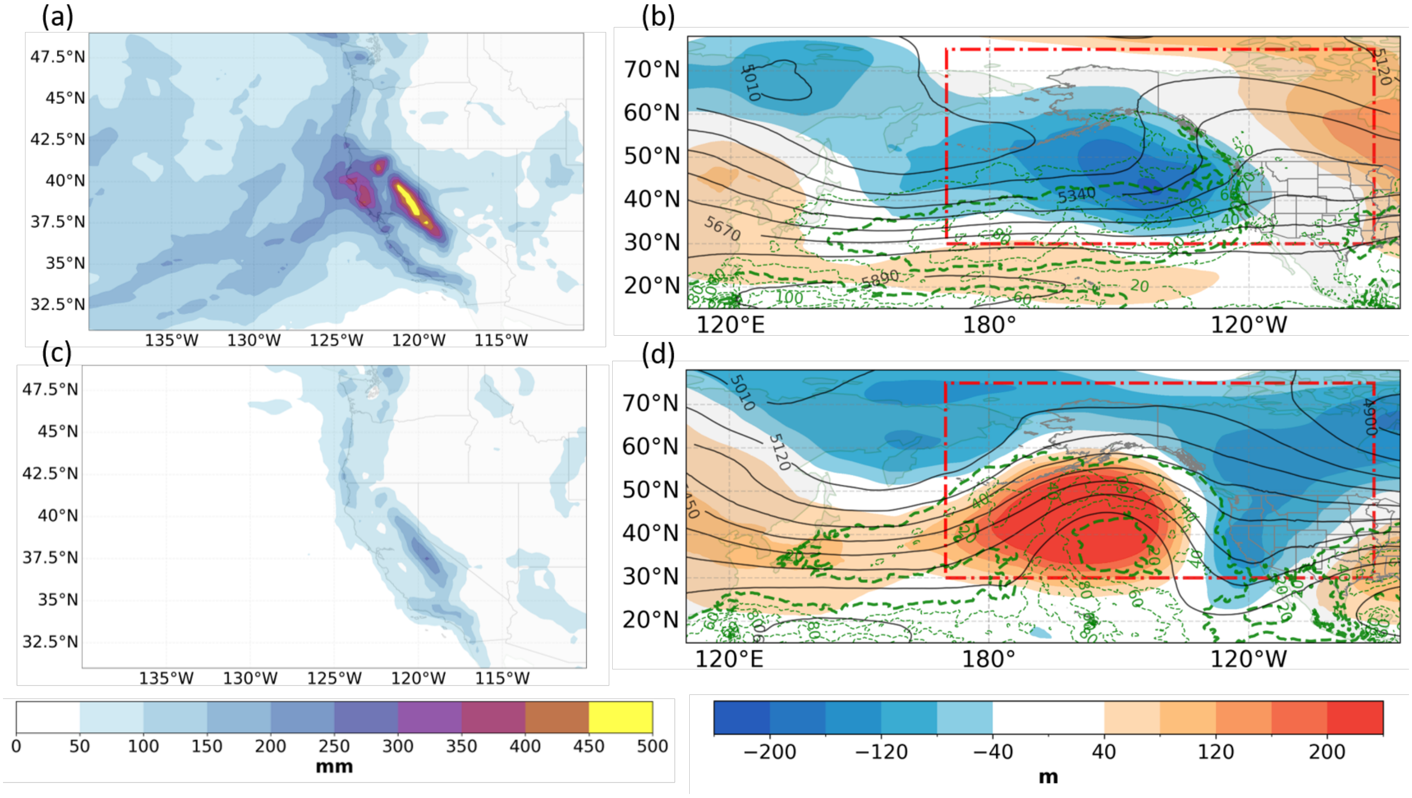

Figure 1. ERA5-based accumulated precipitation (shading in mm) 3–4 weeks after forecast initialization for (a) case study 1 and (c) case study 2 (30 December–13 January for case 1, and 17 February–3 March for case 2). Mean 500-hPa geopotential height anomaly relative to 1970–2019 daily climatology (shading in m), 500-hPa geopotential height (black solid lines in m), and 850-hPa water vapour flux (green dashed lines in $g\,kg^{-1}\,m\,s^{-1}$; 20 and $40\,g\,kg^{-1}\,m\,s^{-1}$ is highlighted in (b) and (d) separately) 3–4 weeks after initialization for (b) case study 1 and (d) case study 2. Red rectangle marks the area for calculating latitude-weighted centered ACCs and MAEs.

subseasonal forecast skill for this case – not only in terms of geopotential height, but also regarding the representation of 850-hPa water vapour flux.

The WTR (Figure 2d–f) and NTR (Figure 2g–i) experiments show marked improvements in reproducing the anomalous 500-hPa geopotential height pattern over the Pacific in all three models. All models better represent the positive geopotential height anomaly over the subtropical North Pacific. Pangu-Weather and NeuralGCM improve the representation of the deep trough over the eastern North Pacific leading to higher ACC values. The presence of this trough is the key distinguishing feature compared to CRL in all three models. The associated enhanced westerly flow leads to a band of high 850-hPa water vapour flux exceeding 40 $g\,kg^{-1}\,m\,s^{-1}$ (highlighted by the bold green dashed line). This moisture transport reaches closer to the Pacific Coast in the WTR and NTR experiments compared to the CRL configuration. Its proximity to the west coast of North America also better matches the verification data (Figure 1b), indicating an improved representation of the precipitation

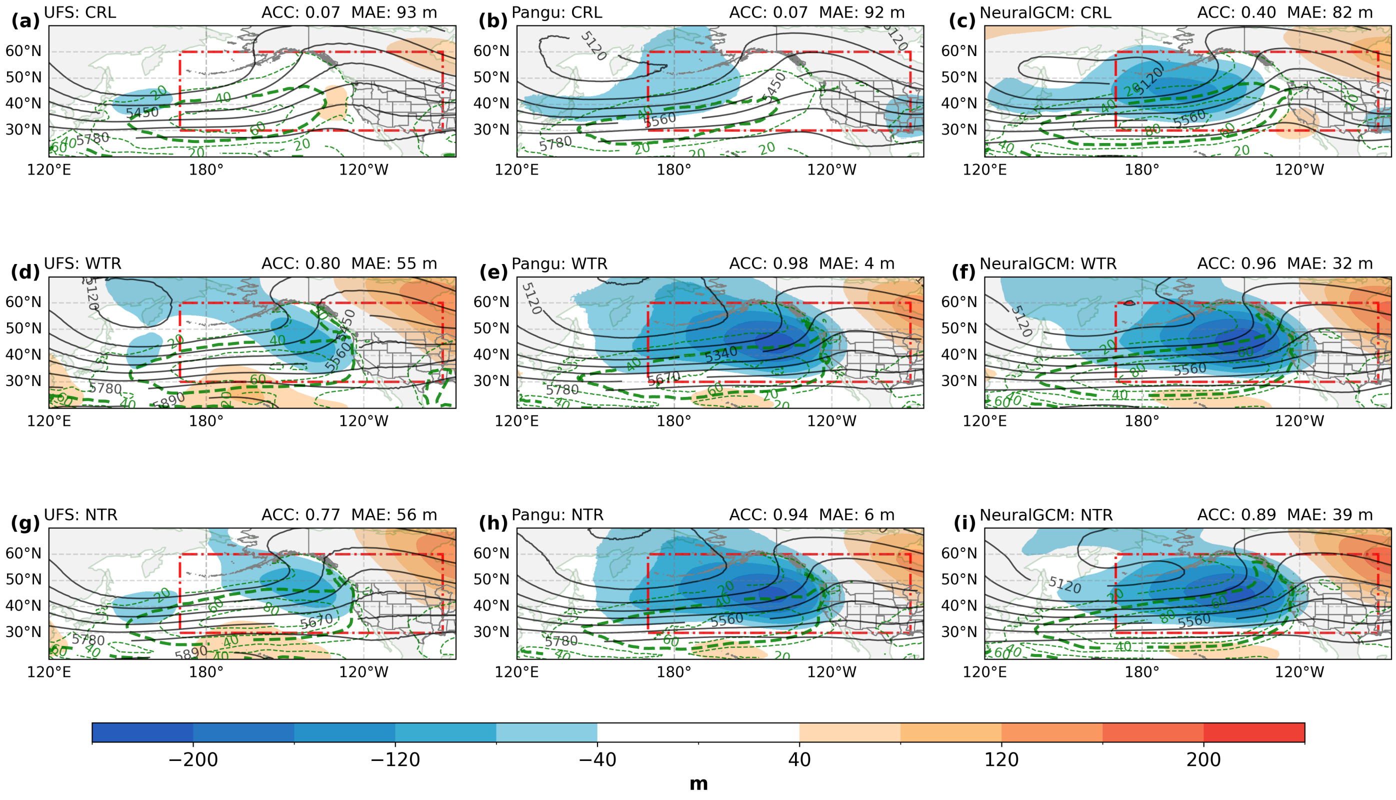
(a) UFS: CRL ACC: 0.07 MAE: 93 m
(b) Pangu: CRL ACC: 0.07 MAE: 92 m
(c) NeuralGCM: CRL ACC: 0.40 MAE: 82 m
(d) UFS: WTR ACC: 0.80 MAE: 55 m
(e) Pangu: WTR ACC: 0.98 MAE: 4 m
(f) NeuralGCM: WTR ACC: 0.96 MAE: 32 m
(g) UFS: NTR ACC: 0.77 MAE: 56 m
(h) Pangu: NTR ACC: 0.94 MAE: 6 m
(i) NeuralGCM: NTR ACC: 0.89 MAE: 39 m
60°N
50°N
40°N
30°N
120°E
180°
120°W
−200
−120
−40
40
120
200
m

Figure 2. Week 3–4 ensemble mean of forecasts initialized on 15 December 2022 showing 500-hPa geopotential height anomaly relative to daily climatology (shading in m), 500-hPa geopotential height (black contours in m), and 850-hPa moisture transport (green dashed lines at intervals of 20 $g\ kg^{-1}\ m\ s^{-1}$, magnitude of 20 $g\ kg^{-1}\ m\ s^{-1}$ is in bold green). The columns show forecasts by (a, d, g) UFS, (b, e, h) Pangu-Weather and (c, f, i) NeuralGCM. The rows show experiments (a, b, c) CRL (d, e, f) WTR and (g, h, i) NTR. Red rectangle denotes the area for calculating the latitude-weighted ACC and MAE from the ensemble mean.

event. The similarity between the WTR and NTR forecasts for all of the models suggests that a better representation of the tropics would have improved the subseasonal forecast skill for this event.

To further understand how the relaxation in the tropics impacts the extratropical Rossby wave forcing, we analyze the RWS (Section 2.4.3) at 200 hPa averaged from 23–30 December 2022 (during week 2; Fig. 3), which is one week earlier than validation periods (week 3 – 4). The focus here is to analyze the establishment of the large-scale flow pattern associated with this extreme precipitation in December in the forecast. Noting that MAEs of RWS are calculated between forecasts and ERA5 over Maritime Continent and western Pacific in the red box (20–38°N, 75–155°E).

For Case 1, UFS CRL predictions consistently overestimate the divergent outflow and the resulting negative vorticity advection to the north of the MJO-related convection over the Maritime Continent and western Pacific, in the days preceding the precipitation events (Moore et al., 2025). Here, results are only shown for the NTR experiments because WTR experiments are

https://doi.org/10.5194/egusphere-2026-35
Preprint. Discussion started: 14 January 2026

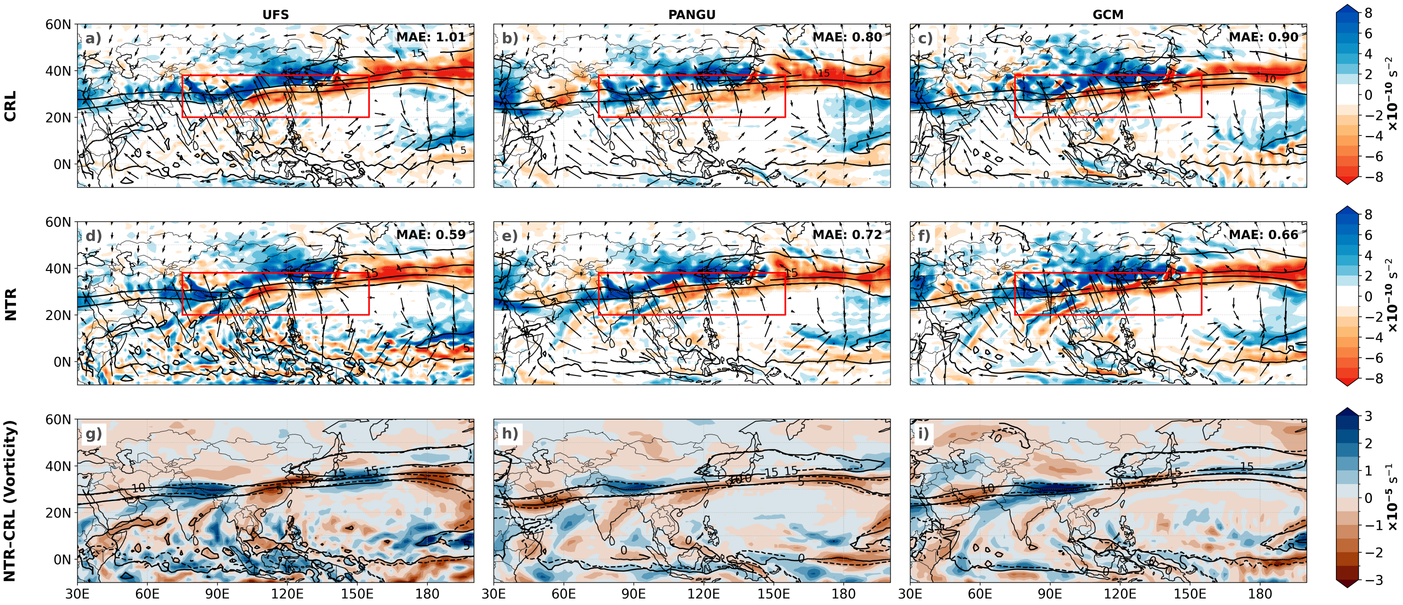

Figure 3. Comparison of ensemble-mean forecasts from CRL (a),(b),(c) and NTR (d),(e),(f) averaged for 23–30 December 2022 for UFS (a,d), Pangu-Weather (b,e), NeuralGCM (c,f). The 200-hPa RWS (10^{-10} s^{-2}, shading), 200-hPa ζ (contours in every 5 interval), 200-hPa divergent wind anomalies. (g),(h),(i), NTR − CTL differences in the 200-hPa ζ (10^{-5} s^{-1}, shading) overlaid by ζ (10^{-5} s^{-1}) contour for NTR (solid) and CRL (dashed) for UFS, Pangu-Weather and NeuralGCM.

qualitatively similar. In NTR and CRL, all models represent the band of negative RWS values over eastern Asia and an area of positive RWS over the western Pacific (Fig. 3a-f). This indicates that the two MLWP models of this study are physically consistent with the UFS model in representing tropical-extratropical teleconnections.

The NTR experiments in UFS exhibits large differences in terms of RWS relative to CRL (Fig. 3a,d) over eastern Asia and the western Pacific. UFS NTR shows an improved Rossby wave source with a smaller forecast error (MAE: 0.59). The negative–positive couplet in this region arises from an eastward displacement and an overestimation of the RWS over the western Pacific in CRL. The related differences in ζ along the waveguide indicate a slight eastward shift of the broad trough–ridge pattern over eastern Asia and the western Pacific (Fig. 3g, Moore et al., 2025).

Pangu-Weather and NeuralGCM CRL exhibit a better prediction of the RWS associated with the divergent outflow with an MAE of 0.80 and 0.90, respectively (Fig. 3b,c). In NTR, RWS is strengthened in the northeastern Indian ocean for both MLWP models (3e, f). Overall, the reduction of the MAE through tropical relaxation is considerably smaller than in UFS (Fig. 3d, e, f). The differences in RWS manifest as dipoles of vorticity differences along the strongest vorticity gradient (Fig. 3k,l), finally affecting moisture transport on week 3 – 4 (Fig. 2h, i) to the west coast of North America.

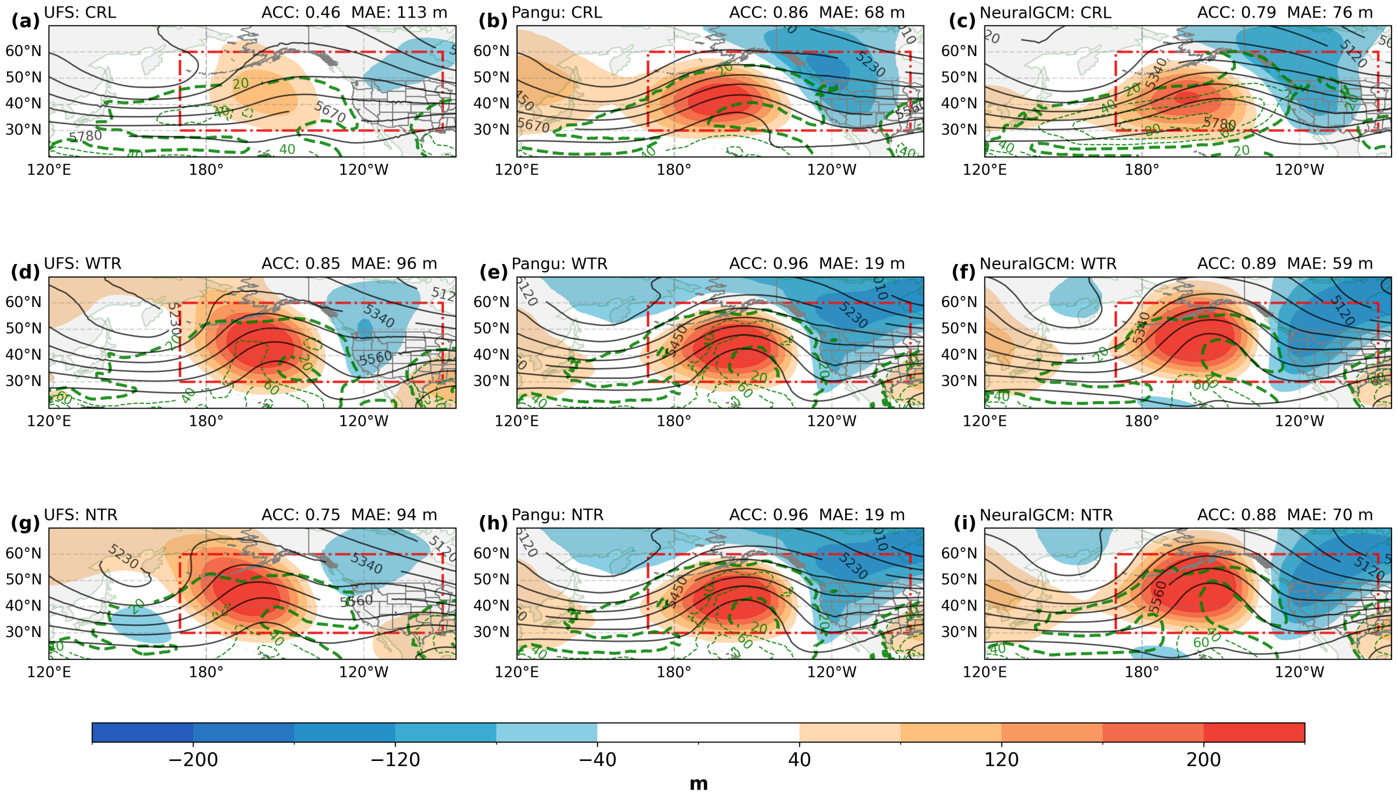
(a) UFS: CRL ACC: 0.46 MAE: 113 m
(b) Pangu: CRL ACC: 0.86 MAE: 68 m
(c) NeuralGCM: CRL ACC: 0.79 MAE: 76 m
(d) UFS: WTR ACC: 0.85 MAE: 96 m
(e) Pangu: WTR ACC: 0.96 MAE: 19 m
(f) NeuralGCM: WTR ACC: 0.89 MAE: 59 m
(g) UFS: NTR ACC: 0.75 MAE: 94 m
(h) Pangu: NTR ACC: 0.96 MAE: 19 m
(i) NeuralGCM: NTR ACC: 0.88 MAE: 70 m
60°N
50°N
40°N
30°N
120°E
180°
120°W
−200
−120
−40
40
120
200
m

Figure 4. Same as Figure 2, but for forecasts initialized on 2 February 2023. The 40 $g\ kg^{-1}\ m\ s^{-1}$ isoline for moisture transport is plotted in bold green.

3.2 Case study 2: February to March 2023

For the second event from mid February to the beginning of March 2023, the ERA5 accumulated precipitation over California reaches approximately 200 mm (Figure 1c). Though the MJO entered simultaneously its active phases 6–7, the midlatitude geopotential height anomalies are very different from case 1. For case 2 (valid from 17 February–3 March), a persistent positive geopotential height anomaly is located over the eastern North Pacific (Figure 1d). ARs are deflected around the associated high pressure anomaly and reach the Pacific Coast in a northwesterly flow. There, the precipitation is produced in connection with severeal upper-level troughs (as manifested by negative anomalies in Fig. 4) on the eastern flank of the Pacific ridge.

All three models exhibit greater subseasonal forecast skill in the CRL experiment (Fig. 4a–c) with higher ACCs and lower MAEs than for Case 1. Pangu-Weather especially depicts the positive geopotential height anomaly over the eastern North Pacific and the surrounding moisture transport, whereas UFS and NeuralGCM underestimate the anomaly amplitude, resulting in lower ACC and higher MAE.

In the UFS model, both the WTR (Fig. 4d) and NTR (Fig. 4g) experiments yield a substantially stronger ridge over the eastern North Pacific, resulting in a considerably improved representation of the large-scale circulation compared to the CRL experiment (Fig. 4a). This finding is consistent with Moore et al. (2025). Pangu-Weather shows modest improvements for the WTR (Fig. 4e) and NTR (Fig. 4h) experiments, particularly in capturing the positive and negative geopotential height anomaly patterns. NeuralGCM predicts a pattern similar to the UFS model, yet with overall higher forecast skill than UFS in this case (cf. Fig. 4a, d, g vs. Fig4 c, f, i). The bands of highest moisture transport around the ridge over the eastern North Pacific, with magnitudes around 40 $g\,kg^{-1}\,m\,s^{-1}$, are consistently well represented. Independent of the relaxation configuration, the forecasts for Pangu-Weather and NeuralGCM with relaxation yield significantly improved representation of the location, amplitude and positive tilt of the trough near the west coast of North America, which may affect precipitation. The UFS forecasts only show a modest improvement regarding the location of the trough.

Overall, the differences in terms of forecast skill between the WTR and NTR are remarkably small across all three models (cf. Fig. 4d–f and Fig. 4g–i. This similarity suggests that forecast skill is not highly sensitive to the width of the tropical nudging region. Nevertheless, the improvements in large positive geopotential height anomalies in the UFS for both WTR and NTR indicate that tropical forecast errors in this model exert a strong influence on predicting the blocking ridge over the eastern North Pacific, even if they did not strongly constrain predictions of downstream wave breaking and trough amplification near the west coast of North America. In the UFS WTR experiment, positive height anomalies are well captured, but negative anomalies near the west coast of North America remain misrepresented. Interestingly, MLWP models better capture the positive geopotential height anomaly over the eastern North Pacific in the CRL without relaxation, possibly due to superior representation of tropical conditions even without nudging.

We also investigate the RWS to assess the impact of the tropical relaxation during week 2 in the Case 2. UFS overestimates RWS over eastern Asia and northern Pacific in CRL (Fig. 5a; Moore et al., 2025) relative to NTR (Fig. 5d). By applying tropical relaxation, RWS is reduced, indicating that the advection of vorticity by the divergent wind is too strong without tropical relaxation.

In the study of Moore et al. (2025), the RWS difference between NTR and CRL in UFS (their Fig. 13c) shows a prominent positive band extending from eastern Asia to the western Pacific. Their findings suggest that vorticity advection by the divergent wind is more adequately represented in NTR and thus the Rossby wave train amplification is enhanced relative to CRL. Here, we show the same evidence as the MAE of RWS decreases from 0.55 to 0.38 when NTR is applied (Fig. 5 a, d). Pangu-Weather and NeuralGCM exhibit a dipole pattern in terms of RWS over Eastern Asia in the CRL experiment with lower MAEs (Fig. 5 b, c). Meanwhile, in the NTR experiments, the RWS in both MLWP models is even slightly deteriorating (Fig. 5 e, f). Higher MAEs in NTR than in CRL are aligned with the hypothesis above that the divergent outflow from the tropics in Case 2 is likely better represented in the MLWP models than in UFS. Finally, noteworthy is an area of intense positive RWS over the eastern Pacific, which is likely associated with enhanced warm conveyor belt activity in this mid-latitude region following MJO phases 6 and 7 (Quinting et al., 2024).

NeuralGCM reaches the lowest RWS MAE in the CRL experiment (0.10). This indicates that the model already captures most of the contributing RWS during the early stage without any nudging. Moreover, the differences in ζ between the CRL

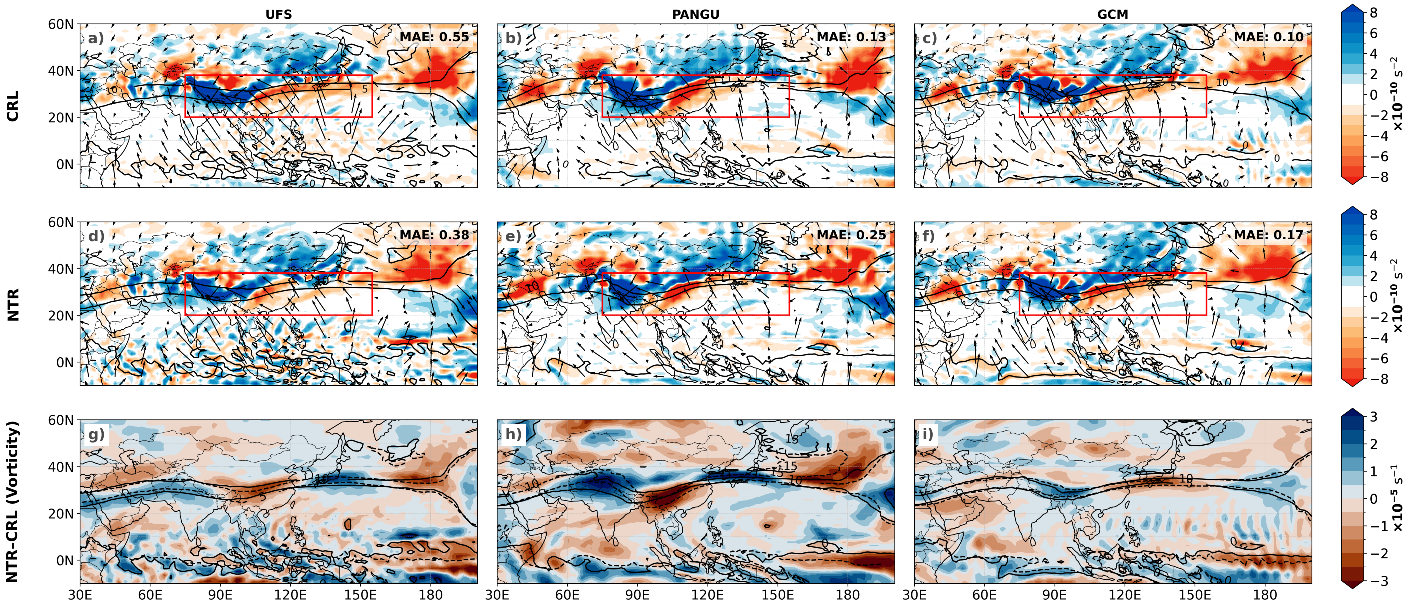

Figure 5. Comparison of ensemble-mean forecasts from CRL (a),(b),(c) and NTR (d),(e),(f) averaged for 9–15 February 2023, same as Figure 3.

and NTR are comparably small (Fig. 5 i), indicating that NeuralGCM's representation of the relevant dynamical fields is less sensitive to the relaxation procedure.

3.3 Synthesis of both cases: forecast uncertainty

Given the different impacts of relaxation in the two cases and the varying contribution of individual members to the ensemble mean, we further examine forecast skill per ensemble member. This provides a more detailed view of forecast performance and model uncertainty across the ensemble for each case.

Both MLWP models demonstrate on average a higher forecast skill in the CRL experiment for the December case compared to UFS (Figs. 6a–c). Still, individual ensemble members in Pangu-Weather and NeuralGCM show negative ACC and thus no forecast skill for this particular event. The application of tropical relaxation (WTR) in December leads to a clear improvement over the control experiment, suggesting a strong influence of tropical forecast errors in all three models on the mid-latitude prediction skill. Also notable is the much smaller range of ACC values between the different ensemble members when tropical relaxation is applied. This suggests an increased confidence in the predictions through tropical relaxation.

In contrast to the December case, the February case shows less sensitivity to tropical relaxation (Figs. 6 d–f) and is better captured by all three models in CRL, especially for the MLWP models. Though a marked improvement in terms of ACC can be seen for UFS, the median ACC of Pangu-Weather and NeuralGCM increases only marginally when tropical relaxation is applied. Also, the smaller range of ACC values between the different ensemble members suggest a higher confidence in the

https://doi.org/10.5194/egusphere-2026-35
Preprint. Discussion started: 14 January 2026

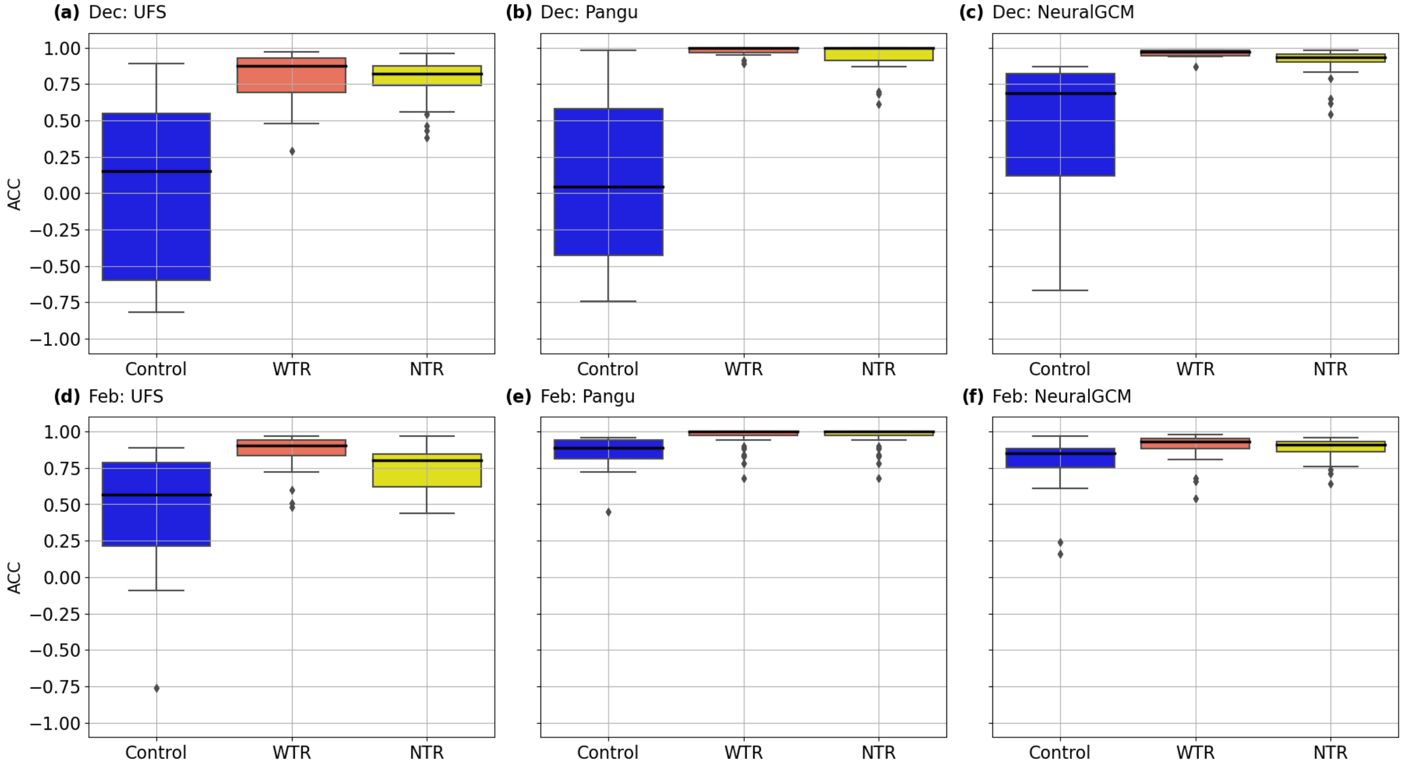

Figure 6. Distribution of the 500-hPa geopotential height latitude weighted centered ACC values for all 30 ensemble members in different relaxation experiments for weeks 3–4. The horizontal line denotes the median, boxes give the 25th to 75th percentile range, whiskers denote the smallest and largest values within 1.5 times the interquartile range, and outliers are given by black dots. Results are shown for the (a),(d) UFS, (b),(e) Pangu-Weather, and (c),(f) NeuralGCM for (top) the December case and (bottom) the February case.

predictions of the two MLWP models. The rather small improvement through tropical relaxation compared to the December case further suggests that tropical forecast errors are less critical for the predictability of the February precipitation event. Since the MLWP models do not predict precipitation, we did not investigate its corresponding predictability directly. Overall, our result indicates that other regions provided predictability for the February case, which could be further investigated with additional nudging experiments in a future study. Such experiments are particularly well suited for MLWP models, because they can produce forecasts through a direct mapping from an initial state to a target lead time. As a result, localized relaxation can be introduced without or less critically on inducing dynamical adjustment processes or transient responses of MLWP models. In contrast, localized relaxation is more challenging in traditional NWP systems, where perturbations introduced in a limited region can interact with upstream disturbances through advection and wave propagation, leading to nonphysical interactions that complicate causal interpretation.

4 Conclusion

This study evaluates the impact of tropical relaxation in UFS, Pangu-Weather and NeuralGCM for two case studies associated with AR landfall in western North America following MJO phases 6–7. Our three central findings are the following.

- The z500 forecast skill of the CRL experiment with Pangu-Weather and NeuralGCM exceeds that of the UFS. These findings underscore the promise of data-driven models in subseasonal forecasting, particularly given their lower computational costs. However, drawing more definitive conclusions will require a systematic evaluation over multiple years and similar events to assess the generalization of these results.
- Relaxation experiments on the subseasonal timescale can be stably conducted in MLWP models, at considerably reduced computational costs in comparison to NWP models. The reference experiments with a traditional NWP models prove useful to establish the necessary confidence in the MLWP relaxation approach at subseasonal scales. Relaxing tropical fields improves forecast skill in MLWP models as in the NWP model. For example, in the December case, tropical relaxation corrects the moisture transport towards western North America in all three models. This suggests that a better representation of the tropical atmospheric state in the models would have improved the prediction of this particular event. Further consistent behaviors are the reduction of the range of ACC values between the different ensemble members and a better representation of Rossby wave source one week earlier. This suggests that the MLWP models used here follow a physically consistent way in generating the Rossby wave.
- The impact of tropical relaxation on mid-latitude forecasts varies between cases. In December, forecasts improve substantially in all three models, suggesting that key tropical processes driving the teleconnection are poorly captured. In February, improvements are smaller, likely due to a combination of better tropical representation in the control runs and an over reduced tropical influence on the event, as also noted by Moore et al. (2025).

In general, the higher forecast skill in NeuralGCM and Pangu-Weather compared to UFS suggests that the NWP model does not fully exploit the predictability inherent in these two events. To identify which relaxation configurations most strongly affect forecast skill will help understand these mechanisms, ultimately guiding targeted improvements in future forecasting systems. Improving the representation of the tropics will likely enhance extratropical prediction skill for similar cases, although a systematic analysis is needed in specific regions to identify where tropical improvements yield the greatest benefit. To translate this insight into improved forecast skill, future research should diagnose the origin of large-scale anomalies, particularly, the pathways through which tropical variability influences extratropical circulation, and assess how predictable these processes are in MLWP models. Such targeted relaxation experiments could also guide MLWP and NWP development by revealing which regions or processes enhance forecast skill most significantly.

Code and data availability. ERA5 reanalysis data are available from ECMWF via Copernicus Climate Change Service, Climate Data Store, 315 (2023). The MLWP models used in this manuscript are available via Rasp et al.(2023). UFS model is available at https://doi.org/10.5281/zenodo.17109573

Appendix A

A1

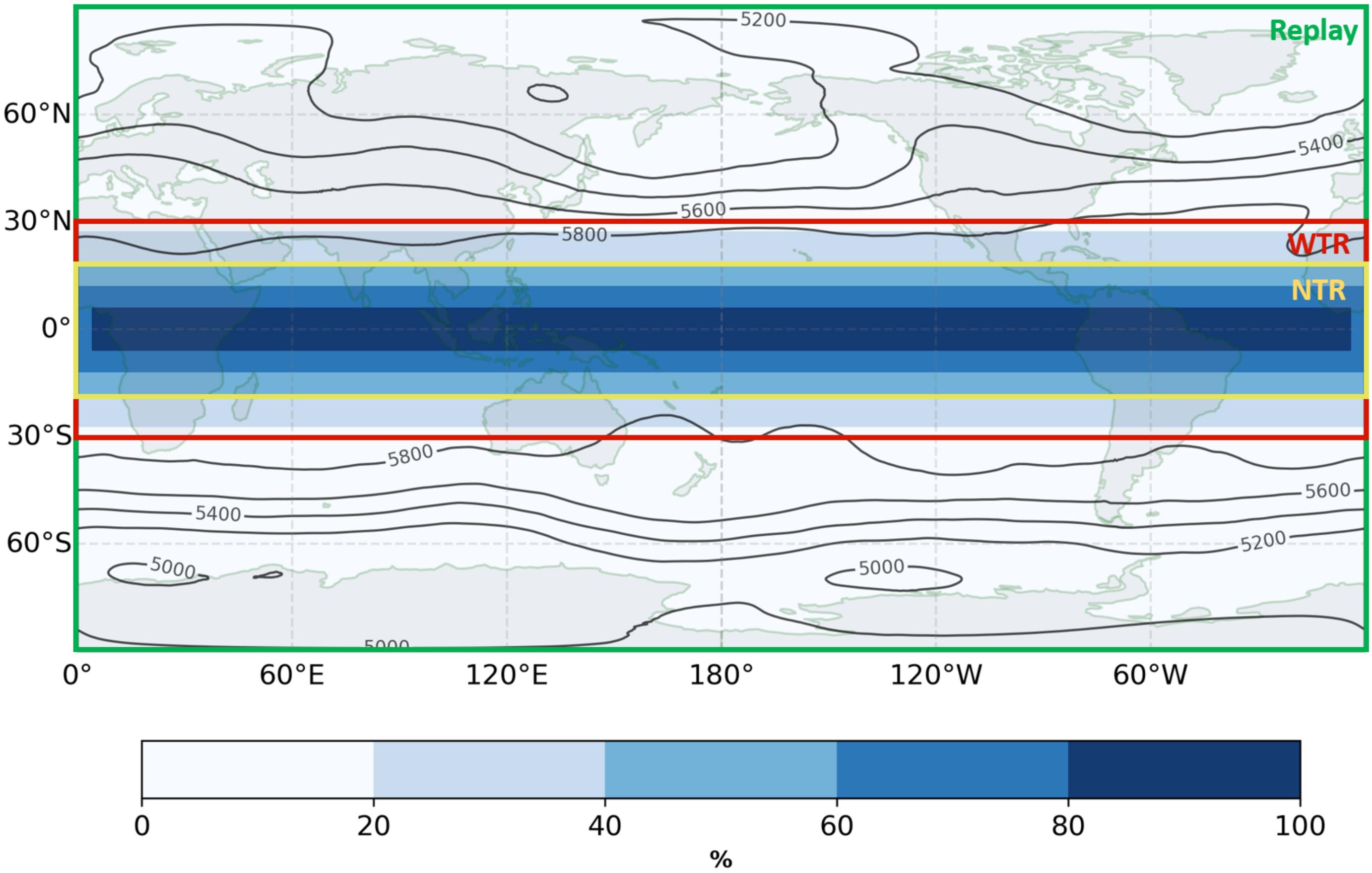

Figure A1. Visualization of relaxing regions during the forecasts

Author contributions. JQ and SL designed the study. SL performed the experiments using Pangu-Weather and NeuralGCM. BM and JD conducted the UFS experiments and provided the corresponding data. SL produced the figures and drafted the manuscript. All authors discussed the results and edited the manuscript.

Competing interests. The authors declare no conflicts of interest relevant to this study.

Acknowledgements. The contribution of SL and JQ was funded by the European Union (ERC, ASPIRE, 101077260). The contribution of JD and BM was supported by the NOAA Physical Sciences Laboratory. We thank Stefan Tulich and Maria Gehne (CIRES/NOAA PSL) for generating the UFS experiments. We thank ECMWF for providing ERA5 reanalysis data. We thank HuaWei and research team of NeuralGCM for sharing MLWP models to the public for research application. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

Ben Bouallègue, Z., Clare, M. C., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T., et al.: The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context, Bulletin of the American Meteorological Society, 105, E864–E883, 2024.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, arXiv preprint arXiv:2211.02556, 2022.

Cassou, C.: Intraseasonal interaction between the Madden–Julian oscillation and the North Atlantic Oscillation, Nature, 455, 523–527, 2008.

Copernicus Climate Change Service: ERA5 hourly data on single levels from 1940 to present, https://doi.org/10.24381/cds.adbb2d47, 2023.

DeFlorio, M. J., Sengupta, A., Castellano, C. M., Wang, J., Zhang, Z., Gershunov, A., Guirguis, K., Luna Niño, R., Clemesha, R. E., Pan, M., et al.: From California's extreme drought to major flooding: evaluating and synthesizing experimental seasonal and subseasonal forecasts of landfalling atmospheric rivers and extreme precipitation during winter 2022/23, Bulletin of the American Meteorological Society, 105, E84–E104, 2024.

Diao, M. T. and Barnes, E. A.: Assessing MJO Tropical-Extratropical Teleconnections in Deep Learning Weather Prediction Models, ESS Open Archive preprint, https://essopenarchive.org/users/631510/articles/1296512-assessing-mjo-tropical-extratropical-teleconnections-in-deep-learning-weather-prediction-models, preprint; submitted / in review, 2025.

Dias, J., Tulich, S. N., Gehne, M., and Kiladis, G. N.: Tropical origins of weeks 2–4 forecast errors during the Northern Hemisphere cool season, Monthly Weather Review, 149, 2975–2991, 2021.

Ennis, K. E., Barnes, E. A., Arcodia, M. C., Fernandez, M. A., and Maloney, E. D.: Turning Up the Heat: Assessing 2-m Temperature Forecast Errors in AI Weather Prediction Models During Heat Waves, arXiv preprint arXiv:2504.21195, 2025.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Doerffer, R., Di Natale, R., Dragani, R., Fuentes, M., Geer, A., Hólm, E. V., Janisková, M., Kaiser, J., Laloyaux, P., Lopez, P., Manrique-Suñén, A., Peubey, C., Radiu, I., Rebetez, O., Thépaut, J.-N., Vitart, F., and De Presanna, P.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hoskins, B. J. and Karoly, D. J.: Teleconnections in the atmosphere and oceans, Journal of Climate, 9, 1049–1072, 1996.

Jacobs, N. A.: Open innovation and the case for community model development, Bulletin of the American Meteorological Society, 102, E2002–E2011, 2021.

Jung, T., Miller, M., and Palmer, T.: Diagnosing the origin of extended-range forecast errors, Monthly Weather Review, 138, 2434–2446, 2010.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al.: Neural general circulation models for weather and climate, Nature, 632, 1060–1066, 2024.

Lin, H., Brunet, G., and Derome, J.: An observed connection between the North Atlantic Oscillation and the Madden–Julian oscillation, Journal of Climate, 22, 364–380, 2009.

Madden, R. A. and Julian, P. R.: Description of global-scale circulation cells in the tropics with a 40–50 day period, Journal of Atmospheric Sciences, 29, 1109–1123, 1972.

Magnusson, L.: Diagnostic methods for understanding the origin of forecast errors, Quarterly Journal of the Royal Meteorological Society, 143, 2129–2142, 2017.

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I., et al.: Current and emerging developments in subseasonal to decadal prediction, Bulletin of the American Meteorological Society, 101, E869–E896, 2020.

Moore, B., Dias, J., Hoell, A., Tulich, S., Gehne, M., Albers, J., Baggett, C., and Lajoie, E.: Impacts of tropical forecast errors on weeks 3-4 extreme precipitation predictions over California during winter 2022-23, Authorea Preprints, https://doi.org/10.22541/essoar.175259860.04608429/v1, 2025.

Perkan, U. and Zaplotnik, Z.: Using gridpoint relaxation for forecast error diagnostics in neural weather models, arXiv preprint arXiv:2506.11987, 2025.

Quinting, J. F., Grams, C. M., Chang, E. K.-M., Pfahl, S., and Wernli, H.: Warm conveyor belt activity over the Pacific: modulation by the Madden–Julian Oscillation and impact on tropical–extratropical teleconnections, Weather and Climate Dynamics, 5, 65–85, 2024.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: a benchmark data set for data-driven weather forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, 2020.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Lopez-Gomez, I., and Yang, V. X.: google-research/weatherbench2: v0.2.0, https://doi.org/10.5281/zenodo.11376271, 2023.

Sardeshmukh, P. D. and Hoskins, B. J.: The generation of global rotational flow by steady idealized tropical divergence, Journal of the Atmospheric Sciences, 45, 1228–1251, https://doi.org/10.1175/1520-0469(1988)045<1228:TGOGRF>2.0.CO;2, 1988.

Schubert, S. D., Chang, Y., DeAngelis, A. M., Lim, Y.-K., Thomas, N. P., Koster, R. D., Bosilovich, M. G., Molod, A. M., Collow, A., and Dezfuli, A.: Insights into the Causes and Predictability of the 2022/23 California Flooding, Journal of Climate, 37, 3613–3629, 2024.

Seo, K.-H. and Son, S.-W.: Rossby wave source of the Pacific–North American teleconnection pattern and its seasonality, Journal of Climate, 29, 548–564, 2016.

Tian, X., Holdaway, D., and Kleist, D.: Exploring the use of machine learning weather models in data assimilation, arXiv preprint arXiv:2411.14677, 2024.

Vitart, F. and Balmaseda, M. A.: Sources of MJO teleconnection errors in the ECMWF extended-range forecasts, Quarterly Journal of the Royal Meteorological Society, 150, 2028–2044, 2024.

Vitart, F., Robertson, A. W., and Anderson, D. L.: Subseasonal to seasonal prediction project: Bridging the gap between weather and climate, Bulletin of the World Meteorological Organization, 61, 23, 2012.

Vonich, P. T. and Hakim, G. J.: Predictability limit of the 2021 Pacific Northwest heatwave from deep-learning sensitivity analysis, Geophysical Research Letters, 51, e2024GL110 651, 2024.

Wheeler, M. C. and Hendon, H. H.: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction, Monthly weather review, 132, 1917–1932, 2004.

White, C. J., Domeisen, D. I., Acharya, N., Adefisan, E. A., Anderson, M. L., Aura, S., Balogun, A. A., Bertram, D., Bluhm, S., Brayshaw, D. J., et al.: Advances in the application and utility of subseasonal-to-seasonal predictions, Bulletin of the American Meteorological Society, 103, E1448–E1472, 2022.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L. M., Chen, X., and Rees, S. L.: Toward convective-scale prediction within the next generation global prediction system, Bulletin of the American Meteorological Society, 100, 1225–1243, 2019.