Supporting Information for "Improved MJO forecasts using the experimental global-nested GFDL SHiELD model."

Breanna L. Zavadoff[1], Kun Gao[2,3], Hosmay Lopez[4], Sang-Ki Lee[4], Dongmin Kim[1,4], and Lucas M. Harris[2]

[1]Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida

[2]NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

[3]Cooperative Institute for Modeling the Earth System, Princeton University, Princeton, New Jersey

[4]NOAA/Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida

Contents of this file

Corresponding author: Breanna L. Zavadoff, Cooperative Institute for Marine and Atmospheric Studies, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149, USA. (bzavadoff@earth.miami.edu)

Text S1: Computing MJO Indices

To isolate the MJO signal from the background a combined EOF analysis of OLR, u850, and u200 from 1979-2019 is performed (Lin et al., 2008; Rashid et al., 2011). Before the EOF analysis can be executed, each of the fields are preprocessed to extract the intraseasonal anomalies from the daily unfiltered data. This involves removing both the seasonal cycle and interannual variability by subtracting the daily climatology and previous 120-day running mean of the subsequent anomalies, respectively, from the raw data. Each of the fields are then latitudinally averaged over an equatorial belt from 15°S - 15°N and normalized by their zonally averaged standard deviation to ensure that the contribution of each variable to the EOF is equally weighted. The leading two EOFs are then calculated and their corresponding spatial structures (Figure S3) are used to derive the real-time multivariate MJO index 1 (RMM1) and 2 (RMM2) for MJO identification.

Observational data, which matches the output time of the model forecasts, is used for verification of the MJO forecasts. Observed and forecast RMM1 and RMM2 values are extracted by projecting their corresponding intraseasonal anomalies onto the combined EOF patterns derived from observations. The daily climatologies of observed zonal wind and OLR calculated from 1979-2019 are also used to remove the seasonal cycle in both the observations and model data. Because each of the model forecasts only extend out to 40 days, 120 previous days of data are not available through the model alone for the removal of the interannual variability. To circumvent this data shortage observations are used to fill in the subsequent missing days preceding the forecast (Xiang et al., 2015; Vitart et al., 2017; Harris et al., 2020), a method that has the added benefit of preventing initial

condition shock at the start of the forecast. For example, for day n of the forecast, 120 - n + 1 days of observational data and n - 1 days of model data are used to calculate the previous 120-day running mean.

Text S2: Deterministic Evaluation Metrics

As noted in Section 1, MJO prediction skill and predictability are sensitive to the amplitude and phase errors of MJO forecasts, which are calculated following the metrics outlined in Rashid et al. (2011). MJO amplitude for observations and forecasts are defined as:

$$AMP_{obs}(t) = \sqrt{RMM1(t)^2 + RMM2(t)^2}, \tag{1}$$

and

$$AMP_{pred}(t,\tau) = \sqrt{RMM1(t,\tau)^2 + RMM2(t,\tau)^2}, \tag{2}$$

respectively, where t is time and τ is the forecast lead time. Amplitude error, which measures how strong or weak the predicted amplitude is compared to the observed amplitude, can then be calculated as a function of lead time as:

$$ERR_{amp}(\tau) = \frac{1}{N}\Sigma_{t=1}^{N}[AMP_{pred}(t,\tau) - AMP_{obs}(t)], \tag{3}$$

where N is the number of predictions and a positive (negative) $ERR_{amp}(\tau)$ corresponds to a predicted amplitude stronger (weaker) than that in observations.

Phase error, which measures the angular difference between the observed and predicted phase in the RMM phase-space diagram, can be calculated as a function of lead time as:

$$ERR_{phase}(\tau) = \frac{1}{N}\Sigma_{t=1}^{N} \tan^{-1}[\frac{RMM1(t)RMM2(t,\tau)-RMM2(t)RMM1(t,\tau)}{RMM1(t)RMM1(t,\tau)+RMM2(t)RMM2(t,\tau)}], \quad (4)$$

where a positive (negative) $ERR_{phase}(\tau)$ indicates a faster (slower) propagating MJO in predictions than observations.

A limitation of these error calculations is that if model errors are positive in some cases but negative in others, they could cancel each other out when averaged. To prevent information loss and more clearly quantify model errors, squared amplitude and squared phase errors are calculated in lieu of amplitude and phase errors (Lim et al., 2018). Their calculations are listed below:

$$ERR^2_{amp}(\tau) = \frac{1}{N}\Sigma_{t=1}^{N}[AMP_{pred}(t,\tau) - AMP_{obs}(t)]^2, \quad (5)$$

and

$$ERR^2_{phase}(\tau) = \frac{1}{N}\Sigma_{t=1}^{N} \tan^{-1}[\frac{RMM1(t)RMM2(t,\tau)-RMM2(t)RMM1(t,\tau)}{RMM1(t)RMM1(t,\tau)+RMM2(t)RMM2(t,\tau)}]^2. \quad (6)$$

MJO prediction skill is commonly measured as a function of lead time using metrics such as the bivariate anomaly correlation coefficient (ACC) and bivariate root-mean-squared error (RMSE). Using the formulas presented in Lin et al. (2008) and Rashid et al. (2011) these can be calculated as:

$$ACC(\tau) = \frac{\Sigma_{t=1}^{N}[RMM1(t)RMM1(t,\tau)+RMM2(t)RMM2(t,\tau)]}{\sqrt{\Sigma_{t=1}^{N}[RMM1(t)^2+RMM2(t)^2]}\sqrt{\Sigma_{t=1}^{N}[RMM1(t,\tau)^2+RMM2(t,\tau)^2]}}, \quad (7)$$

and

$$RMSE(\tau) = \sqrt{\frac{1}{N}\Sigma_{t=1}^{N}([RMM1(t) - RMM1(t,\tau)]^2 + [RMM2(t) - RMM2(t,\tau)]^2)}. \quad (8)$$

MJO prediction is considered skillful until the forecast lead time at which $ACC(\tau)$ falls below a threshold value of 0.5. In the case of RMSE, skillful MJO prediction is achieved when the predicted $RMSE(\tau)$ remains below a threshold value of $\sqrt{2}$.

Text S3: Probabilistic Evaluation Metrics

The reliability diagram summarizes several probabilistic metrics by showing the full joint distribution of observed frequency $\bar{o}_k$ and and forecast probability f_k in terms of calibration refinement (Murphy & Winkler, 1977; Atger, 1999). The equation below summarizes the Brier score (BS; Toth et al., 2003; Kharin & Zwiers, 2003) as a function of its components, reliability (left), resolution (center), and uncertainty (right), which are central in the reliability diagram analysis.

$$BS = \frac{1}{n}\Sigma_{k=1}^{I}(N_k[f_k - \bar{o}_k]^2) - \frac{1}{n}\Sigma_{k=1}^{I}([\bar{o}_k - \bar{o}]^2) + \bar{o}(1-\bar{o}), \quad (9)$$

where $\bar{o}$ is the overall sample climatological frequency, I is a discrete forecast sample, N_k is the number of instances each forecast f_k is being verified at each discrete I bin, and n is is the number of forecasts-events pairs ($n = \Sigma_{k=1}^{I}(N_k)$). For forecasts probabilities from 0.0 to 1.0 at 0.1 intervals, $I = 11$ discrete forecast bins.

By breaking the BS into its components, we can asses the particular strengths and weaknesses in each of the forecast systems. For instance, reliability measures the calibration or conditional bias in the forecast such that in a reliable, or well-calibrated, forecast the reliability term should be very small or near zero, which is depicted as a reliability curve that lies near the diagonal. Resolution measures the ability of the forecast to discern subsample forecast periods with different relative frequencies of the event. Ideally, we want the resolution term to be large (i.e., the forecast sorts the observations into subsamples having different relative frequencies than climatology). This is shown in the reliability diagram as a line that spans a large range of observed frequencies (i.e., large range in the vertical), and is far away from the horizontal line labeled "no resolution". Since uncertainty depends only on the observations it is equal in both model runs as the forecast has no saying in how uncertain the truth is. The uncertainty term in our case for a tercile forecast equals 2/9.

The Brier skill score (BSS) can be obtained from the three previously discussed components presented in equation 9 (Toth et al., 2003; Kharin & Zwiers, 2003) as:

$$BSS = \frac{Resolution - Reliability}{Uncertainty} \tag{10}$$

ROC is a method to evaluate the probabilistic skill of a prediction system based on a 2 x 2 contingency table (Swets, 1973; Mason & Graham, 1999), which contains four possible outcomes: a hit (h) if a warning is issued and the event occurs; a miss (m) if no warning is issued for an event that occurs; a false alarm (f) if a warning is issued and no event occurs; and a correct rejection (c) if no warning is issued and no event occurs.

The probabilistic skill of a prediction system can be evaluated by comparing hit rates to false alarm rates. The hit rate is the proportion of events for which warnings were issued correctly; it provides an estimate of the probability that an event is correctly predicted. On the other hand, the false alarm rate is the proportion of non-events for which warnings were issued incorrectly. Equations for both the hit rate and false alarm rate are depicted below:

$$Hit\ Rate = \frac{h}{h+m} \tag{11}$$

$$False\ Alarm\ Rate = \frac{f}{f+c} \tag{12}$$

.

The contingency table can be constructed and further used to plot the ROC curve, which compares hit rates and false alarm rates for a range of warning threshold values. The ROC score is the area between the ROC curve and the diagonal (i.e., equal hit and false alarm rates). It ranges between -1 and 1, and measures the utility of the forecasts compared to the utility of a perfect forecast (Swets, 1973; Mason & Graham, 1999). A ROC score of zero generally indicates no forecast skill relative to random guesses from the climatological probability density function, while a ROC score above 0.4 indicates that the forecast fairly well discriminates between events and non-events better than a random guess from the climatological probability density function, so that the system is more likely to correctly predict an actual event than to issue a false alarm.

References

Atger, F. (1999). The skill of ensemble prediction systems. *Mon. Wea. Rev.*, *127*(9), 1941–1953.

Harris, L., Zhou, L., Lin, S.-J., Chen, J.-H., Chen, X., Gao, K., ... others (2020). GFDL SHiELD: A unified system for weather-to-seasonal prediction. *J. Adv. Model. Earth Syst.*, *12*(10), e2020MS002223.

Kharin, V. V., & Zwiers, F. W. (2003). Improved seasonal probability forecasts. *J. Climate*, *16*(11), 1684–1701.

Lim, Y., Son, S.-W., & Kim, D. (2018). MJO prediction skill of the subseasonal-to-seasonal prediction models. *J. Climate*, *31*(10), 4075–4094.

Lin, H., Brunet, G., & Derome, J. (2008). Forecast skill of the Madden–Julian oscillation in two Canadian atmospheric models. *Mon. Weather Rev.*, *136*(11), 4130–4149.

Mason, S. J., & Graham, N. E. (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, *14*(5), 713–725.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *J. R. Stat. Soc.*, *26*(1), 41–47.

Rashid, H. A., Hendon, H. H., Wheeler, M. C., & Alves, O. (2011). Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. *Clim. Dyn,*, *36*(3), 649–661.

Swets, J. A. (1973). The Relative Operating Characteristic in Psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, *182*(4116), 990–1000.

Toth, Z., Talagrand, O., Candille, G., & Zhu, Y. (2003). *Probability and ensemble. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Jolliffe IT and Stephenson, DB.* Wiley.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., ... others (2017). The subseasonal to seasonal (S2S) prediction project database. *Bull. Am. Meteorol. Soc.*, *98*(1), 163–173.

Xiang, B., Zhao, M., Jiang, X., Lin, S.-J., Li, T., Fu, X., & Vecchi, G. (2015). The 3–4-week MJO prediction skill in a GFDL coupled model. *J. Climate*, *28*(13), 5351–5364.

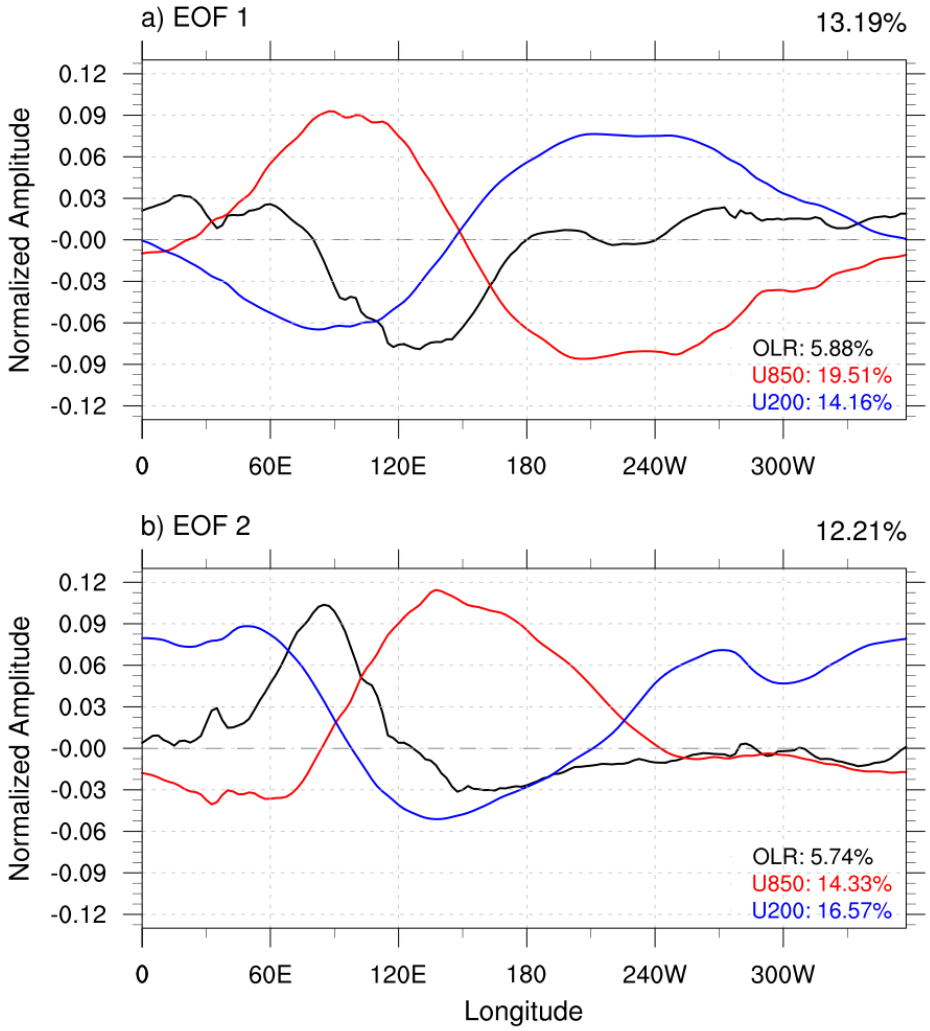

Figure S1. Combined (a) EOF 1 and (b) EOF 2 of the latitudinally averaged (15°S - 15°N) intraseasonal anomalies of observed OLR (black curves), u850 (red curves) and u200 (blue curves). The variances explained by each EOF are provided in the top right corner, while the variances of each individual variable are presented in the bottom right corner.

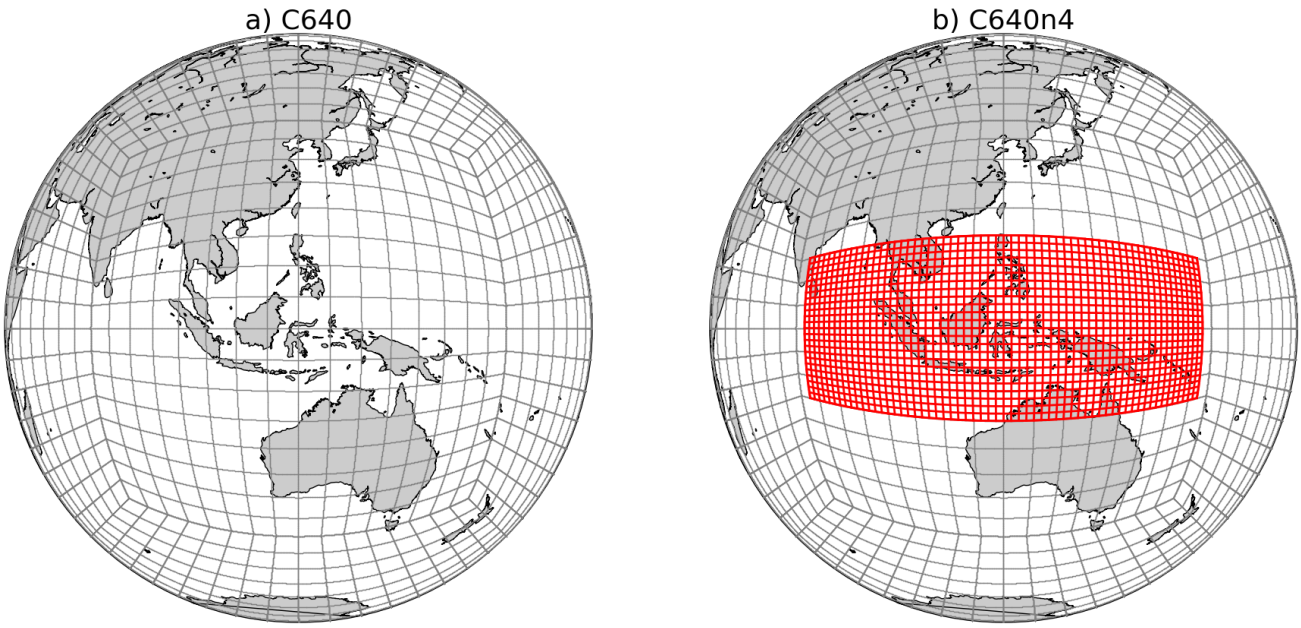

Figure S2. The SHiELD (a) control (C640) and (b) nested (C640n4) grid configurations used in this study. Red grids in (b) represent the two-way nested grids. Each plotted cell represents 48 x 48 actual grid cells. The horizontal resolution is about 16 km in the global tile and 4 km in the nest (a refinement ratio of 4 is applied).

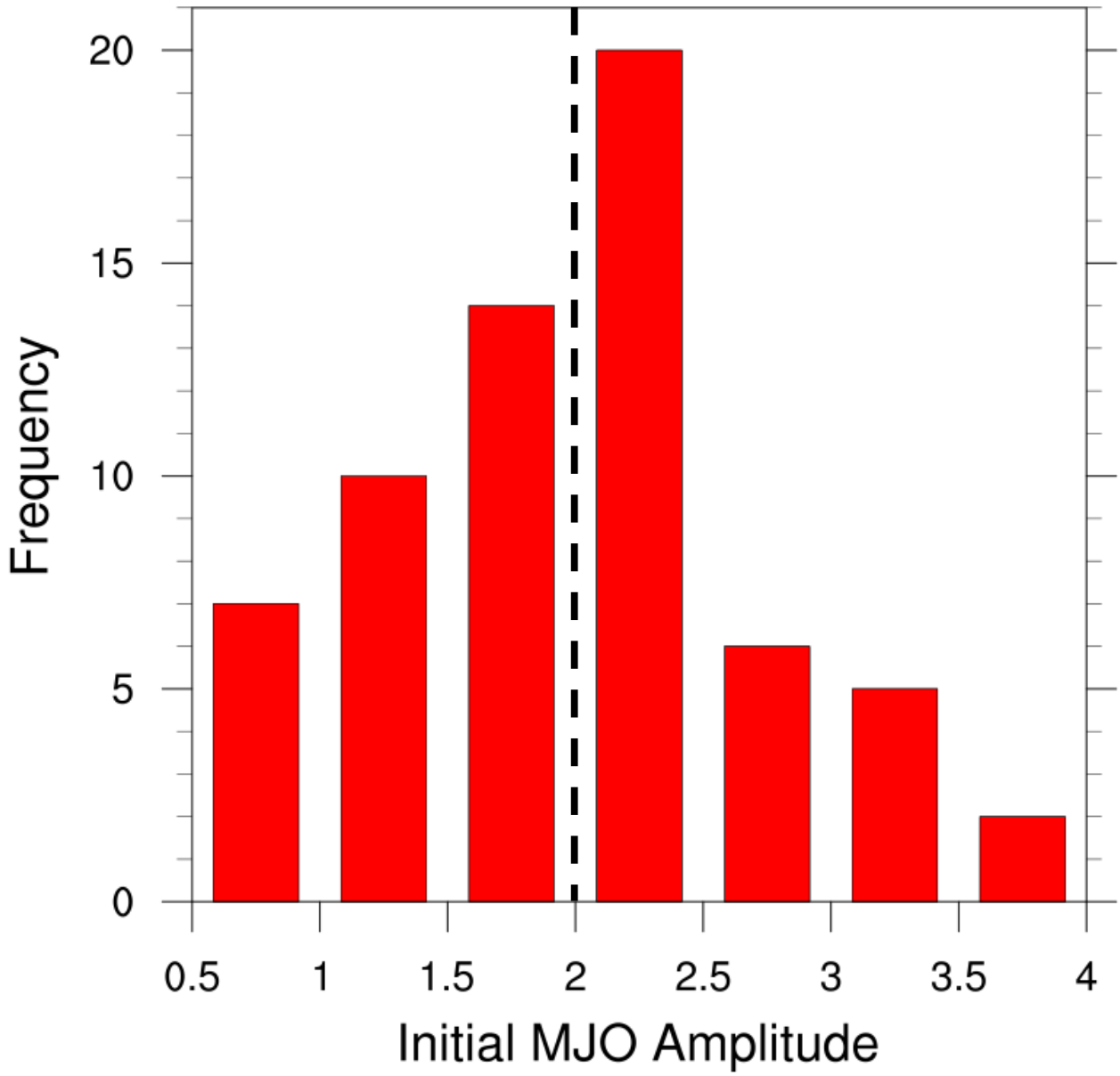

Figure S3. Histogram of MJO amplitude at initialization for each of the 64 SHiELD forecasts. The dashed black line denotes the amplitude (2.0) chosen as a threshold to separate neutral/weak and moderate MJOs at initialization from strong MJOs at initialization.

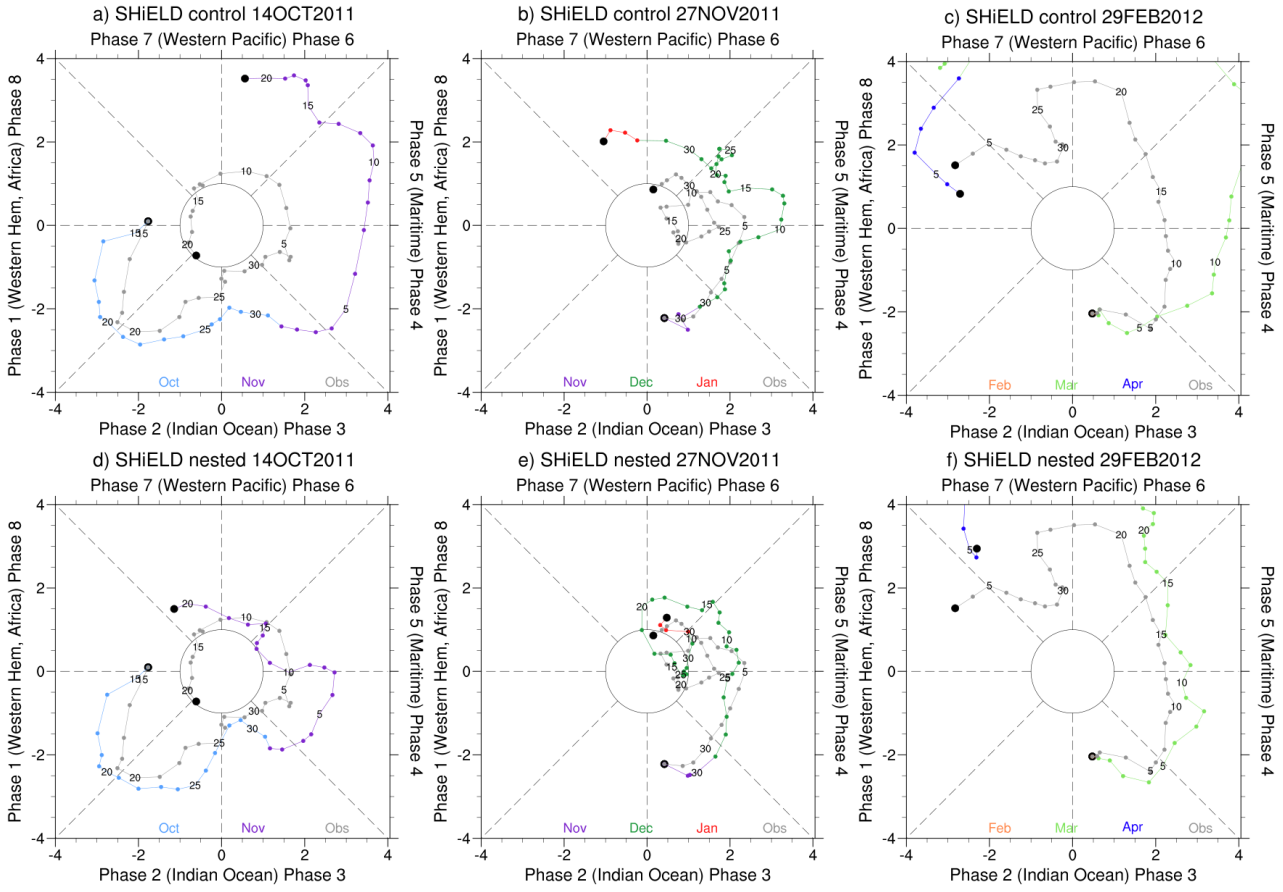

Figure S4. RMM diagrams for control (top) and nested (bottom) 40 day forecasts initialized on (a,d) October 14, 2011, (b,e) November 27, 2011, and (c,f) February 29, 2012. Dots represent the predicted amplitude and phase for each day following the initialization time and are color coded by month, while the corresponding observed amplitude and phase are plotted in grey. An MJO with an amplitude less than 1 is considered to be in the neutral (inactive) phase, which is depicted as the area within the circle in the center of the RMM-phase diagram. Days of the month are labeled in increments of 5. Open and closed circles denote the beginning and end of the RMM time series, respectively.

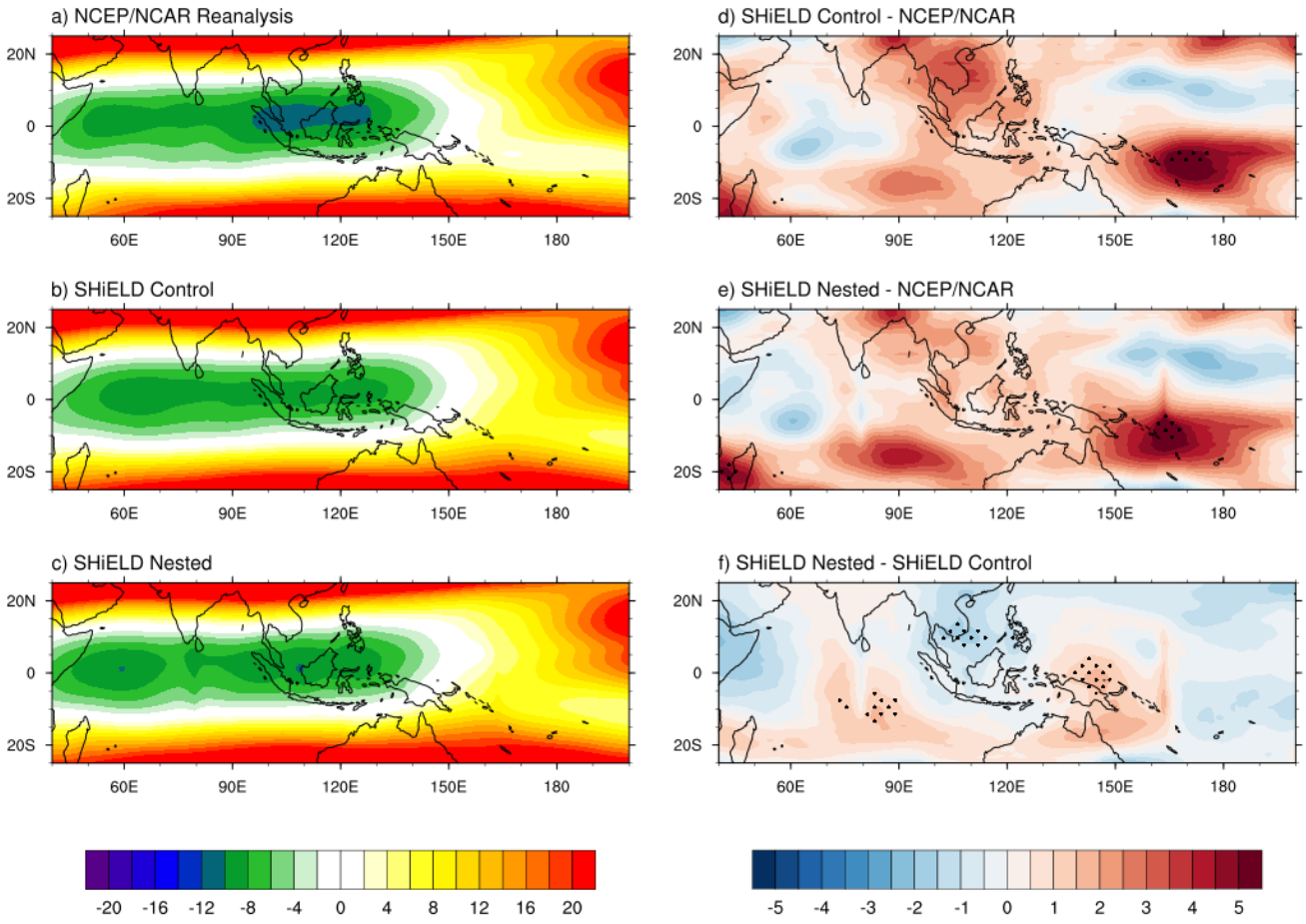

Figure S5. Daily u200 (m s^{-1}; shading according to colorbar) averaged for (a) observations time matched to all 64 (b) SHiELD control, and (c) nested 40 day forecasts. Differences (m s^{-1}; shading according to colorbar) between observations and the SHiELD control and nested forecasts are presented in (d) and (e), respectively, while (f) shows how the anomalies differ between the two forecasts. Stippling in (d,e,f) indicates regions where these differences are statistically significant at the 95% confidence level according to a Student's t test.

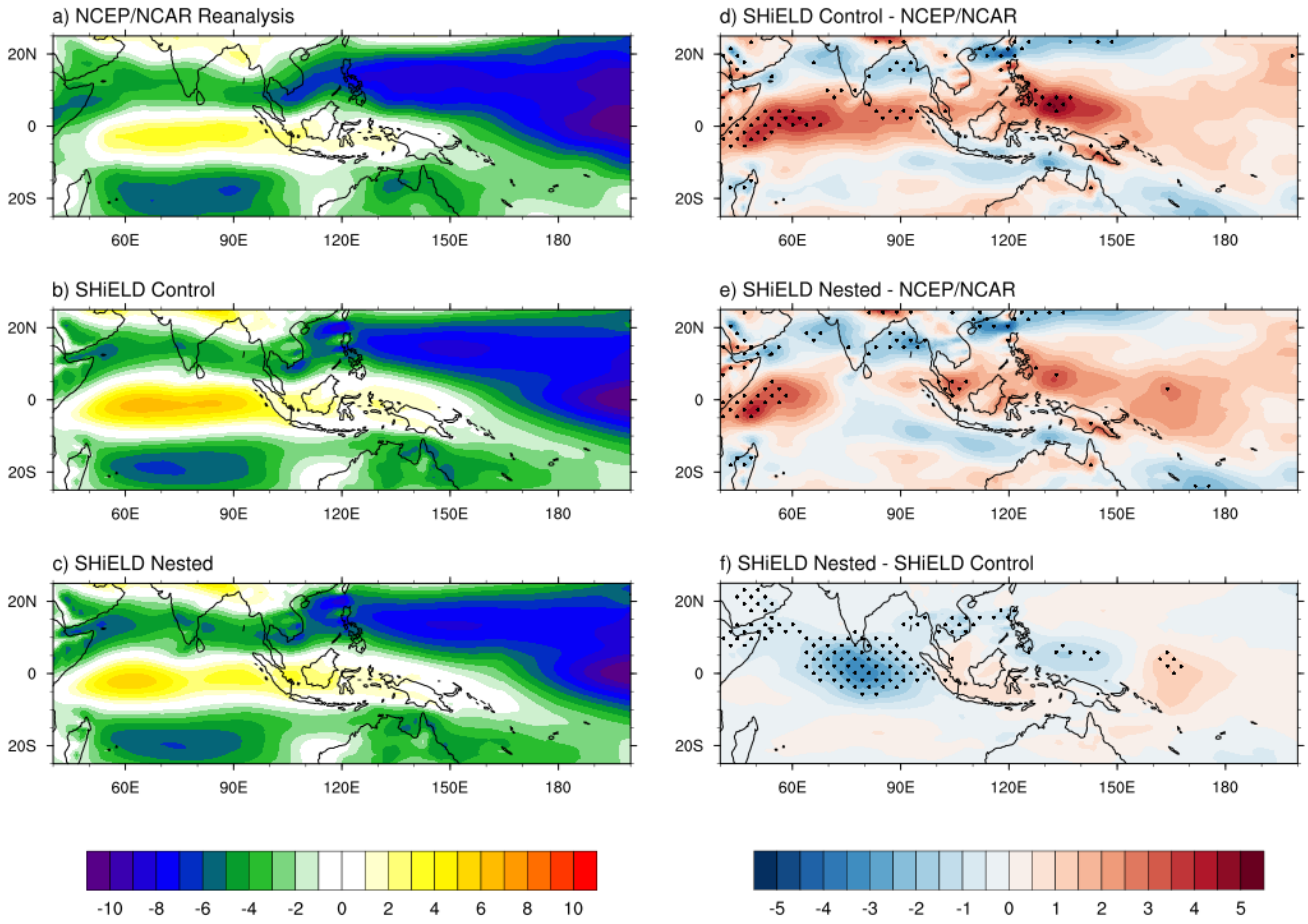

Figure S6. As in Figure S3, but for u850.

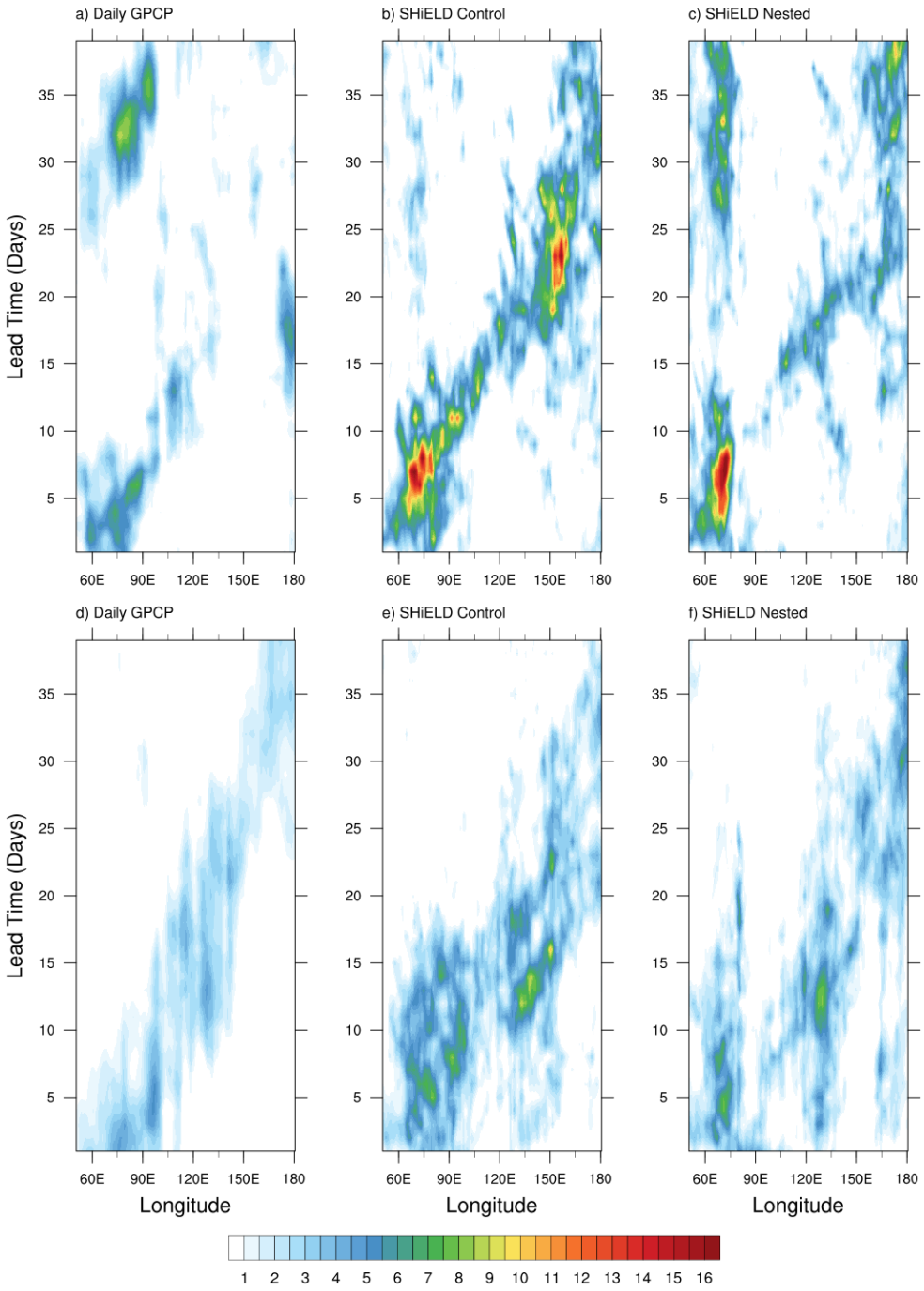

Figure S7. Composite Hovmöller diagrams of latitudinally averaged (15°S - 15°N) daily intraseasonal anomalies of precipitation rate (mm day^{-1}; shaded according to color bar) as a function of lead time and longitude for (a,d) observations time matched to the 35 (b,e) SHiELD control and (c,f) nested forecasts initialized with a moderate (top) or strong (bottom) MJO in phase 2 or 3.

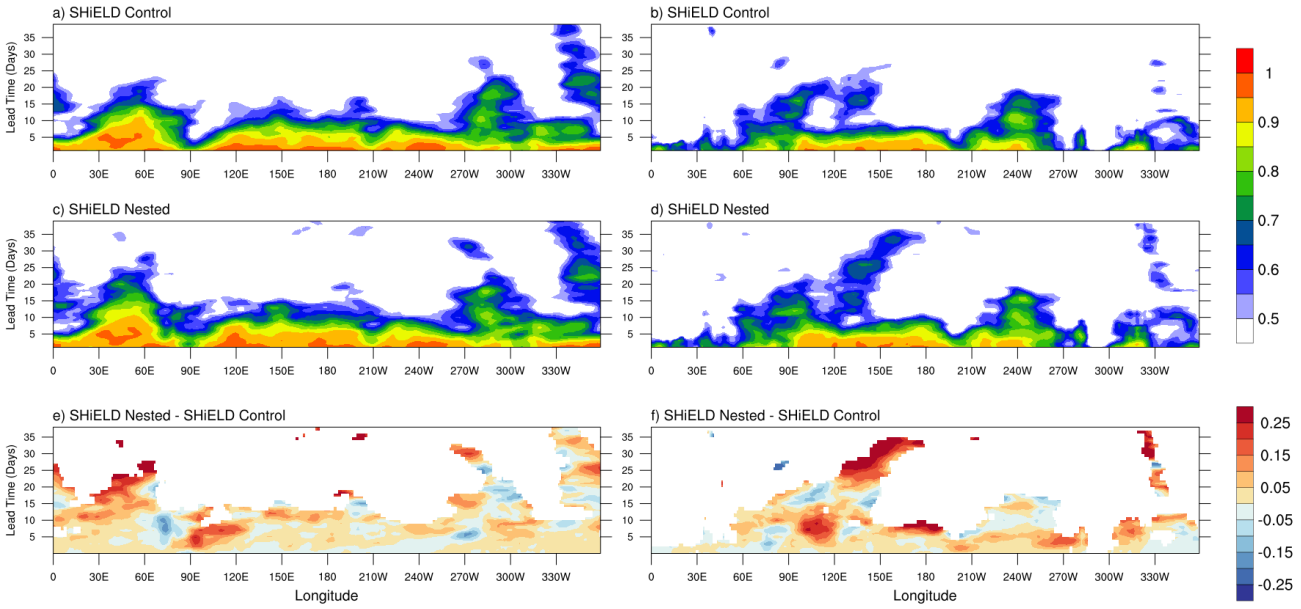

Figure S8. ACC (shaded) for latitudinally averaged (15°S - 15°N) daily averaged intraseasonal anomalies of u200 (left) and u850 (right) as a function of lead time and longitude for all (a,b) SHiELD control and (c,d) nested forecasts initialized with an active MJO and (e,f) their differences.

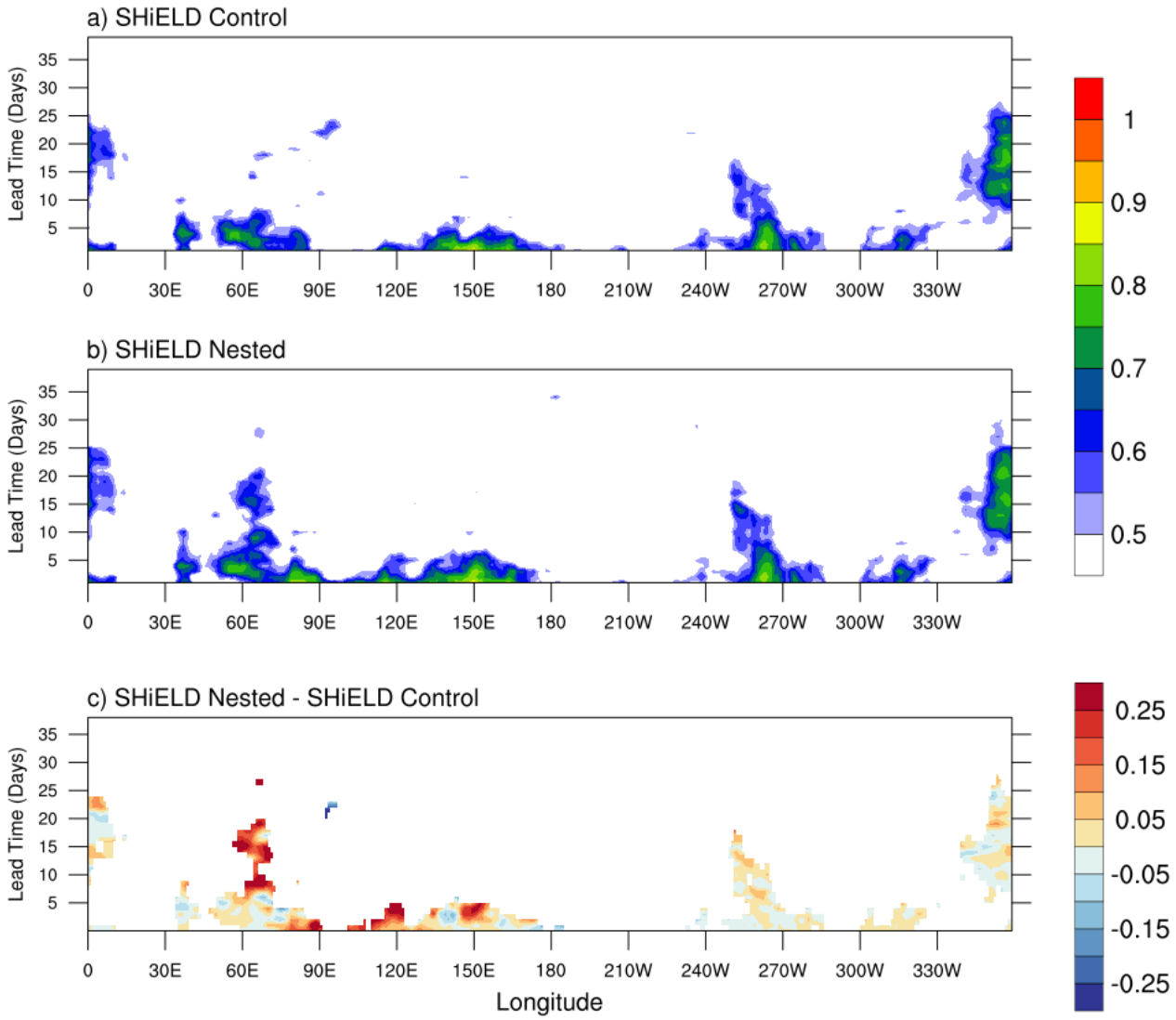

Figure S9. As in Figure S8, but for OLR.