Title

Subseasonal Forecasting and MJO Teleconnections in Machine Learning Weather Prediction Models

Journal

Journal of Geophysical Research: Atmospheres, 131(3)

ISSN

2169-897X

Authors

Peings, Yannick
Dong, Cameron
Mahesh, Ankur
et al.

Publication Date

2026-02-16

DOI

10.1029/2025jd044910

Subseasonal forecasting and MJO teleconnections in machine learning weather prediction models

Yannick Peings[1], Cameron Dong[1,2], Ankur Mahesh[3,4], Michael Pritchard[1,5],
William Collins[3,4] and Gudrun Magnusdottir[1]

[1] Department of Earth System Science, University of California Irvine, Irvine, CA, USA
[2] Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA
[3] Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA, USA
[4] Department of Earth and Planetary Science, University of California, Berkeley, USA
[5] NVIDIA Corporation, Santa Clara, CA, USA

Key Points

- Two machine learning weather forecast models exhibit state-of-the-art prediction skill at the subseasonal time scale in the Pacific sector.

- The models equal ECWMF in terms of MJO prediction skill, and they accurately predict the MJO propagation and associated teleconnections.

- The two machine-learning models represent key physical processes for subseasonal prediction, despite being trained for weather forecasting.

Abstract

In recent years, machine-learning (ML) models trained on reanalysis data have rivaled physics-based forecast models in terms of performance skill for global weather forecasting. With increased rollout stability, the question of how these models perform for subseasonal to seasonal (S2S, week 3 to 8) forecasting has emerged. In this study we run a large set of subseasonal hindcasts over 2004-2023 to evaluate two ML weather forecast models at the S2S time scale, SFNO-HENS (Nvidia, fully ML) and NeuralGCM (Google Research, hybrid). Corresponding hindcasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) are used as a baseline for comparison to a physics-based model. Because our focus is on predicting moisture transport over the Western United States between October and March, we evaluate the models' prediction skill for the Madden-Julian Oscillation (MJO) and its associated teleconnections in the North Pacific. We find that both ML models are competitive with the ECWMF model, with comparable skill in predicting the North Pacific large-scale circulation and the MJO at week 3 and beyond. Even though overall the mid-latitude subseasonal prediction skill remains low, the ML models exhibit interesting behavior such as a realistic propagation of the MJO across the Maritime Continent and realistic teleconnections. A SFNO-HENS sensitivity experiment with altered initial conditions in the tropics demonstrates the stability of the model, and it illustrates the capability of ML models to represent important physical processes of the atmosphere at the S2S time scale.

Plain Language summary

Predicting weather patterns and precipitation a few weeks in advance (subseasonal time scale) is of great interest for stakeholders such as water managers in the Southwest United States, where arid conditions prevail. Subseasonal forecasts from traditional weather forecast models exhibit low skill in the region, limiting their applicability. Here we examine whether the recent breakthrough in weather forecasting made with machine learning/artificial intelligence models can translate to improved subseasonal forecasts. Recently-developed machine learning models exhibit comparable skill to a state-of-the-art physics-based model for predicting weather patterns in the North Pacific/North America region, and associated moisture transport. The same applies to their skill in predicting the tropical pattern, the Madden-Julian Oscillation, and its important remote perturbations over the midlatitude East Pacific and Southwest United States. Additionally, a perturbation experiment carried out with one of the machine learning models illustrates their ability to not only predict the evolution of atmospheric fields, but also to learn and represent physical processes such as tropics-extratropics Rossby wave propagation.

1. Introduction

Machine learning (ML), or data-driven, weather forecast models represent a recent breakthrough for global weather prediction (de Burgh-Day and Leeuwenburg 2023). Unlike traditional weather forecasts built upon general circulation models (GCMs) that replicate atmospheric physical processes by solving the fundamental equations of fluid dynamics (physics-based models), ML weather forecast models simply predict the weather from past observations or reanalyses. After a training stage that uses global historical data of atmospheric fields, for example from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5, Hersbach et al. 2020), ML models initialized with reanalysis atmospheric conditions are able to accurately predict the evolution of atmospheric fields over medium-range weather forecast time scales. Although their development is relatively recent, ML models already equal or even outperform physics-based weather forecast models for a diversity of metrics (e.g., Pathak et al. 2022, Lam et al. 2023, Bi et al. 2023, Chen L. et al. 2023, Rasp et al. 2024, Price et al. 2023), not only for deterministic but also probabilistic weather forecasting (Lang et al. 2024, Oskarsson et al. 2024, Mahesh et al. 2025a, Price et al. 2025). One advantage of the ML models is that they are orders-of-magnitude less computationally expensive than physics-based models, which allows for running large to "huge" ensembles (Pathak et al. 2022, Mahesh et al. 2025a). The models are trained using sets of fields from reanalyses to learn the high-frequency evolution of the atmosphere, and they provide output forecast data at the same time scale and spatial resolution as the training data (for instance, same 0.25° spatial resolution and 6-hourly time frequency as ERA5). Given the compelling medium-range prediction skill achieved by the weather ML models, a natural question that arises is whether these models can also be used for climate applications (Eyring et al. 2024) and longer-range prediction, in particular for subseasonal to seasonal (S2S) forecasting (He et al. 2021, Weyn et al. 2021).

Some modern ML global models exhibit stable rollouts over long periods of time, which allows for running long forecasts beyond the weather time scale (Weyn et al. 2021). For instance, Kent et al. (2025) used the Ai2 Climate Emulator version 2 (ACE2) ML model to demonstrate its skill in predicting the seasonal winter North Atlantic Oscillation (NAO) with a few weeks lead times. The Fuxi-S2S ML model has also been shown to outperform one of the best dynamical forecast models in terms of predicting the Madden Julian Oscillation (MJO) (Chen et al. 2024), the main source of subseasonal climate predictability in the tropics. The MJO is an intermittent wave packet of enhanced tropical convection that transits west to east through the entire tropics with a period of 30 to 60 days, with repercussions in the mid and high latitudes through climate teleconnections and the propagation of tropics-extratropics Rossby waves (Madden and Julian 1971, Zhang 2013). In particular, the MJO impacts precipitation and temperature extremes over the western United States (US) (e.g., Wang et al. 2023, Wang et al. 2024). Accurate predictions of the MJO are critical for S2S forecasting of the mid-latitude climate, in particular over the region of interest for this study, the eastern North Pacific and western North America (referred to as the Eastern North Pacific-Western North America sector hereafter, or ENP-WNA) (e.g., Peings et al. 2022). In addition to a skillful MJO prediction, anticipating the persistence of certain weather patterns over the ENP-WNA a few weeks in advance necessitates accurate representation of MJO teleconnection mechanisms in the forecast models (Vitart et al. 2017). Physics-based models generally struggle to accurately capture MJO teleconnections as in observations, in particular due to biases in their mean state (Lim et al. 2018, Kim et al. 2019, Xiang et al. 2022). Models are initialized with our best knowledge of the observed climate state at one given time, but as the forecast evolves further from initialization, variability that is inherent to the model develops and model biases emerge. This hampers their long-range prediction skill since it relies on specific climate teleconnections that will be less accurate in the forecast as errors in the atmospheric background flow develop. For instance, Gong et al. (2023) illustrated how slight differences in the North Pacific zonal flow modify the propagation of Rossby waves generated in the vicinity of the east Asia subtropical jet, resulting in vastly different precipitation patterns over western North America. In addition, the MJO predictability and its associated impacts differ depending on the type of MJO propagation, which has been classified into four categories (stand, jump, fast, and slow propagation; Wang et al. 2019). Higher predictability is achieved for the fast-propagating MJO events than for slow-propagating or standing MJO events (Xiang et al. 2022, Abhik et al. 2023), so that certain periods of time are more favorable to skillful S2S forecasts than others. Such periods are called "windows of opportunity for prediction" (Mariotti et al. 2020), whose identification before a forecast is produced is a remaining goal for climate forecasters (Liu et al. 2024).

The western United States is very dependent on seasonal precipitation for water resources, for which S2S and seasonal forecast skill is notoriously low (e.g. Kumar and Chen 2020, Jiang et al. 2022, Sun et al. 2022, DeFlorio et al. 2023). Accurate long-term forecasts depend on accurate prediction of the frequency and intensity of atmospheric rivers (ARs, Ralph and Dettinger 2011) and of the location of their landfall, which depends on predicting large-scale weather patterns with precision. This remains challenging for the current generation of dynamical forecast systems because of model biases that hamper accurate representation of key large-scale teleconnections driving precipitation variability (e.g., Dong et al. 2024, Peings et al. 2024). Since the ML models do not rely on fluid mechanics equations and parametrizations to perform a forecast, they may be less subject to developing model biases during a forecast numerical calculation. This is however an open question as to this date few studies have analyzed long-range predictions of ML weather forecast models, in particular for mid-latitudes and the ENP-WNA region.

In this study we explore this question using two sets of long-range hindcasts from two different ML models, which we compare to state-of-the-art S2S hindcasts from ECMWF. We begin by evaluating and comparing the prediction skill of the different models, focusing on the North Pacific large-scale circulation and the MJO. We then assess how the models represent MJO teleconnections that are critical for accurate prediction skill of S2S climate variability over the Western US. Finally, we demonstrate the potential for ML models to carry out attribution studies of specific S2S climate events with large ensembles of runs.

2. Methods

2.1 S2S hindcast data

We use three sets of S2S hindcasts to explore S2S prediction skill, model biases and S2S teleconnections. First, hindcasts have been run using Spherical Fourier Neural Operators (SFNO) (Bonev et al., 2023), a purely data-driven ML model. We use the variant of the SFNO architecture originally introduced by Bonev et al. (2024) with updates for probabilistic forecasting by Mahesh et al. (2025a) leading to their version fit for purpose for "Huge

ENSemble” forecasting (hereafter SFNO-HENS). Briefly, this approach optimizes intrinsic dispersion by accessing high effective horizontal resolution, by minimizing spectral truncation in conjunction with increased ML model capacity through enlarged embedding dimensions. Dispersion is further optimized using a combination of multiple model checkpoints and bred vectors to generate ensembles of forecasts. The model is trained using the ERA5 reanalysis over the 1979-2015 period, with 2016 and 2017 as test years. Therefore, our validation period for this model is 2018-2023. Training is done using the ERA5 6-hourly data and 74 atmospheric fields, and 6-hourly outputs are generated for each of the 74 fields in the forecasts. Checkpoints are the same as those provided in Mahesh et al. (2025b). Hindcasts have been run over the period 2004-2023, initializing using ERA5 initial conditions, for the months of October to March (ONDJFM) which is our season of interest. Hindcast dates are provided in Table 1 for each of the models. They include 23 initialization dates for each year between 2004 and 2022, and 10 hindcasts between January and March 2023. This represents a total of 447 hindcasts. The 2004-2023 period for the hindcasts has been chosen to obtain sufficient sample size to evaluate model skill and biases. This creates an overlap with the training period (2004-2015), but because we primarily focus on long-range prediction (week 3 and beyond) when memory of initial conditions is mostly lost, we consider this as acceptable to address the research questions of the present study. Each SFNO-HENS hindcast includes 58 ensemble members, generated using 29 model checkpoints available at the time and two bred vectors for each. More details on the model and ensemble generation can be found in Mahesh et al. (2025a). Note that we found SFNO-HENS to be unstable at times, with large instabilities emerging near the surface and developing to unphysical values in the run (extremely negative or below absolute zero temperature). This concerns 329 ensemble members, out of the 447x58= 25926 hindcast ensemble members that were run with SFNO-HENS, or about 1.3% of the members. These members were discarded from the analyses.

The following fields have been saved and are analyzed in this study: the zonal and meridional wind at 850 hPa (U850 and V850), the zonal and meridional wind at 200 hPa (U200 and V200), the geopotential height at 500 hPa (Z500) and the total column water vapor (TCWV), all included in the list of 74 atmospheric fields used by SFNO-HENS for training and forecast inference. Precipitation and outgoing longwave radiation (OLR) are not part of these 74 fields and we derive the velocity potential at 200 hPa (VP200), from U200 and V200, to examine MJO variability (see section 2.2).

The second ML model is a hybrid model, the NeuralGCM model from Google Research (Kochkov et al. 2023). NeuralGCM combines a simple atmospheric dynamical core with machine-learning components. The ML component is a learned physics module that parameterizes physical processes with a neural network. A single ML parameterization is learnt that can be used universally at any given spatial location. While the dynamical core of the model solves the governing equations of the atmosphere, like a traditional GCM, the learned physics module predicts the effects of unresolved processes on simulated fields, effectively replacing traditional physics parameterizations. NeuralGCM achieves stable long rollout capability and good probabilistic calibration by leveraging the differentiability of their dynamical core to allow for multi-timestep optimization over time horizons from 6-hours to 5-days, including strategic terms in its objective function designed to promote a mixture of probabilistic skill and spectral fidelity. We use the stochastic version that injects noise at each time step to generate ensemble spread. It has been trained using ERA5 over the 1979-2019 period, so that the validation period for NeuralGCM is 2020-2023. The same set of hindcasts as for SFNO-HENS have been performed, only initialized a day earlier (Table 1). The initialization dates for SFNO-HENS and NeuralGCM were chosen to match the most recent ECMWF real-time forecasts. Because the SFNO-HENS hindcasts were run before NeuralGCM, the two sets of dates are not exactly the same because of different initialization dates in the updated ECMWF real-time forecasts. This is taken into account when comparing the hindcasts to ERA5 to evaluate the models. For each of the 447 hindcasts, 10 ensemble members have been run, saving the same output variables as for SFNO-HENS.

Finally, as a physics-based baseline for comparison, we use S2S hindcasts from the ECMWF Integrated Forecasting System (IFS), a state-of-the-art physics-based model and one of the best in terms of MJO prediction skill (Vitart et al. 2017, Lim et al. 2018, Kim et al. 2019, Liu et al. 2024). The hindcasts we have downloaded are from the Cy48R1 cycle of IFS forecasts (https://confluence.ecmwf.int/display/S2S/ECMWF+Model). The analyzed ECMWF hindcasts were run with a weekly frequency (Table 1), each including 11 ensemble members (a control and

10 perturbation runs), and we analyze the same atmospheric fields listed above. The ECMWF hindcasts are 45-days long, so we have run the SFNO-HENS hindcasts for 45 days too. The Neural-GCM hindcasts are a bit shorter, 43 days versus 45 days (Table 1), but this is sufficient for the forecast lead times analyzed in this study.

Model	**Initialization dates**	**Ensemble size**	**Length**
SFNO-HENS	JFM: 01-02 01-09 01-16 01-23 01-30 02-06 02-13 02-20 02-27 03-06 OND (no 2023) : 10-03 10-10 10-17 10-24 10-31 11-07 11-14 11-21 11-28 12-05 12-12 12-19 12-26	58	45 days
Neural-GCM	JFM: 01-01 01-08 01-15 01-22 01-29 02-05 02-12 02-19 02-26 03-04 OND (no 2023) : 10-02 10-09 10-16 10-23 10-30 11-06 11-13 11-20 11-27 12-04 12-11 12-18 12-25	10	43 days
ECMWF	JFM: 01-01 01-08 01-15 01-22 01-29 02-05 02-12 02-19 02-26 03-04 OND (no 2023) : 10-02 10-09 10-16 10-23 10-30 11-06 11-13 11-20 11-27 12-04 12-11 12-18 12-25	11	45 days

Table 1. Description of the three sets of 2004-2023 hindcasts.

2.2 Calculation of the MJO

The MJO is typically characterized using the Real-time Multivariate MJO (RMM) index (Wheeler and Hendon 2004), which is derived from a combined empirical orthogonal function (EOF) of three combined fields (U200, U850, OLR). Since OLR is not available for SFNO-HENS and Neural-GCM, we instead use the *velocity potential MJO (VPM)* index (Ventrice et al. 2013) which requires VP200 rather than OLR. Ventrice et al. (2013) found that using VP200 to compute the VPM index yields consistent results with the original RMM index using OLR. By using the VPM, we focus on the predictability of the wind-component of the MJO. The two leading EOFs of the combined field of equatorially averaged VP200, U850 and U200 are calculated using ERA5. The daily anomalies in VP200, U200 and U850 are then projected onto the two leading modes to retrieve the VPM1 and VPM2 timeseries that are used to assess the MJO phase and amplitude (Wheeler and Hendon 2004). The same multivariate EOF modes from ERA5 were used to construct the model hindcast VPM indices, i.e., the same reference modes are used for all datasets. They were calculated for the 1980-2017 period using daily anomalies meridionally averaged over 15°S to 15°N for each field, filtered by removing the previous 100-day running mean, and normalized by the ERA5 standard deviation over all longitudes.

The real-time MJO index requires applying a moving average, rather than a bandpass filter, for isolating S2S variability in the fields (because a bandpass filter requires future data which is not available in real time, or for 45-day hindcasts in our case). We remove a moving 100-day mean and we also calculate a moving 5-day mean from the daily anomalies in VP200, U850 and U200 before performing the combined EOF analysis. For this, we combine ERA5 daily anomalies with daily anomalies from the hindcasts to obtain timeseries longer than 45 days. More specifically, daily anomalies from the prior 100 days to a hindcast initialization date are concatenated with the daily anomalies from the hindcast, and for each day, the 100-day mean of the last 100 days is subtracted. This allows for a suppression of seasonal and longer time scales from the data, removing the influence of the El Niño Southern Oscillation in particular, while the additional 5-day moving average filters high-frequency (weather) variability. Prior studies have used various lengths for the moving average window (e.g., 40 days in Kiladis et al. 2014 , 120 days in Lin et al. 2008), we found that 100 days was enough to suppress low-frequency variability in our hindcasts. Daily anomalies for the hindcasts are derived from each model's respective climatology as a function of lead time, using the 2004-2023 set of hindcasts, which account for each model’s drift and biases. For consistency, daily anomalies from ERA5 are also calculated relative to the 2004-2023 climatology. Because 2004-2023 is a relatively short period to construct a climatology - climatologies are typically defined over 30 years - we have replicated the analyses using the 1991-2020 ERA5 climatology to calculate both ERA5 and the model anomalies. This has the advantage of using a longer period to define the climatology, but is inconvenient in that it does not account for model drift in their climatology, with lead time. The results are very similar for all the analyses presented in this paper, i.e. the findings are robust regardless of the climatology used to derive the anomalies (model versus ERA5, 2004-2023 versus 1991-2020). Consequently we only present results using anomalies that were derived from respective 2004-2023 climatologies. In the end, the fact that our IFS baseline will exhibit the expected decorrelation timescale associated with MJO predictability can be viewed as evidence that specifics of the above methodological choices are mostly immaterial.

Composites on the MJO are performed after filtering atmospheric fields with the same method as for the MJO VPM index (both 100-day and 5-day moving averages). For a certain field, prior 100-day ERA5 daily data is concatenated with the ensemble mean of a hindcast and a moving 100-day average is subtracted to suppress low-frequency variability. MJO events are identified as days with a value of the VPM greater than 0.7, with a minimum of 5 days spacing between each occurrence. The ENP-WNA region is defined as the [180°E/90°W ; 25°N/65°N] domain. The IO-WP (Indian ocean–Warm pool) domain is defined as the [40°E/145°W; 25°S/25°N] domain.

3. Results

3.1 Skill in predicting subseasonal North Pacific atmospheric patterns

We begin our model evaluations by calculating their skill score in predicting the average weekly Z500. We evaluate the skill for fall (October to December, OND) and winter (January to March, JFM) separately, to account for the seasonality in prediction skill in the ENP-WNA region (Peings et al. 2025), as well as in the MJO teleconnections in the North Pacific (Wang et al. 2023). A true skill score assessment requires using validation data only, i.e., not including any data from the training period. Given that we want to compare the models fairly by using a similar evaluation period for all, this would greatly limit sampling and the robustness of the results since it would have to be done over 2020-2023, the validation period for NeuralGCM. For the purpose of this study, which is to evaluate long-range prediction skill beyond the weather forecast time scale on which the ML models were trained, we choose to measure prediction skill on the maximum time period, 2004-2023, at the expense of including training data in the ML model hindcasts. Our assumption is that the memory of training is mostly lost beyond week 2 and that it should not enhance prediction skill too much at the S2S time scale. We aim to verify this assumption as thoroughly as possible throughout the remainder of the paper, by replicating the results over the shorter validation periods.

Figure 1 shows the Z500 prediction skill for week 3 and week 5 (i.e., weekly averages for lead time 14-21 and 28-35 days) in the 2004-2023 hindcasts initialized in JFM, measured from the grid-point correlation between observed and predicted Z500 anomalies (447 data points in each time serie). Fig. S1 completes this view with the week 1 prediction skill, and also shows the impact of using 2004-2023 (training data included in SFNO-HENS and NeuralGCM) versus 2020-2023 (validation data only). At week 1, all three models have very high skill with an average grid-point correlation in the ENP-WNA domain greater than 0.98 over the 2004-2023 period, greater than 0.97 over the 2020-2023 period (real skill of the ML models) (Fig. S1). Note that ECMWF also exhibits slightly lower skill over 2020-2023, illustrating that the higher skill over 2004-2023 in the ML models does not necessarily result only from including the test period in the data. Regardless of the evaluation period, such prediction skill is expected from the known performance of these models for weather forecasting, their main application. At week 3 (Fig. 1a-c), the skill has dropped, and it is maximum in the central North Pacific. In the ENP-WNA domain, a moderate average correlation of about 0.5 is found in all three models. At week 5, the skill has dropped further and is about 0.2 in the ENP-WNA domain. Remembering this skill is an estimate that includes training data, we calculated the equivalent prediction skill as in Figure 1 for the 2020-2023 period alone (Fig. S2). For both week 3 and week 5, the pattern of the grid-point correlations is more noisy and the average correlation drops by about 0.1 in the ENP-WNA domain. Again, the fact that a similar drop (and noisy correlation pattern) is also found in ECMWF points to lower inherent predictability during the 2020-2023 period, rather than to the sole influence of including training data in the ML models. We have also replicated the analyses for the OND season (Figure S3). In OND, the prediction skill is generally lower than in JFM, with an average correlation in the ENP-WNA domain of about 0.4 and lower correlation on and off the coast of the southwest US. Both JFM and OND correlation skill scores remain limited for reliable predictions of S2S climate variability in the Western US, given that a good prediction of large-scale patterns is critical to forecast temperature and precipitation anomalies.

Moderate skill aside, it is very interesting to note that the two ML models compare very well to ECMWF, at each lead time. The three models exhibit comparable prediction skill for both JFM and OND subseasonal Z500. A remarkable feature of the ML models is that their geographical repartition of high/low prediction skill resembles the one in ECMWF, especially at week 3 for both evaluation periods (Fig. 1 a-c, S2a-c and S3a-c). This highlights that ML models can be used for S2S forecasting, without significant skill loss but with lower computational costs and greater ability to run larger ensembles. However, they seem to confirm that achieving actionable S2S prediction skill in the ENP-WNA region is very challenging, both in physics-based, hybrid, and purely data-driven models. We now turn to the tropics to examine how the models predict the MJO and address potential causes for low prediction skill in the mid-latitudes beyond week 3.

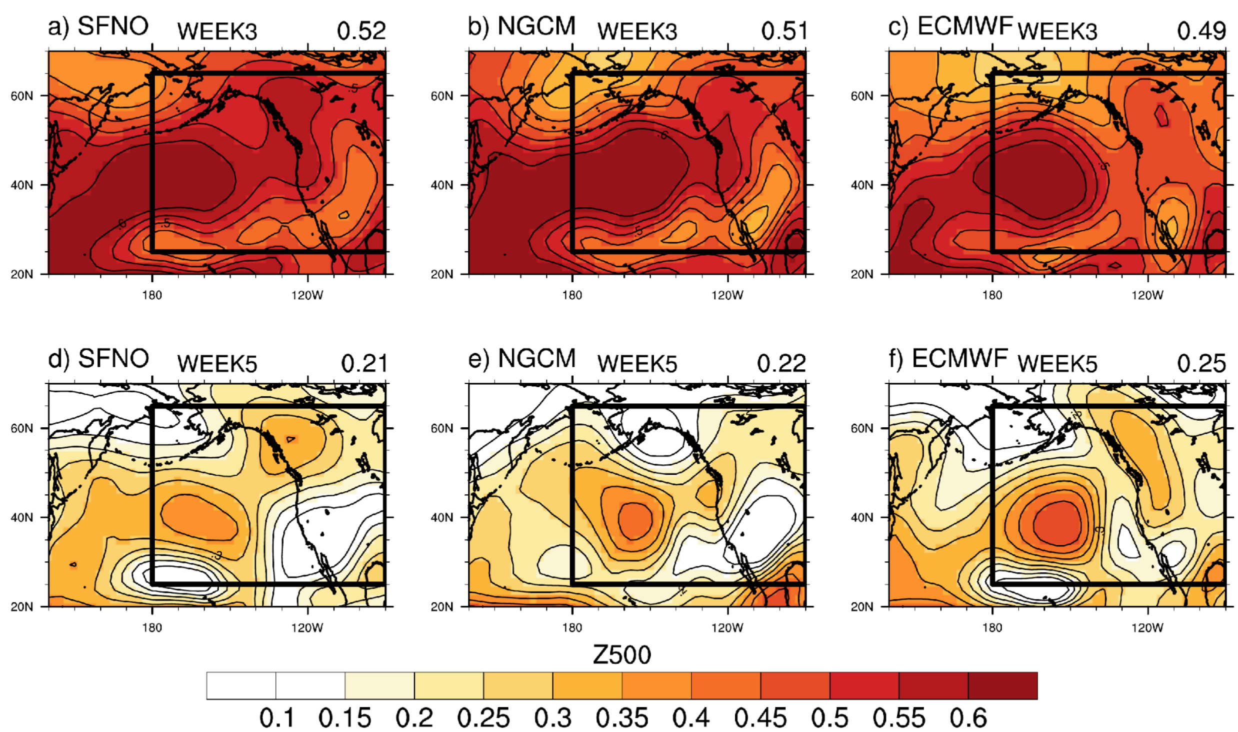

Figure 1. Grid-point correlation between observed (ERA5) and ensemble-mean prediction of average weekly Z500 (m) for the three different models and two lead times (weeks 3 and 5), for the JFM season (2004-2023 period). Numbers in the upper right indicate the average correlation value in the ENP-WNA domain (black box).

3.2 Skill in predicting the MJO and large-scale tropical convection

3.2.1 VPM prediction skill

The skill in predicting the MJO is first evaluated by computing the bivariate correlation coefficient between the two observed and predicted EOF components of the VPM index (see Text S1 for the definitions of the skill metrics). Figure 2 shows the MJO prediction skill for the three models, averaged over all the ONDJFM 2004-2023 hindcasts (solid lines) and the ONDJFM 2020-2023 hindcasts only (dashed lines). A threshold of R=0.5 is used to easily compare prediction skills at longer lead times. By design, the prediction skill is close to 1 after initialization, and it then decreases with forecast lead times. The 2004-2023 correlation skill crosses the R=0.5 threshold between lead time 31 days (NeuralGCM and ECMWF) and 33 days (SFNO-HENS). The threshold is met a bit later for each model when using the 2020-2023 hindcasts only (32, 33 and 34 days for NeuralGCM, ECMWF and SFNO-HENS, respectively). After 30 days, SFNO-HENS exhibits slightly higher skill than the two other models, especially over 2020-2023, but it also has lower skill between lead times 15 and 30 days over the latter period. We believe that such differences in model skill should not be overinterpreted as small data sampling when evaluating a model over a limited number of hindcasts limits the robustness of the evaluation. Evaluating the skill using the first 10 ensemble members of SFNO-HENS only (versus 58 in total) yields comparable results to Fig. 2a (not shown), i.e. disparity in ensemble sizes is not a very significant factor here, when comparing the models. This is illustrated in Fig. 2b in which we plot a 40-point moving correlation over 2004-2023, for the correlation skill between observed and predicted ONDJFM VPM index shown in Fig. 2a, for lead time 25 days (when SFNO-HENS has lower skill over 2020-2023). Correlation skill is calculated for each group of consecutive 40 hindcasts, starting from the initial one in 2004, and shifting the subset by one hindcast until the last one in 2023. There is considerable variability in the correlation skill of all three models due to internal variability (or noise), even during the training periods of the ML models. SFNO-HENS exhibits a noticeable drop when entering the validation period (marked by the vertical dashed blue line), but the correlation skill increases again at the very end, suggesting that there is a large amount of internal variability behind the differences in skill between 2004-2023 and 2020-2023 in Fig. 2a (rather than a clear impact of including training data in the model evaluation). Additional metrics that evaluate prediction skill are shown in Fig. S4. In terms of bivariate root-mean square error (RMSE, Fig. S4a) which measures the combined error in both VPM components, the three models exhibit comparable errors and the results align closely with conclusions drawn from calculating bivariate correlations (Fig. 2). The most striking difference between the models is for the error in VPM amplitude (or magnitude of the VPM vector), with ECMWF showing a faster decay in the amplitude of the VPM ensemble mean prediction (Fig. S4b). We expand on this at the end of this subsection. Phase errors are noisier and are included for completeness (Fig. S4c).

Unlike FuXi-S2S that was shown to extend the skillful forecast lead time from 30 days to 36 days compared to ECMWF, surpassing the performance of ECMWF S2S (Chen L. et al. 2024), the two ML models examined here exhibit more similar skill to ECMWF. Remember however that our periods of evaluation are different (2004-2023/2020-2023 versus 2017-2021 in Chen L. et al.), that we only evaluate the MJO between October and March (versus year-round in Chen L. et al. 2024) and that our MJO index uses VP200 (versus OLR in Chen L. et al. 2024). Our conclusion is that the two ML models exhibit comparable skill to ECMWF for predicting the MJO between October and March. ECMWF being a state-of-the-art S2S forecast model, this is an impressive achievement for recently developed models that are primarily designed for weather forecasting.

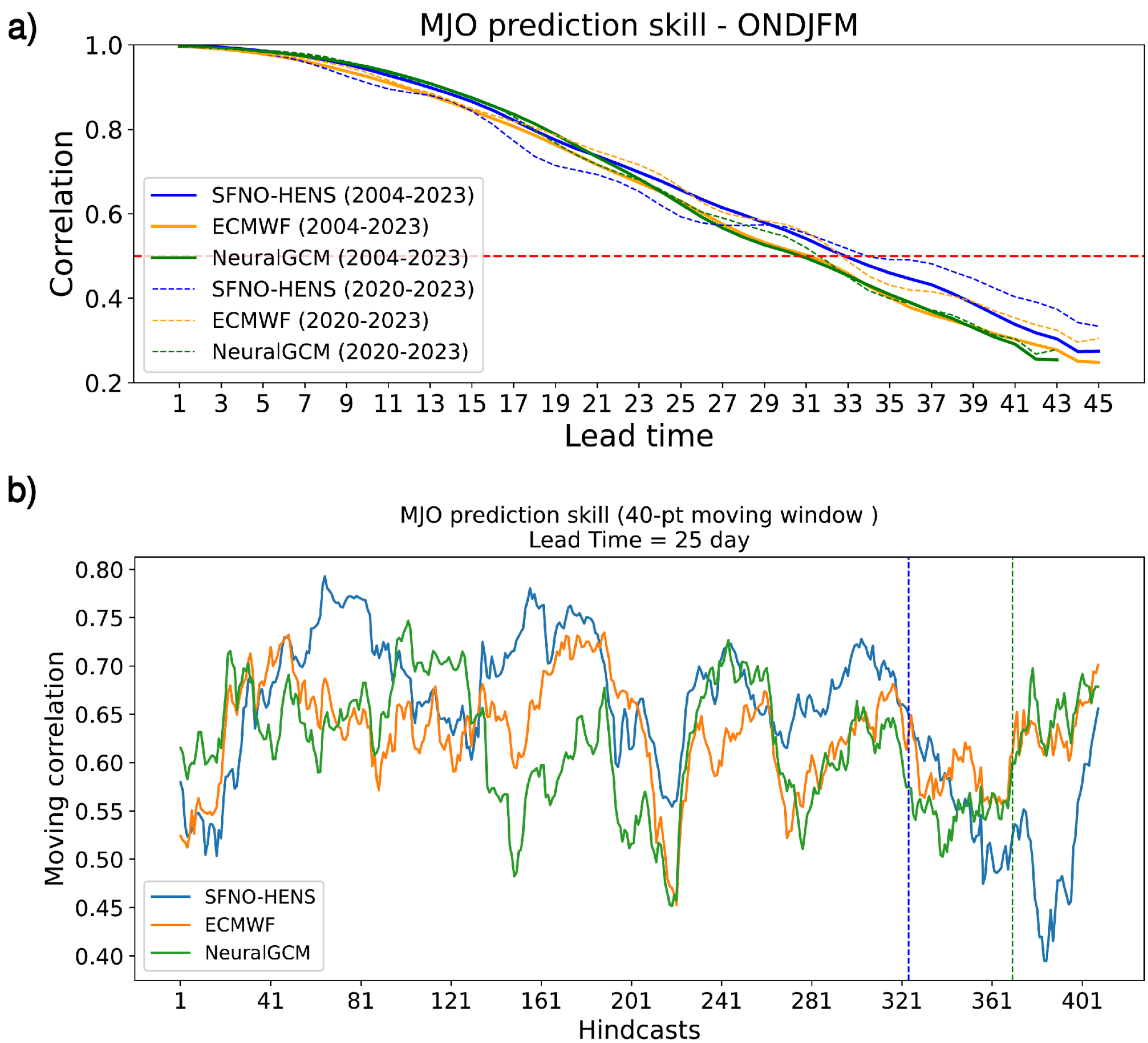

Figure 2. a) Skill in predicting the MJO between October and March (ONDJFM) as a function of lead time (days) for SFNO-HENS (blue), NeuralGCM (green) and ECMWF (orange), as a bivariate correlation between the observed (ERA5) and predicted VPM index. Solid lines are for the full 2004-2023 period of the hindcasts, dashed lines are for 2020-2023 (validation period only). The 0.5 correlation threshold is marked as a horizontal red dashed line. b) 40-point moving correlation between observed and predicted VPM index at lead time 25 days, from 2004 to 2023 (the first value corresponds to the correlation skill measured over the first 40 hindcasts, starting in 01-2004, and so on), for SFNO-HENS (blue), NeuralGCM (green) and ECMWF (orange). The vertical blue and green bars mark the beginning of validation-only data in the moving correlation (blue for SFNO-HENS, green for NeuralGCM).

In order to further characterize model skill in predicting S2S tropical variability, Figure 3 shows the grid-point correlation between observed and predicted VP200 for the JFM season, in the Indo-Pacific tropical sector. As expected from higher potential predictability in the tropics, prediction skill is higher than for ENP Z500, with a correlation of about 0.7 between observed and predicted IO-WP VP200 at week 3, and about 0.4 at week 5. While the three models have comparable skill at week 3, we note that NeuralGCM is outperformed by SFNO-HENS and ECMWF at week 5 (R~0.35 versus ~0.45 in the two other models, Fig. 3e), although this is not associated with significantly lower MJO index prediction skill (Fig. 2b). When measured over 2020-2023, the correlation skill score for IO-WP VP200 reduces to about 0.6 in all three models at week 3, with no truly remarkable difference in the spatial patterns of skill score among the models (Fig. S5). In OND, the prediction skill for IO-WP VP200 is essentially comparable among models at week 3 and 5, both when calculated over 2004-2023 (Fig. S6) and 2020-2023 (Fig. S7).

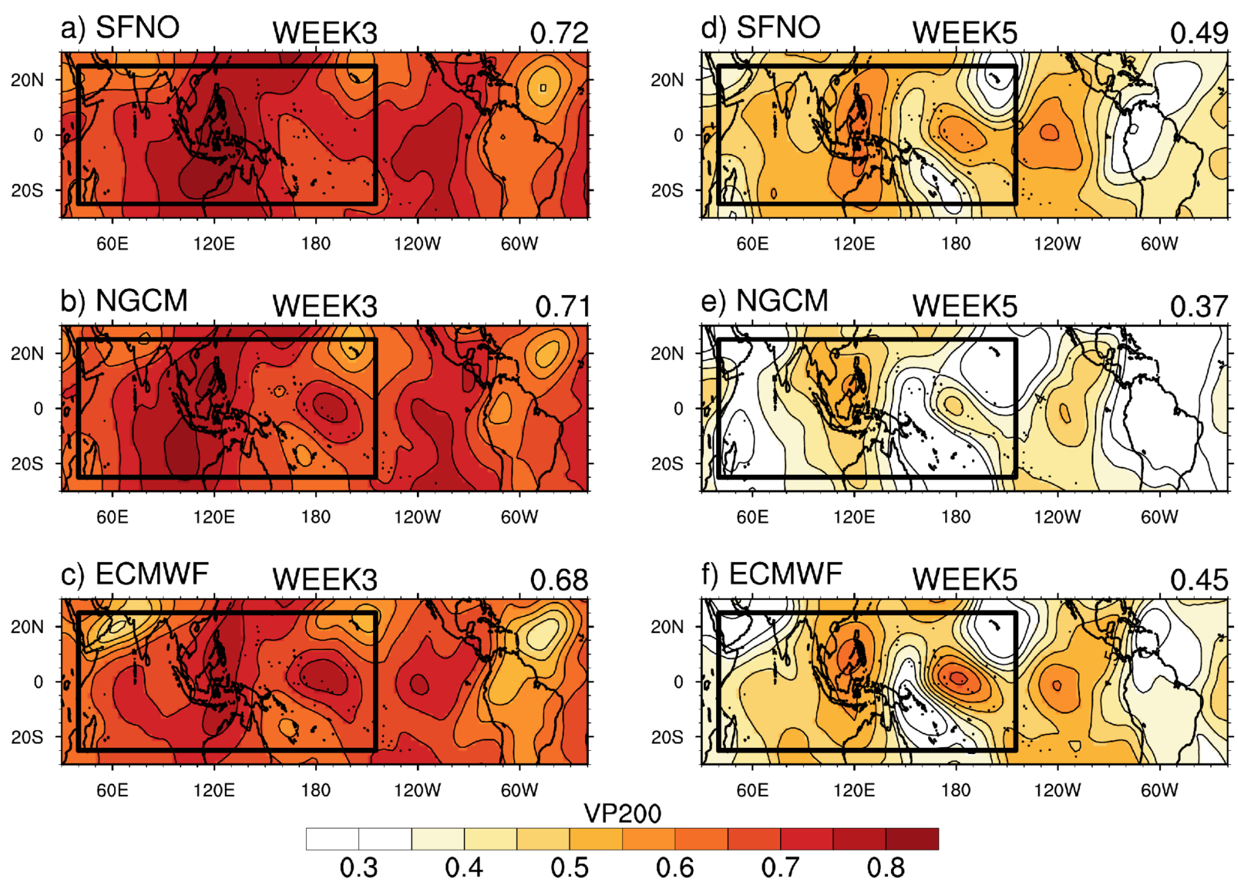

Figure 3. Grid-point correlation between observed (ERA5) and ensemble-mean prediction of average weekly VP200 (m^2/s) for the three different models and two lead times (weeks 3 and 5), for the JFM season (2004-2023 period). Numbers in the upper right indicate the average correlation value in the IO-WP domain (black box).

To further refine our evaluation of the MJO in the models, we illustrate in Figure 4 how the VPM amplitude (magnitude of the VPM vector, text S1) evolves in the ensemble mean prediction as a function of lead time. After aggregating all the hindcast ensemble means for a given model (2004-2023), we plot the distribution of the VPM amplitude as a function of lead time, including prior days from ERA5 that are merged with hindcasts to subtract the 100-day moving average. Before lead time=0, the distribution represents the variability in amplitude of the VPM index in ERA5, for the 447 hindcasts period analyzed in this study. The slight differences between the three curves prior to initialization are due to the 100-day moving average and also calculation of the VPM index that involves using the whole period for some of the parameters (ERA5+hindcast). The hindcasts start at day 0 on the x-axis, and we observe a decrease in VPM amplitude with lead time. Because we show a distribution of hindcast ensemble means (447 hindcasts) in this plot, the decrease in both mean (solid lines) and spread (shaded envelopes) is expected as in each individual hindcast the multimember spread - or uncertainty in the forecast - increases with lead time, leading to reduced amplitude in its ensemble mean. Interestingly, ECMWF shows a faster loss of amplitude in the ensemble-mean VPM during the first 10 days compared to the two ML models. This reduction is evident in both the ensemble mean of hindcasts, and the spread of ensemble means among hindcasts. Similar fast decay in the ensemble mean VPM amplitude in ECMWF is also observed when considering the 2020–2023 hindcasts only (see amplitude error, Fig. S4b). After approximately day 10, the rate of amplitude decay in the ECMWF ensemble mean becomes more comparable to that of the two ML models. Note that the accelerated decay of the predicted MJO amplitude in ECMWF forecasts is not necessarily representative of all physics-based models (e.g., Chen et al., 2025). Indeed, as earlier ECMWF hindcast versions displayed the same behavior, with a notably faster decay than other S2S physics-based models (e.g., Fig. 1 in Kim et al. 2019), this characteristic appears to be a recurring feature of ECMWF's S2S model family.

An important remark to be made here is that we discuss the amplitude of the ensemble mean VPM, not the amplitude of the VPM in individual ensemble members of the models. A critical feature of a well-balanced prediction system is that prediction error from the ensemble mean matches the spread in the ensemble (or the system is said to be overconfident, or underconfident, depending on their ratio, see for example Fortin et al. 2014). Figure 4b shows the ratio between the spread (the square root of the VPM ensemble variance) and the RMSE of the VPM as a function of lead time (see Text S1 for the formulation). Ideally, a well-calibrated forecast ensemble should yield a spread-to-error ratio of one, and overall, the three models remain close to this target. Initially, all three models exhibit ratios below one, indicating underdispersion (overconfidence) in the predictions. The ratio then increases toward unity as the ensemble spread begins to match the forecast error. Beyond these initial days, SFNO-HENS maintains a nearly constant ratio around one throughout the hindcast period, suggesting a well-calibrated ensemble. NeuralGCM remains slightly underdispersive, with a ratio around 0.9 over the 43-day hindcast period. In contrast, ECMWF shows ratio values exceeding one during the first 10 days, indicating overdispersion—i.e., the ensemble spread exceeds the actual forecast error—despite good predictive skill. This behavior is consistent with ECMWF’s faster loss of amplitude in the ensemble-mean VPM compared to the two ML models, as excessive spread tends to dampen the ensemble-mean amplitude.

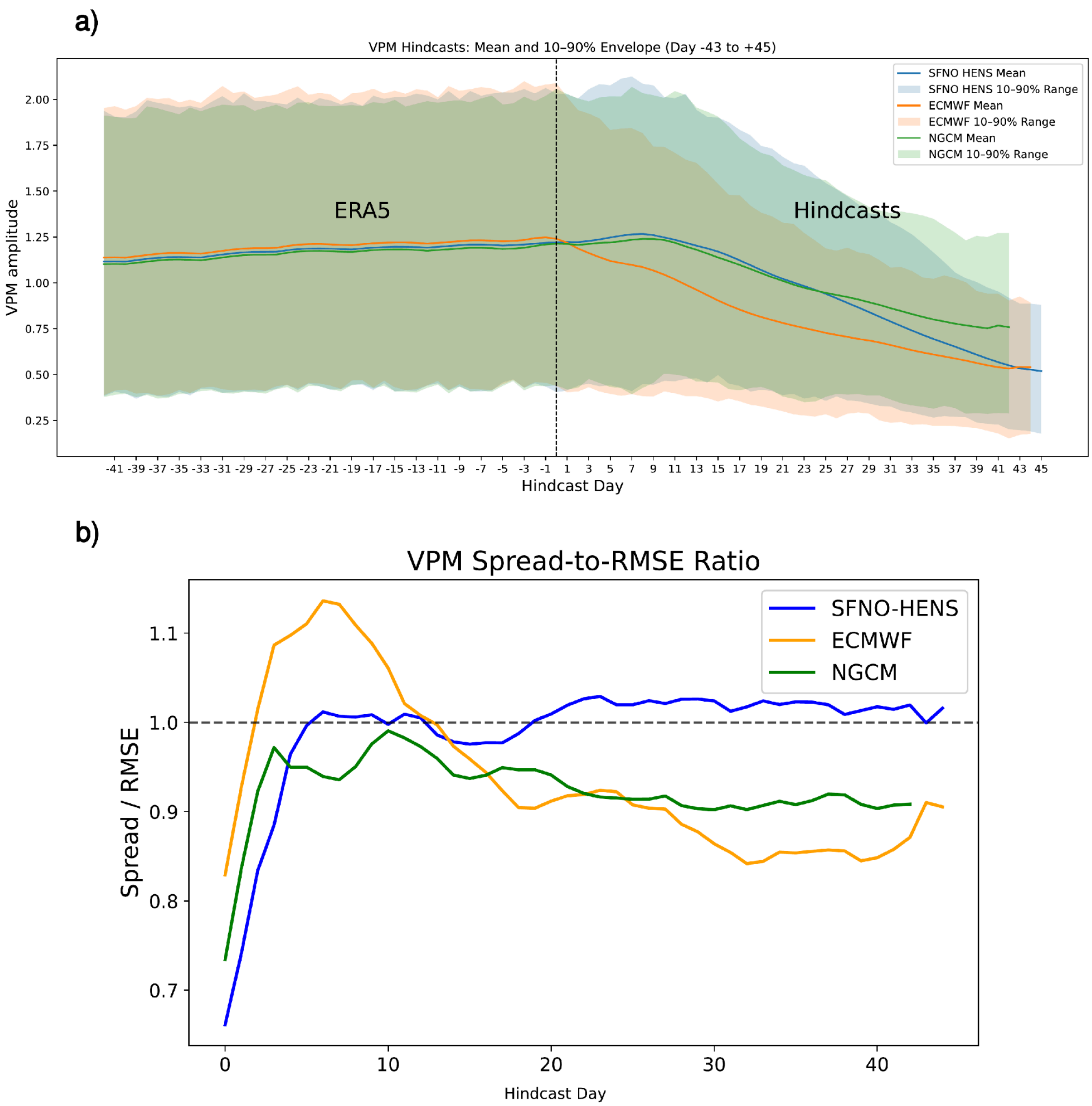

Figure 4. a) Distribution of the amplitude of the ensemble mean VPM index, among the 2004-2023 ensemble of ONDJFM hindcasts of a given model (447 hindcasts), as a function of lead time. Negative lead times correspond to ERA5, which is concatenated to the hindcast ensemble means before filtering the fields and computing the VPM index. The envelopes show the 10th/90th percentiles of the distributions (spread in the 447 hindcast ensemble means). Blue is for SFNO-HENS, orange is for ECMWF, green is for NeuralGCM. b) Average spread-error ratio among the 447 hindcasts when predicting the VPM index as a function of lead time, as measured as the ratio between ensemble spread (using all

available ensemble members for each model) and the RMSE in the ensemble mean predictions. Blue is for SFNO-HENS, orange is for ECMWF, green is for NeuralGCM.

3.2.2 Representation of MJO propagation

MJO propagation for different initial MJO phases are compared in Figures 5 and 6, for the three models and ERA5. We plot composites of VP200 and U200 daily anomalies as an average of all MJO events among the hindcasts/reanalysis, with an initial VPM>0.7 in phases 2 or 3 (Fig. 5) and in phases 6 or 7 (Fig. 6). Compared to ERA5, the propagation after phases 2 and 3 of the MJO (initial convection in the Indian ocean sector) is well predicted in SFNO-HENS and NeuralGCM, with accurate amplitude and speed of the propagation (as can also be seen in the U200 anomalies, black contours). Compared to ERA5, SFNO-HENS slightly overestimates the amplitude of the anomalies while NeuralGCM slightly underestimates them, but both ML models are able to represent a realistic MJO propagation. ECMWF also exhibits a realistic MJO propagation but it tends to underestimate the amplitude of the VP200 and U200 anomalies in the Indian sector, as expected from too fast decay in the MJO identified in Fig. 4.

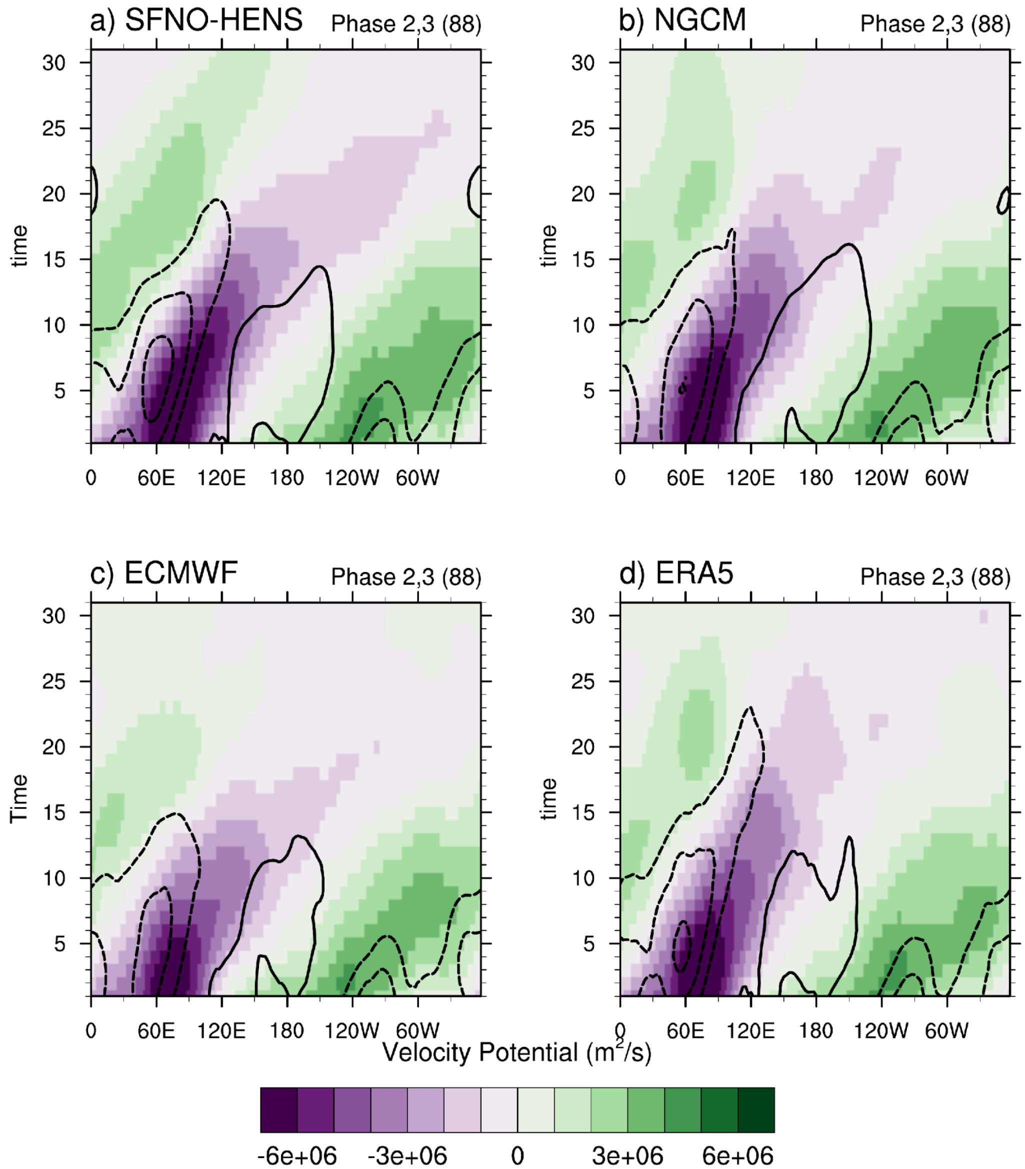

Figure 5. MJO propagation from phases 2 and 3 in ONDJFM, for : a) SFNO-HENS ; b) Neural-GCM ; c) ECMWF ; d) ERA5. Shading shows VP200 anomalies (m^2/s), contours show U200 anomalies (2 m/s contour interval). The number of MJO events used to create the composite is given into parentheses.

Similar conclusions apply for MJO propagation starting from phases 6-7 (initial convection in the western Pacific, Fig. 6). Again, SFNO-HENS and NeuralGCM exhibit remarkably realistic MJO propagation, while ECMWF lacks amplitude in the wind and deep convection anomalies. This is also observed during the 2020–2023 period (Fig. S8, phases 2–3), although the composite includes only a limited number of MJO events over such a short timeframe, preventing a robust assessment. These analyses confirm that ML models are able to accurately represent MJO propagation, both in atmosphere-only S2S forecasts as here, and in longer climate simulations with prescribed sea surface temperature (Watt-Meyer et al. 2025). As to why it does not result in higher S2S skill in the mid-latitudes at week 3 and beyond, there are multiple potential explanations, one of which being how MJO teleconnections are predicted. This is explored in the next section.

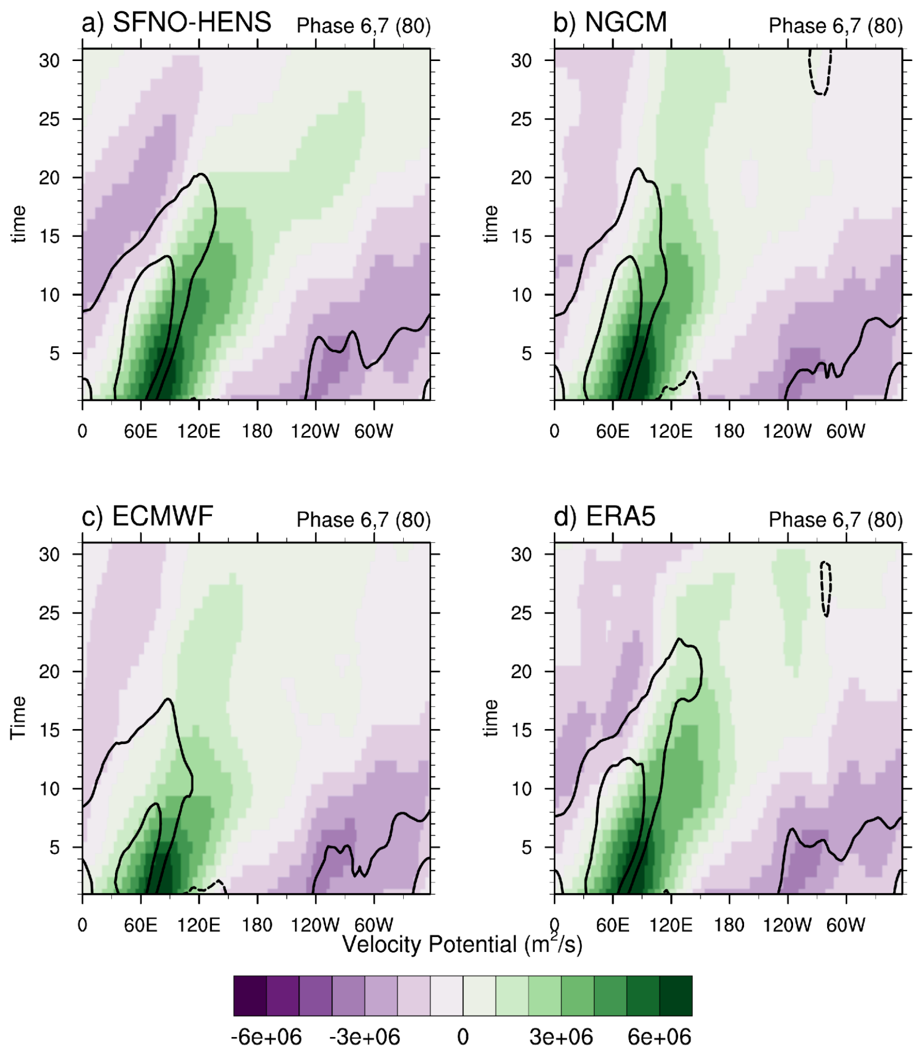

Figure 6. MJO propagation from phases 6 and 7 in ONDJFM, for : a) SFNO-HENS ; b) Neural-GCM ; c) ECMWF ; d) ERA5. Shading shows VP200 anomalies (m^2/s), contours show U200 anomalies (2 m/s contour interval).

3.3 Evaluation of MJO teleconnections in the models

3.3.1 Week-1 MJO teleconnection

In this section we evaluate the short-term (week 1) MJO teleconnection in each of the models. This is first done by compositing Z500, VP200 and wind at 850 hPa filtered daily anomalies for MJO events that pass the 0.7 sigma threshold criterion of the VPM index (derived from ERA5), under a given phase. Figure 7 shows the resulting anomalies for the difference between MJO phases 2 and 3 combined, minus 6 and 7 combined, which represent distinct locations for deep convection, i.e. in the Indian ocean for phases 2-3, and in the western Pacific for phases 6-7. A MJO event is selected based on its VPM index at day 0, and the fields are averaged for days 5 to 9 (end of week 1/beginning of week 2) following this event, to account for the time that MJO-induced extratropical wave trains take to develop. The composites in this subsection represent the MJO teleconnection in the hindcasts before model biases have had the time to develop. Mean U200 climatological biases of the models corresponding to selected events are plotted (black contours), along with the corresponding U200 climatology in ERA5 (Fig. 7d), but because we are so close to initialization, no bias can be seen in any of the three models in the extratropics (Fig. 7a-c), they emerge later in the hindcasts (see next subsection). The difference between MJO phases 2-3 and 6-7 corresponds to a dipole in the tropical VP200 anomalies, with enhanced upper-level divergence (negative VP200 values) over the Indian Ocean/Maritime continent and upper-level convergence anomalies in the central/eastern tropical Pacific. The pattern and amplitude of the VP200 anomalies is realistic in all models, although their amplitude is already substantially weaker in ECMWF, as expected from the fast decay in the predicted ensemble mean MJO in this model (section 3.2.1).

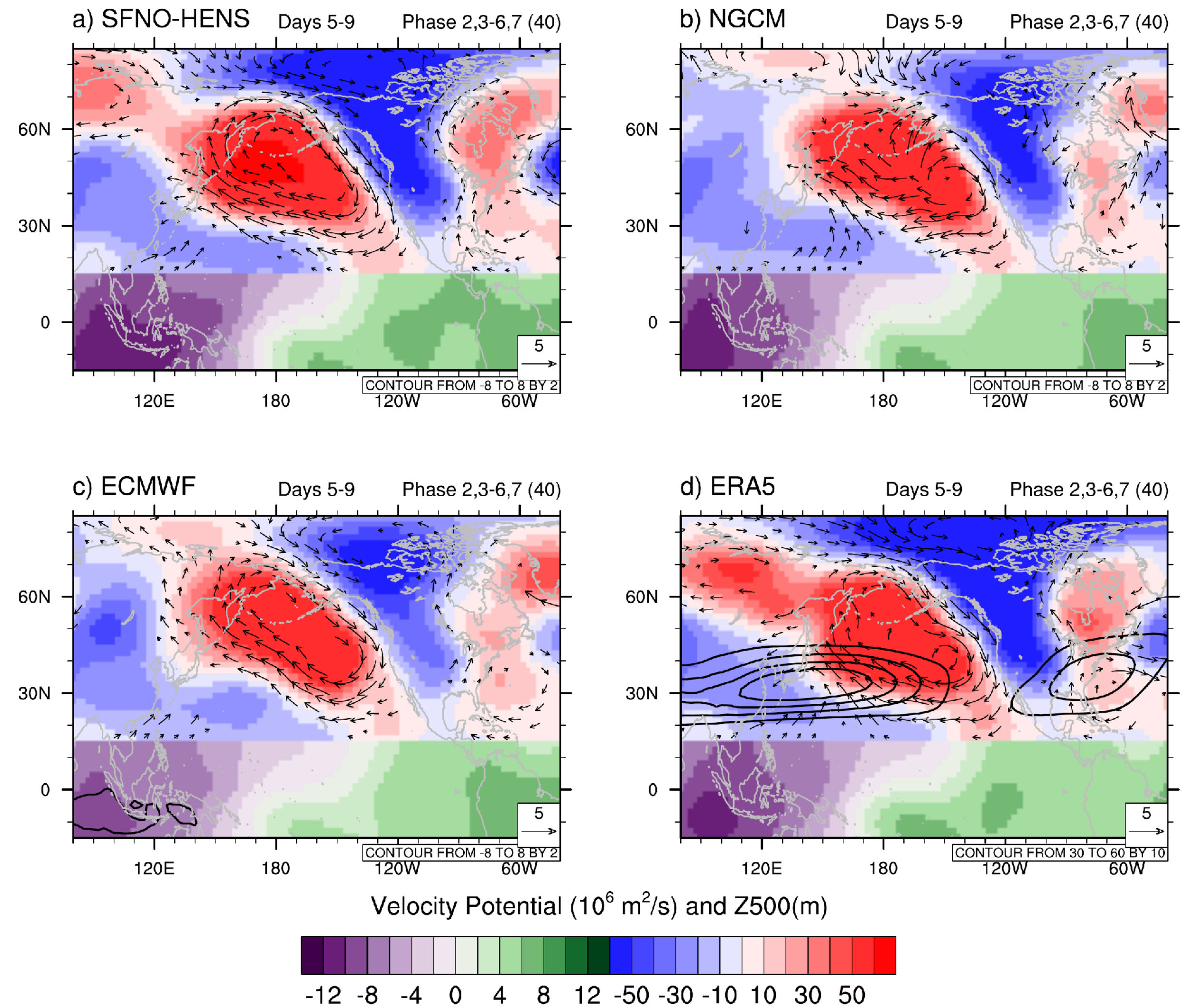

Figure 7. a) MJO teleconnection for phases 2-3 minus phases 6-7, ***in the JFM hindcasts***, as a composite of Z500 anomalies (shading above 15°N, m), VP200 anomalies (shading below 15°N, m^2/s) and wind at 850 hPa (vectors, m/s), 5-9 days after MJO days with amplitude of the VPM index greater than 0.7, ***at lead time 0*** *(week 1)*, for : a) SFNO-HENS ; b) Neural-GCM ; c) ECMWF ; d) ERA5. 850 hPa wind vectors under 1 m/s are masked. The number of events selected for the composite is given in the top right of the panel (in parentheses). Black contours in d show the U200 climatology (10 m/s contour interval), and the model biases (difference from d) in a, b and c (2 m/s contour interval).

The associated Z500 anomalies (red/blue shading north of 15°N, Fig. 7) exhibit a pronounced wave train that originates in the East Asia jet stream entrance region (see U200 climatology in Fig. 7d), with a dipole of positive (ridge) and negative (trough) Z500 anomalies over the North Pacific and North America and associated wind circulation (850 hPa wind vectors). The strong ridging in the North Pacific prevents the propagation of mid-latitude storms and ARs over the western US, inducing dry conditions over the region as confirmed by similar composite analyses of TCWV (Fig. 8, TCWV is not available for NeuralGCM but we can look into TCWV composite anomalies in SFNO-HENS and ECWMF). As can be seen on the maps of Fig. 8, the ridge circulation over the North Pacific transports air masses from the northeast Pacific over the western US, when wet conditions in the western US and more so the southwest US are typically associated with a zonal or southwest-northeast wind circulation that advects moist air from the storm track or from the subtropics. These results are consistent with Wang et al. (2023) who performed similar analyses using reanalyses but with slight differences in methodology (in particular a different period of analysis and the use of the OLR-based MJO index), which supports the robustness of the results. In summary, at week 1 the three models accurately predict the 5-9 days extratropical large-scale circulation anomalies associated with the MJO (Fig. 7a-c), although ECMWF tends to underestimate the amplitude of deep convection MJO anomalies.

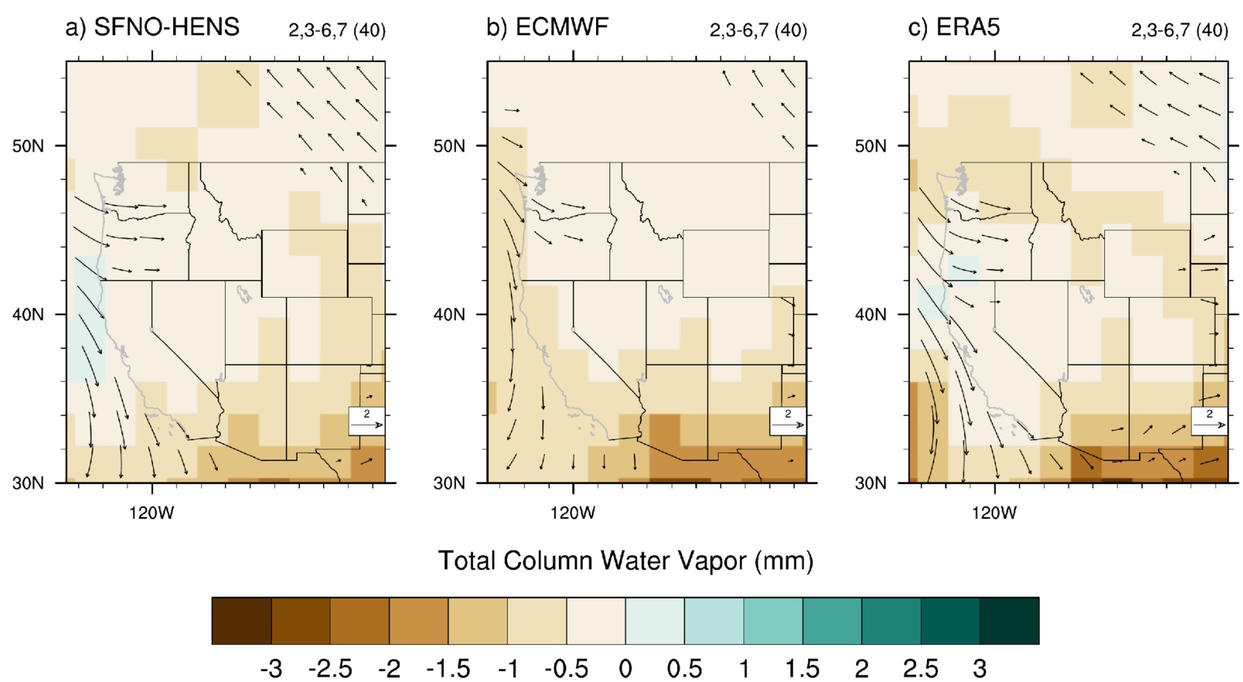

Figure 8. a) MJO teleconnection for phases 2-3 minus phases 6-7, ***in the JFM hindcasts***, as a composite of total column water vapor anomalies (shading, mm) and wind at 850 hPa (vectors, m/s), 5-9 days after MJO days with amplitude of the VPM index greater than 0.7*, **at lead time 0** (week 1)*, for : a) SFNO-HENS

; b) ECMWF ; c) ERA5. 850 hPa wind vectors below 0.5 m/s amplitude are masked. The number of events selected for the composite is given in the top right of the panel (in parentheses).

3.3.2 Week-3 MJO teleconnection

Now that we have verified that the model hindcasts provide accurate MJO teleconnections at short lead time, we turn towards a longer S2S lead time. Figures 9 and 10 are similar to Fig. 7 and 8, except this time the composites are based on days that meet the MJO identification criteria in ERA5 *14 days* after initialization (week 3). In other words, we evaluate similar MJO teleconnections as in Fig. 7 and 8, but two weeks after initialization when the influence of initial conditions has diminished and model biases have emerged. This allows us to explore two things. First, whether the models are able to make skillful forecasts of the 5-9 days MJO teleconnection at week 3 (accounting for the 5-9 day lag between a selected MJO day and the composited fields, the composite represents days 19-23 after initialization, i.e. the end of week 3/beginning of week 4). Second, whether model biases in the mean flow and in particular in the East Asian Jet Stream (EASJ; U200 contours) affect the representation of MJO teleconnections in each model. Because the composite analysis of ERA5 is not affected by lead time (Fig. 9d), we expect relatively similar results as in Fig. 7d because different days, but with similar MJO characteristics, are selected in the calculation of the composite. This is the case (Fig. 7d and Fig. 9d resemble each other), and this is also the case for the TCWV anomalies (Fig. 8c versus Fig. 10c) that exhibit similar patterns although slightly different amplitudes in the anomalies. As expected from the results of section 3.2.1, the models exhibit reduced MJO amplitude versus ERA5, as seen from the VP200 anomalies (Fig. 9a-d, south of 15°N). This is particularly the case for ECMWF which retains about half of the amplitude in the anomalies, compared to the same anomalies in SFNO-HENS and NeuralGCM. The three models capture the general Z500 pattern in the mid-latitudes, but it is again underestimated in ECMWF, pointing to a better representation of the MJO teleconnection in the ML models.

An interesting feature of the ML models is that they are less affected by systematic drift to biased states than ECMWF, which begins to exhibit significant model biases in U200 in the EAJS region (black contours, Fig. 9c). NeuralGCM only exhibits a small positive U200 bias in that region (Fig. 9b), and model biases are difficult to detect in SFNO-HENS over the Central Pacific; farther west, modest zonal flow anomalies do develop (Fig. 9a). This points to an interesting advantage of ML models, since reduced biases in the mean flow provide more accurate large-scale teleconnections, such as the ones associated with the MJO. However, in the present analysis we do not see the clear benefit of reduced systematic drift in the two AI models (in terms of average Z500 prediction skill for example, Fig. 1), presumably because the skill remains low and other effects are at play, such as too low MJO amplitude in the ensemble mean prediction.

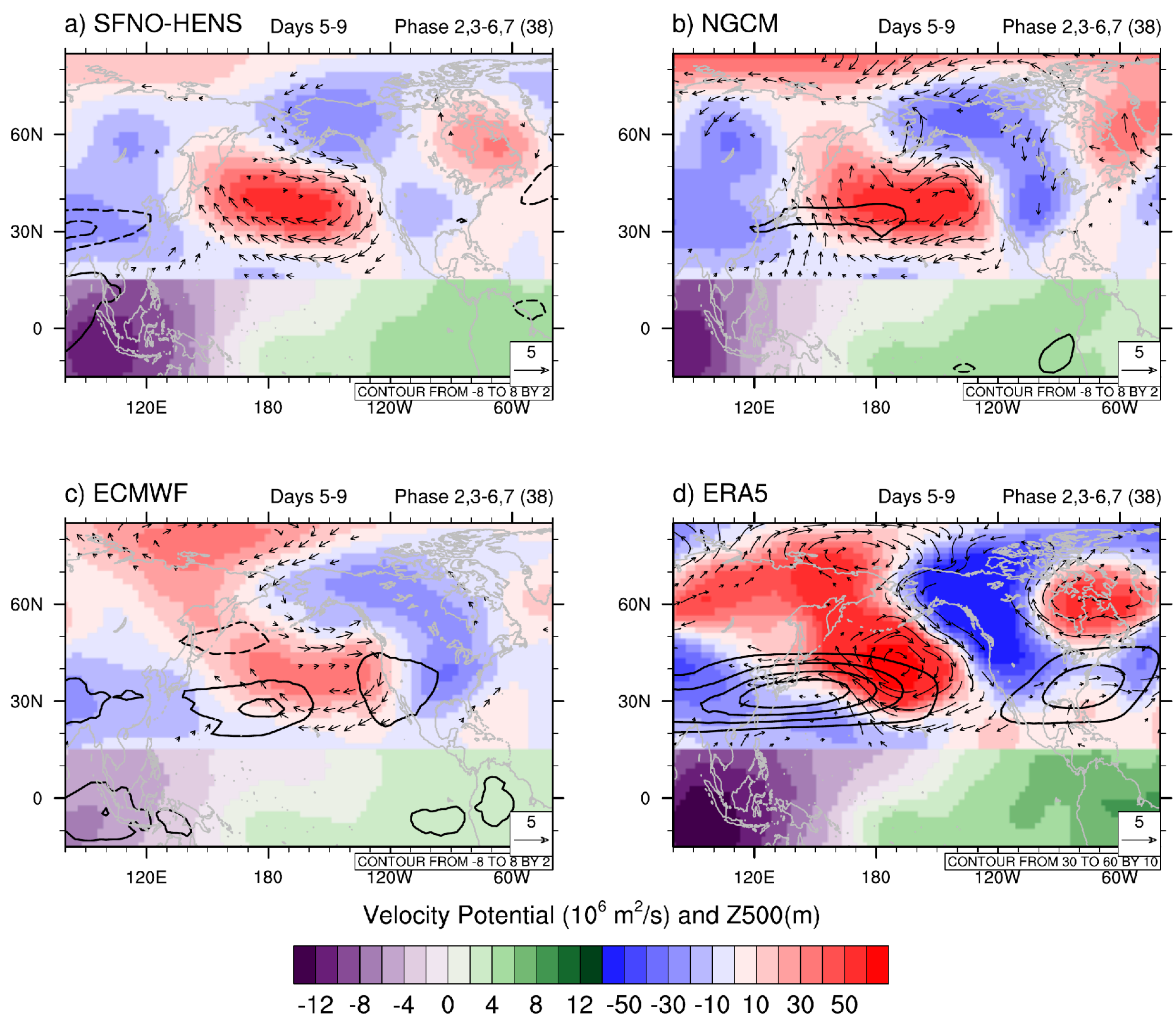

Figure 9. a) MJO teleconnection for phases 2-3 minus phases 6-7, ***in the JFM hindcasts***, as a composite of Z500 anomalies (shading above 15°N, m), VP200 anomalies (shading below 15°N, m^2/s) and wind at 850

hPa (vectors, m/s), 5-9 days after MJO days with amplitude of the VPM index greater than 0.7*, **at lead time 14** (week 3)*, for : a) SFNO-HENS ; b) Neural-GCM ; c) ECMWF ; d) ERA5. 850 hPa wind vectors under 1 m/s are masked. The number of events selected for the composite is given in the top right of the panel (in parentheses). Black contours in d show the U200 climatology (10 m/s contour interval), and the model biases (difference from d) in a, b and c (2 m/s contour interval).

When compared to Fig. 8, the impact of a less accurate MJO teleconnection at week 3 is clearly visible in the TCWV composites (Fig. 10). SFNO-HENS captures dry conditions in the Southwest US (SWUS) but it underestimates dryness in the Pacific Northwest region (Fig. 10a), compared to ERA5 (Fig. 10c). ECMWF also accurately predicts dry conditions in the SWUS but with underestimated amplitude, and it predicts wet conditions in the Northwest (Fig. 10b). This is due to a weaker ridge and inaccurate wind circulation in the eastern North Pacific in ECMWF (Fig. 9c), which favors storm propagation and wet conditions over the Northwest US.

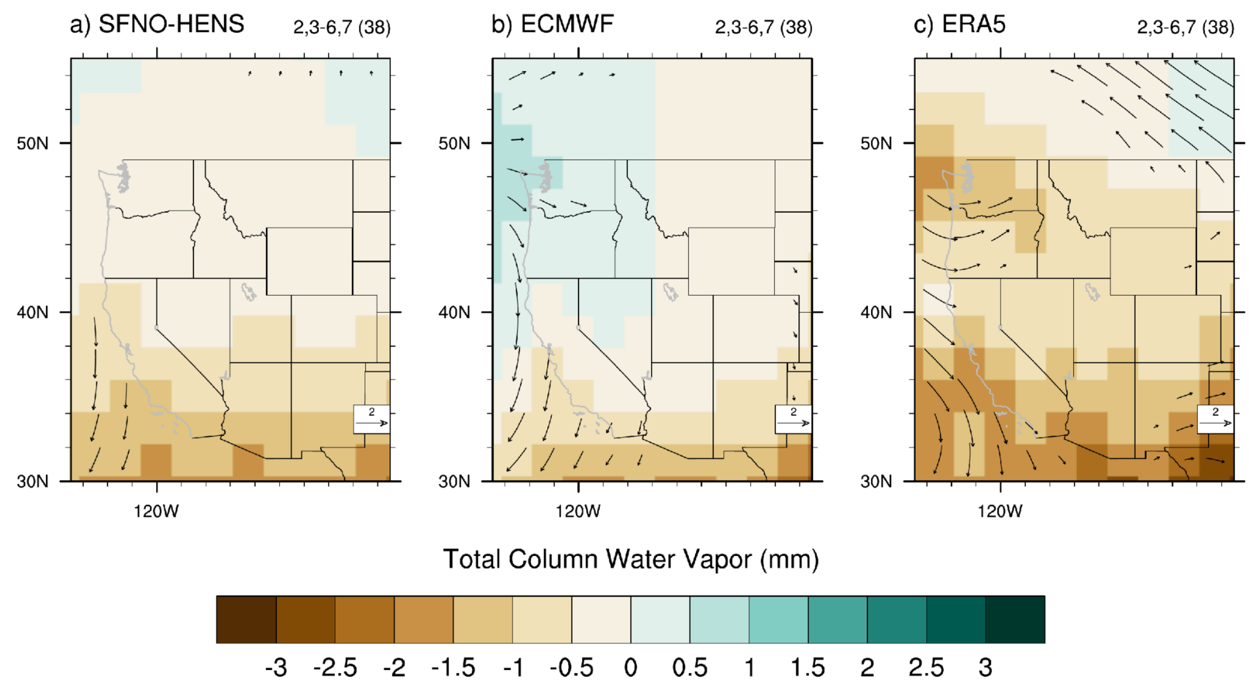

Figure 10. a) MJO teleconnection for phases 2-3 minus phases 6-7, ***in the JFM hindcasts***, as a composite of total column water vapor anomalies (shading, mm) and wind at 850 hPa (vectors, m/s), 5-9 days after MJO days with amplitude of the VPM index greater than 0.7*, **at lead time 14** (week 3)*, for : a) SFNO-HENS ; b) ECMWF ; d) ERA5. 850 hPa wind vectors under 0.5 m/s are masked. The number of events selected for the composite is given in the top right of the panel (in parentheses).

For the sake of completeness, similar composites as Fig. 7 and 9 are shown in SI, using the 2020-2023 hindcasts only (Fig. S9 and S10). In line with the 2004-2023 analyses, the models exhibit good accuracy at week 1. The results for week 3 are more contrasted. Note however that data samples are too small for any robust assessment of the models versus ERA5 (8 phases 2-3 and 6-7 MJO events for week 1, 7 for week 3). We also provide similar composites as Fig. 7 to 10 for the OND season (Fig. S11 to S14). In ERA5, the MJO teleconnection measured as phases 2-3 minus 6-7 is quite different from the JFM teleconnection shown above. Rather than a strong ridge in the North Pacific, a smaller ridge is observed in the central North Pacific, followed by a trough centered north of Alaska and another ridge over North America (Fig. S11). This results in a dipole of TCWV with dry conditions in the Southwest United States (SWUS), like in JFM, and wet conditions in the Pacific Northwest (opposite to JFM) (Fig. S12). This illustrates how the impact of the MJO is season dependent and must be evaluated accordingly (Wang et al. 2023). As in JFM, the week-1 MJO teleconnection is accurate, although ECMWF also underestimates the amplitude of the MJO anomalies (Fig. S11). At week 3, the two ML models struggle to predict the amplitude of the trough-ridge pattern over Alaska and North America (Fig. S13), which results in too weak westerlies advecting moisture over the US west coast (Fig. S14) compared to ERA5. Although of larger amplitude, the Z500 anomaly patterns in ECMWF are not accurate (Fig. S13c), resulting in a wrong prediction of TCWV over the US West Coast (Fig. S14b, dry conditions versus wet in ERA5). Of note, U200 model biases are mostly absent (Fig. S13), including in ECMWF, so that they do not play a significant role in OND.

3.3.3 Discussion

Overall, this section suggests the following: 1) accurate S2S predictions of Western US weather conditions beyond week 2 remain challenging for both physics- and AI-based forecast models, even in the presence of significant MJO activity ; 2) precise representation of the MJO teleconnections is critical because slight differences in atmospheric patterns make for large differences in moisture advection and precipitation. To this date, this high degree of accuracy is out of reach for the two ML models, as it is for ECMWF and physics-based models in general; 3) overall, the ML models perform equally well to ECWMF in terms of S2S prediction of the MJO and its teleconnections in the North Pacific, with the advantage of maintaining ensemble mean amplitude of predicted MJO activity longer than ECMWF. Moreover, the ML models are less affected by systematic background flow biases, so if their ongoing rapid development leads to greater MJO prediction skill in the future, improved mean-state ducting may allow for more realistic MJO teleconnections in the North Pacific and associated weather patterns over North America. On this last point, it is important to note that model biases in the ML models will depend on the period used for training, but assuming training is updated with the latest years of reanalyses, this feature should persist in future model versions. It is important to keep in mind that the comparison with ERA5 has limitations due to the relatively short sample size (20 years) when exploring MJO teleconnections. ERA5 represents a single realization of the climate, whereas the model hindcasts provide an ensemble, so internal variability (weather noise) inevitably influences the composites derived from ERA5.

3.4 Perturbation experiment for attribution

Aggregate scores of S2S skill obscure anomalous periods of enhanced predictability associated with forecasts of opportunity. To investigate such a time period, we now turn to a case study that represents a novel sensitivity test of physical credibility and robustness for a ML model in the context of S2S prediction. This also helps illustrate the main advantage of ML models, which is the ability to run large ensembles of runs with relatively low computational cost. This feature is of course beneficial to ensemble forecasting, but also it makes ML models very convenient tools for performing perturbation experiments and exploring physical mechanisms. Sensitivity experiments can be done by simply altering initial conditions that are provided to a ML model after training. This is the same protocol that could be followed using a GCM, but here with a straightforward (simply altering initial conditions from ERA5) and with the capability of running very large ensembles with reasonable computational costs. This is done here using the SFNO-HENS model.

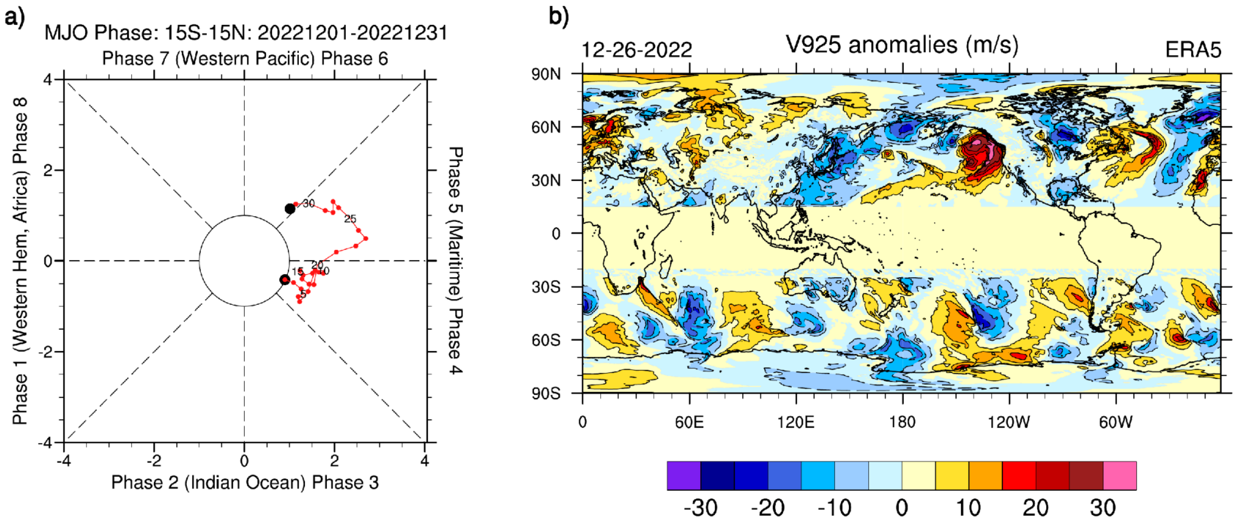

Figure 11. a) Observed MJO phase between December 1st and December 31st 2022. b) Initial conditions for the 2022-12-26 perturbation experiment (SFNO-HENS) with climatology imposed in the tropics, seen here through 6-hourly V925 anomalies (m/s) on December 26th 2022, 00:00 UTC.

We selected December 26th, 2022, as our hindcast of interest for this experiment because: 1) it was followed by three weeks of trough conditions in the eastern North Pacific and wet weather over the west coast of the US; 2) it is an example of accurate week-2 forecast in the ensemble of SFNO hindcasts; 3) it corresponds to a pronounced phase 5-6 MJO (Fig. 11a). The week-2 ensemble mean prediction is shown in Fig. 12b. The trough anomaly is well captured by the forecast, with an ACC of 0.89 over the ENP-WNA domain, as part of a Rossby wave pattern (alternating ridge/trough/ridge anomalies) that is accurately represented in the forecast. Prediction skill remains high beyond week-2, with an ACC of 0.78 and an accurate representation of the persistent trough in the ENP at week-3, such that the entire 3-week wet episode in the SWUS is well predicted (not shown).

In order to explore the potential role of tropical forcing and in particular of the MJO, we conduct another hindcast initialized on December 26th 2022, similar to the original hindcast but with altered ERA5 initial conditions. For this, we prescribe the 6-hourly climatological values of the 74 fields, derived from the 1979-2015 period (training period of the SFNO model), in the tropics. The climatology is prescribed between 15°N and 25°S along all the longitudes, with 5°N buffer zones at the limits of the tropical domain that includes a 11-pixel moving latitudinal average. This smooths out latitudinal gradients and limits sharp edges at the boundaries of the forcing domain. An illustration of the protocol is shown in Fig. 11b, where the meridional wind 6-hourly anomalies at 925 hPa in the initial conditions are plotted. By design, anomalies are absent or close to zero in the tropics, effectively suppressing any tropical variability forcing from the initial conditions (the same is true for all 74 fields). Suppressing tropical anomalies in the IC significantly degrades the week-2 forecast, with an ACC of 0.18 in the ENP-WNA domain in the altered clim-trop hindcast (Fig. 12c). The Rossby wave and ENP trough is replaced by a dipole of low and high Z500 anomalies in the North Pacific, and a ridge centered off the west coast of the US. Precipitation is not available in SFNO, but we can extrapolate that such large-scale pattern would lead to dry conditions in the SWUS, at odds with the observed anomaly. These results suggest that tropical forcing played an active role in driving the observed Rossby wave train and the ENP trough. This forecast period illustrates a window of opportunity, when S2S prediction skill exists for consequential and persistent weather conditions.

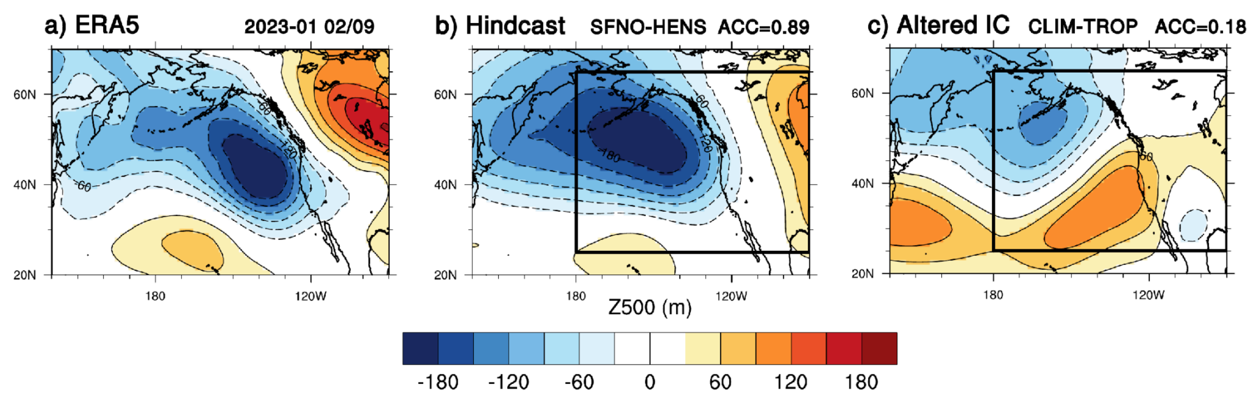

Figure 12. Impact of tropical variability on the week-2 forecast initialized on December 26th 2022, using SFNO-HENS. a) Observed Z500 anomaly (week-2 average, between January 2nd and January 9th 2023), from ERA5. b) Predicted anomaly in the standard SFNO-HENS hindcast. The spatial correlation (ACC) between the observed and predicted Z500 pattern in the black box domain is given in the top right. c) Same a but for the hindcast carried out after suppressing tropical variability from the initial condition.

In order to refine our understanding of the atmospheric dynamics leading to the persistent ENP trough and an accurate forecast in the original hindcast, we plot on Fig. 13 the difference in Z500 and V200 between the two hindcast ensemble means (original minus clim-trop) at different lead times. By design, no anomalies are present at day 1 except in the tropics since both hindcasts are initialized with exactly the same conditions in the extratropics. The two ensembles diverge with time and at week 1, different Z500 and V200 wave train patterns have developed in the subtropics, north of the forcing domain. At week 2, we retrieve the ENP trough discussed before, along with a well pronounced Rossby wave signal emanating from the subtropical western Pacific, better identified in the V200 anomalies through an alternation of northward and southward meridional wind anomalies. At week 3, the Rossby wave starts to decay upstream of the ENP trough, which still persists and maintains wet conditions over the Western US. This comparison between the original and altered hindcast ensemble means strongly suggests that tropical variability and the MJO phase 5-6 that occurred in December 2022 played a role for the subsequent persistence of the ENP trough in the following weeks. While such a wave response pattern is expected in physics-based models, it is encouraging to see a similar sensible dynamical pathway in the SFNO-HENS model. This supports previous work that has demonstrated realistic atmospheric physics behavior in ML weather forecast models (Hakim et al. 2024).

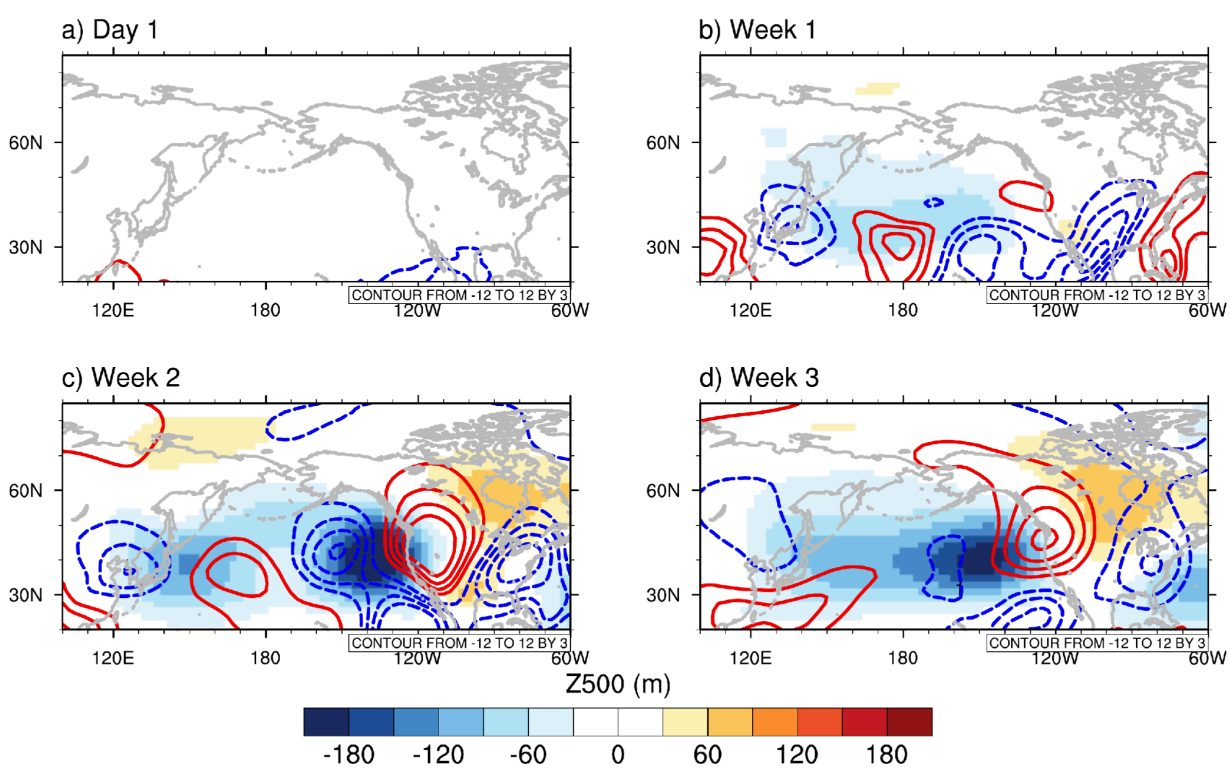

Figure 13. Evolution of the difference in Z500 (shading, m) and V200 (contours, m/s) anomalies between the original and clim-trop SFNO-HENS hindcasts for December 26th 2022 at : a) day 1 , b) week 1, c) week 2 and d) week 3. The anomalies shown here correspond to original minus clim-trop, highlighting the impact of including tropical anomalies, in particular the MJO phase 5, in the initial conditions.

4. Conclusion

In this study we have used three sets of S2S hindcasts from two ML models, SFNO-HENS and NeuralGCM, and a traditional physics-based GCM from ECMWF. Our goal is to evaluate the ML models and compare them to a state-of-the-art physics-based S2S forecasting model, focusing on the North Pacific atmospheric patterns and associated moisture transport over the western US during the cool season (October to March). We find that, in terms of S2S prediction of the North Pacific atmospheric patterns, the two models perform at least as well as ECMWF, i.e., all forecasting models have similarly limited aggregate S2S skill beyond week 3. Encouragingly the ML models have comparable skill to the ECMWF and are capable of competitive MJO prediction skill, as has been demonstrated using the Fuxi-S2S model (Chen et al. 2024). In addition, SFNO-HENS and NeuralGCM exhibit a longer persistence than ECWMF in the amplitude of the MJO signal in their ensemble mean forecasts. They also slightly better represent the propagation of the MJO across the Maritime Continent and they exhibit less wintertime climatological bias in the North Pacific mean flow and subtropical jet than ECMWF. While our 2004-2023 set of hindcasts overlap with the training periods of SFNO-HENS and NeuralGCM, we have shown that similar conclusions apply when using the 2020-2023 set of hindcasts that is outside their training periods.

Despite these promising characteristics, the prediction skill for mid-latitude weather patterns remains limited and comparable to ECMWF. As illustrated in Section 3.3, S2S prediction of western US climate conditions during the cool season is very challenging since it requires a precise prediction of North Pacific weather patterns and their background flow. The main source of predictability for such patterns is the MJO through the tropics-extratropics teleconnection, but even under strong MJO activity, the three S2S models do not demonstrate enough skill to accurately and consistently predict moisture transport (and precipitation, which was not available as an output field for this study) beyond week 2. Interestingly, the current ML models seem to hit the same prediction ceiling as ECMWF, one of the best dynamical models for the MJO and S2S prediction in general. The outstanding question is whether the ML models have led us to a fundamental limit in prediction skill that the dynamical models took decades to reach, or if they have not yet delivered their full potential for S2S prediction. Because they were developed for weather forecasting and not specifically for S2S prediction, it is likely that much progress is still to be made. An interesting feature of these models is that they are less subject to systematic model biases that can hamper the representation of large-scale climate teleconnections. Although the present study has not pointed out benefits due to less systematic biases, it is possible that future gain in MJO prediction skill may also translate into higher performance in the mid-latitudes, with more realistic mean flow and wave guides, as well as ocean coupling.

A straightforward perturbation experiment shows that ML models can also be useful tools for climate attribution and for our understanding of climate teleconnections. SFNO-HENS survived an intervention that removed *all* tropical variability from its initial condition – itself a test of robustness that it is not obvious ML models should pass, given the associated deviation from training data properties. We were able to generate a 58-member ensemble of an altered forecast to explore the role of tropical variability on a major S2S climate event during a window of long-term predictability over the ENP-WNA region. Our experiment suggests that the persistence of trough conditions in the ENP between December 26th 2022 and January 17th 2023 was predictable a few weeks in advance, and that tropical weather activity was critical to this predictability. Anomalies between the intervention and control experiments reveal expected wave train signatures that are known to bridge the tropics and extratropics on S2S time scales in physics-based simulations. Repeating similar perturbation experiments - which becomes easy under the fast inference of ML models - may lead to increased understanding of the conditions that generate a window of opportunity for S2S forecasting. The ease of use and flexibility of ML models also mean that they represent an appealing opportunity for future research.

To conclude, given their recent development to the modeling toolbox it is remarkable that ML models compare so well with state-of-the-art S2S forecast models. Rapid development means that future improvement in S2S prediction skill is possible (Eyring et al. 2024). Some of the models being developed include an ocean component along with the atmospheric model (Cresswell-Clay et al. 2024, Watt-Meyer et al. 2025, Zhou et al. 2025), similar to coupled ocean-atmosphere dynamical models. Various ML techniques are still under development and model training has mostly been focused on medium-range weather prediction, using high-frequency sub-daily input data. We hypothesize that future ML models designed for S2S time scales, for example using different input fields and temporal/spatial resolutions, may be effective in making progress towards more skillful S2S forecasting. This will have to be verified though, as it is also apparent that this is a complex problem that deals with chaotic and unpredictable variability which is inherent to the climate system.

Acknowledgements

YP, CD and GM were supported by the CA Department of Water Resources (Contract 4600013127). YP and GM were additionally in part supported by NSF (Grant PHY-2309135 to KITP). WC and AM were supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy (DOE) under Contract No. DE-AC02-05CH11231 and by the Regional and Global Model Analysis Program area within the Earth and Environmental Systems Modeling Program. MP was supported by NVIDIA Research. The research used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the DOE, under Contract No. DE-AC02-05CH11231. This research was supported in part by grant NSF PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP). The computation for this paper was supported in part by the DOE Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC) 2023-2024 award ‘Huge Ensembles of Weather Extremes using the Fourier Forecasting Neural Network’ to William Collins (LBNL) and the 2024-2025 award ‘Huge Ensembles of Weather Extremes using the Fourier Forecasting Neural Network’ to William Collins (LBNL).

Open Research

The Nvidia SFNO-HENS model version used in this study is openly available at Mahesh et al. (2025b). The Google NeuralGCM model can be downloaded from https://github.com/neuralgcm/neuralgcm. S2S hindcasts from ECMWF are available at https://apps.ecmwf.int/datasets/data/s2s, and the ERA5 reanalysis at Hersbach et al. (2023). The hindcast data used in this study is too large to be shared on an open research repository but they are available upon request to the main author.

Conflict of Interest Statement

The authors have no conflicts of interest to disclose.

5. References

Abhik, S., Zhang, C., & Hendon, H. H. (2023). The Indo-Pacific Maritime Continent barrier effect on MJO ensemble prediction. *Geophysical Research Letters*, 50, e2023GL105462. https://doi.org/10.1029/2023GL105462

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

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., and Anandkumar, A. (2023) Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere, https://doi.org/10.48550/ARXIV.2306.03838.

Brenowitz, N. D., Cohen, Y., Pathak, J., Mahesh, A., Bonev, B., Kurth, T., et al. (2025) A practical probabilistic benchmark for AI weather models. Geophysical Research Letters, 52, e2024GL113656. https://doi.org/10.1029/2024GL113656

Chen, L., Zhong, X., Zhang, F. *et al.* (2023) FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim Atmos Sci* **6**, 190. https://doi.org/10.1038/s41612-023-00512-1

Chen, L., Zhong, X., Li, H. *et al.* (2024) A machine learning model that outperforms conventional global subseasonal forecast models. *Nat Commun* **15**, 6425. https://doi.org/10.1038/s41467-024-50714-1

Chen, H., Constantinescu, E., Rao, V., & Stan, C. (2025) Improving the predictability of the Madden-Julian oscillation at subseasonal scales with Gaussian process models. Journal of Advances in Modeling Earth Systems, 17, e2023MS004188. https://doi.org/10.1029/2023MS004188

Cresswell-Clay, N., B. Liu, D.R. Durran, R.A. Moreno, Z. Liu, Z. Espinosa, and M. Karlbauer 2024: A Deep Learning Earth System Model for Stable and Efficient Simulation of the Current Climate, arXiv:2409.16247

de Burgh-Day C.O. and T. Leeuwenburg (2023) Machine learning for numerical weather and climate modelling: a review. Geoscientific Model Development 16(22), 6433–6477. https://doi.org/10.5194/Gmd-16-6433-2023

DeFlorio, M.J., Sengupta, A., Castellano, C.M., Wang, J., Zhang, Z., Gershunov, A., Guirguis, K., Niño, R.L., Clemesha, R.E., Pan, M. and Xiao, M. (2023) 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 2022-2023. Bull. Amer. Meteor. Soc., https://doi.org/10.1175/BAMS-D-22-0208.1

Dong, C., Y. Peings, and G. Magnusdottir (2024) How Do Forecast Model Biases Affect Large-Scale Teleconnections That Control Southwest U.S. Precipitation? Part I: S2S Models. J. Climate, 37, 5409–5426, https://doi.org/10.1175/JCLI-D-24-0051.1.

Eyring, V., Collins, W.D., Gentine, P. et al. (2024) Pushing the frontiers in climate modelling and analysis with machine learning. Nat. Clim. Chang. 14, 916–928. https://doi.org/10.1038/s41558-024-02095-y

Fortin, V., M. Abaza, F. Anctil, and R. Turcotte (2014) Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?. J. Hydrometeor., 15, 1708–1713, https://doi.org/10.1175/JHM-D-14-0008.1.

He, S., Li, X., DelSole, T., Ravikumar, P. & Banerjee, A. (2021) Sub-seasonal climate forecasting via machine learning: challenges, analysis, and advances. Proc. AAAI Conf. Artif. Intell. 35, 169–177.

Hersbach H, Bell B, Berrisford P, et al. (2020) The ERA5 global reanalysis. Q J R Meteorol Soc., 146: 1999–2049. https://doi.org/10.1002/qj.3803

Hersbach, H., Comyn-Platt, E., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N., Cagnazo, C., Cucchi, M. (2023) ERA5 post-processed daily-statistics on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.4991cf48

Jiang, X., Waliser, D.E., Gibson, P.B., Chen, G. and Guan, W., (2022) Why Seasonal Prediction of California Winter Precipitation Is Challenging. Bull. Amer. Meteor. Soc., 103(12), E2688-E2700.

Kent, C., Scaife, A.A., Dunstone, N.J. et al. (2025) Skilful global seasonal predictions from a machine learning weather model trained on reanalysis data. npj Clim Atmos Sci 8, 314. https://doi.org/10.1038/s41612-025-01198-3

Kiladis, G. N., J. Dias, K. H. Straub, M. C. Wheeler, S. N. Tulich, K. Kikuchi, K. M. Weickmann, and M. J. Ventrice, 2014: A Comparison of OLR and Circulation-Based Indices for Tracking the MJO. *Mon. Wea. Rev.*, **142**, 1697–1715, https://doi.org/10.1175/MWR-D-13-00301.1.

Kim, H., Janiga, M. A., & Pegion, K. (2019). MJO propagation processes and mean biases in the SubX and S2S reforecasts. *Journal of Geophysical Research: Atmospheres*, 124, 9314–9331. https://doi.org/10.1029/2019JD031139

Kumar A. and M. Chen (2020) Understanding skill of seasonal mean precipitation prediction over California during boreal winter and role of predictability limits. J. Climate, 33(14), 6141-6163.https://doi.org/10.1175/JCLI-D-19-0275.1.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023

Lang, S. et al. (2024) Aifs-ecmwf's data-driven forecasting system. arXiv preprint arXiv:2406.01465.

Lim, Y., S. Son, and D. Kim, 2018: MJO Prediction Skill of the Subseasonal-to-Seasonal Prediction Models. *J. Climate*, **31**, 4075–4094, https://doi.org/10.1175/JCLI-D-17-0545.1.

Lin, H., G. Brunet, and J. Derome, 2008: Forecast Skill of the Madden–Julian Oscillation in Two Canadian Atmospheric Models. Mon. Wea. Rev., 136, 4130–4149, https://doi.org/10.1175/2008MWR2459.1.

Liu, X., Peng, Y., Su, J., & Liu, X. (2024). To identify the forecast skill windows of MJO based on the S2S database. *Geophysical Research Letters*, 51, e2024GL109903. https://doi.org/10.1029/2024GL109903

Madden, R. A. , and P. R. Julian , 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, 28, 702– 708, https://doi.org/10.1175/1520-0469(1971)028<0702:DOADOI>2.0.CO;2.

Mahesh, A., Collins, W. D., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., Harrington, P., Kashinath, K., Kurth, T., North, J., O'Brien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J. (2025a) Huge ensembles – Part 1: Design of ensemble weather forecasts using spherical Fourier neural operators, Geosci. Model Dev., 18, 5575–5603, https://doi.org/10.5194/gmd-18-5575-2025.

Mahesh, A., Collins, W., Bonev, B. et al. (2025b) Huge ensembles part I design of ensemble weather forecasts with spherical Fourier neural operators; Huge ensembles part II properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators [Dataset]. Dryad. https://doi.org/10.5061/dryad.2rbnzs80n

Mariotti, A., C. Baggett, E. A. Barnes, E. Becker, A. Butler, and Coauthors (2020) Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. Bull. Amer. Meteor. Soc., 101, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Oskarsson J., T. Landelius, M. Deisenroth, F. Lindsten (2024) Probabilistic weather forecasting with hierarchical graph neural networks. Advances in Neural Information Processing Systems 37, 41577-41648.

Pathak, J. et al. (2022) Fourcastnet: a global data-driven high-resolution weather model using adaptive Fourier neural operators. Preprint at arxiv.org/abs/2202.11214.

Peings, Y., Y. Lim, and G. Magnusdottir (2022) Potential Predictability of Southwest U.S. Rainfall: Role of Tropical and High-Latitude Variability. J. Climate, 35, 1697–1717, https://doi.org/10.1175/JCLI-D-21-0775.1.

Peings, Y., C. Dong, and G. Magnusdottir (2024) How do model biases affect large-scale teleconnections that control Southwest US precipitation? Part II: seasonal models. J. Climate, https://doi.org/10.1175/JCLI-D-24-0287.1, in press.

Peings, Y., C. Dong, and G. Magnusdottir (2025) Seasonality in the predictive skill of Southwest United States precipitation in the seasonal forecast systems. Geophys. Res. Lett., 10.1029/2025GL115972.

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M. (2023) GenCast: Diffusion-based ensemble forecasting for medium-range weather, https://doi.org/10.48550/ARXIV.2312.15796.

Price, I., Sanchez-Gonzalez, A., Alet, F. et al. (2025) Probabilistic weather forecasting with machine learning. Nature 637, 84–90. https://doi.org/10.1038/s41586-024-08252-9

Ralph F.M. and M.D. Dettinger (2011) Storms, Floods, and the Science of Atmospheric Rivers. Eos Trans. AGU 92, 2009–2011.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., et al. (2024) WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16, e2023MS004019. https://doi.org/10.1029/2023MS004019

Sun, L., Hoerling, M.P., Richter, J.H., Hoell, A., Kumar, A. and Hurrell, J.W. (2022) Attribution of North American Subseasonal Precipitation Prediction Skill. Weather and Forecasting, 37(11), 2069-2085. https://doi.org/10.1175/WAF-D-22-0076.1

Ventrice, M. J., M. C. Wheeler, H. H. Hendon, C. J. Schreck III, C. D. Thorncroft, and G. N. Kiladis, 2013: A modified multivariate Madden–Julian oscillation index using velocity potential. Mon. Wea. Rev., 141, 4197–4210, doi:10.1175/MWR-D-12-00327.1.

Vitart, F. , 2017: Madden–Julian Oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, 143, 2210– 2220, https://doi.org/10.1002/qj.3079.

Wang B., G. Chen and F. Liu (2019) Diversity of the Madden-Julian Oscillation. Sci. Adv. 5, eaax0220, DOI:10.1126/sciadv.aax0220

Wang, J., M. J. DeFlorio, B. Guan, and C. M. Castellano (2023) Seasonality of MJO Impacts on Precipitation Extremes over the Western United States. *J. Hydrometeor.*, **24**, 151–166, https://doi.org/10.1175/JHM-D-22-0089.1.

Wang, J., DeFlorio, M.J., Gershunov, A. *et al.* (2024) Association of western US compound hydrometeorological extremes with Madden-Julian oscillation and ENSO interaction. *Commun Earth Environ* **5**, 314. https://doi.org/10.1038/s43247-024-01449-w

Watt-Meyer, O., Henn, B., McGibbon, J. et al. ACE2 (2025) Accurately learning subseasonal to decadal atmospheric variability and forced responses. npj Clim Atmos Sci 8, 205. https://doi.org/10.1038/s41612-025-01090-0

Weyn, J. A., Durran, D. R., Caruana, R. & Cresswell-Clay, N. (2021) Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. J. Adv. Model. Earth Syst. 13, e2021MS002502 .

Wheeler, M. C., and H. H. Hendon (2004) An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. Mon. Wea. Rev., 132 , 1917–1932.

Xiang, B., and Coauthors, 2022: S2S Prediction in GFDL SPEAR: MJO Diversity and Teleconnections. *Bull. Amer. Meteor. Soc.*, **103**, E463–E484, https://doi.org/10.1175/BAMS-D-21-0124.1.

Zhang, C., 2013: Madden–Julian oscillation: Bridging weather and climate. *Bull. Amer, Meteor. Soc.*, **94**, 1849–1870, https://doi.org/10.1175/BAMS-D-12-00026.1.

Zhou C., L. Chen, X. Zhong, B. Lu, H. Li, L. Wu, J. Wu, J. Hu, Z. Dou, P.-C. Hsu, X. Zhang (2025) A machine learning model for skillful climate system prediction. arXiv preprint, arXiv:2505.06269.