Subseasonal predictions of tropical cyclone occurrence and ACE in the S2S dataset

Chia-Ying Lee*

Lamont-Doherty Earth Observatory, Columbia University Palisades, NY

Suzana J. Camargo

Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY

Fréderic Vitart

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

Adam H. Sobel

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY

Joanne Camp

Met Office Hadley Centre, Exeter, UK

Shuguang Wang

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

Michael K. Tippett

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

Qidong Yang

Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY

**Corresponding author address:* Chia-Ying Lee, Lamont-Doherty Earth Observatory, Columbia University, 61 route, 9W, Palisades, NY, 20964

E-mail: cl3225@columbia.edu

ABSTRACT

Probabilistic tropical cyclone (TC) occurrence, at lead times of week 1 to 4, in the Subseasonal to Seasonal (S2S) dataset are examined here. Forecasts are defined over 15^{o} in latitude $\times$ 20^{o} in longitude regions, and the prediction skill is measured using the Brier skill score with reference to climatological reference forecasts. Two types of reference forecasts are used: a seasonally constant one and a seasonally varying one with the latter used for forecasts of anomalies from the seasonal climatology. Models from the European Centre for Medium-Range Weather Forecasts (ECMWF), Australian Bureau of Meteorology, and Météo-France/Centre National de Recherche Météorologiques have skill in predicting TC occurrence four weeks in advance. In contrast, only the ECMWF model is skillful in predicting the anomaly of TC occurrence beyond one week. Errors in genesis prediction largely limit models' skill in predicting TC occurrence. Three calibration techniques, removing the mean genesis and occurrence forecast biases, and a linear-regression method, are explored here. The linear-regression method performs the best and guarantees a higher skill score when applied to the in-sample dataset. However, when applied to the out-of-sample data, especially in areas where the TC sample size is small, it may reduce the models' prediction skill. Generally speaking, the S2S models are more skillful in predicting TC occurrence during favorable Madden–Julian oscillation phases. Lastly, we also report accumulated cyclone energy predictions skill using the Ranked probability skill score.

1. Introduction

Tropical cyclone (TC) predictions are evaluated differently at different time-scales. Short–term (weather prediction time-scale) track and intensity forecasts are usually verified against best-track records at the same time via mean absolute error (e.g., DeMaria et al. 2014). Seasonal storm predictions, on the other hand, are often verified over a basin using correlations of observed and forecast TC counts or accumulated cyclone energy (ACE; e.g., Chen and Lin 2013). Only recently have global weather prediction systems started to generate forecasts at subseasonal time-scales (Vitart et al. 2010). Therefore, there are no widely accepted standards for verifying and evaluating subseasonal TC predictions (Camargo et al. 2019). Similarly to short–term weather predictions, Elsberry et al. (2011) and Tsai et al. (2013) verified subseasonal predictions from the European Centre for Medium-Range Weather Forecasts (ECMWF) by comparing the forecast and observed TCs at times and locations at which the storms were very close to each other. Yamaguchi et al. (2015) defined forecasts of weekly storm occurrences over $0.5^{o} \times 0.5^{o}$ grids. Vitart et al. (2010), Camp et al. (2018), and Gregory et al. (2019) examined weekly storm occurrence over 15^{o} in latitude $\times$ 20^{o} in longitude boxes with 7.5^{o} and 10^{o} buffer ranges. Others, such as Li et al. (2016), Lee et al. (2018) and Gao et al. (2019) considered basin–wide TC activity.

Verification methods are, on one hand, limited by the skill of the forecasts, and on the other hand, they reflect, implicitly, what information is expected from the forecasts. One guiding principle in designing verifications is to consider the desired socio-economic value of the forecasts. For example, which kind of information would be useful for disaster preparedness with two to three weeks lead-time? This information could be used, e.g., to plan the distribution and storage of emergency supplies or deploy emergency personnel (Vitart and Robertson 2018). Forecasts of basin–wide TC activity clearly do not provide the ideal type of forecast information at these time-scales as they do not provide the kind of regional information that is essential for regional disaster preparedness. Conversely, due to the limitations of current prediction systems, it is not reasonable to expect reliable forecasts of the exact time, location or intensity of landfalling TCs weeks in advance. The verification method used by Vitart et al. (2010), Camp et al. (2018), and Gregory et al. (2019) is therefore a reasonable compromise, since it balances the capability of current weather prediction systems with the needs of the user on subseasonal time-scales.

Many studies have shown that forecasts of TC position and genesis can have skill beyond 10 days. Elsberry et al. (2011) and Tsai et al. (2013) found that the ECMWF ensembles were able to predict most of the named typhoons' tracks out to 4 weeks in advance in the 2009 and 2010 Northwestern Pacific typhoon seasons, although there was a 50% false alarm rate. Vitart et al. (2010) showed that a calibration that removes the mean forecast bias could increase the ECMWF's track predictions skill in the Southern Hemisphere TC basins from two to four weeks. Similar results are found in two recent papers (Camp et al. 2018; Gregory et al. 2019), which evaluated reforecasts and real-time forecasts of the Australian Bureau of Meteorology seasonal forecasting system (ACCESS–S1) over the Southern Oceans. In the subseasonal to seasonal (S2S) dataset (see Section 2), Lee et al. (2018) showed that reforecasts run by six operational centers can predict genesis weeks in advance.

TCs have a strong climatological seasonal cycle, and subseasonal variability of TCs is defined as the anomaly (fluctuation) that deviates from that cycle. Thus, accurately predicting TCs at subseasonal time-scales requires models to forecast both the seasonal cycle and anomalies. Generally speaking, global models can predict the seasonal cycle reasonably well because they are good at simulating the low-frequency large-scale atmospheric and oceanic patterns. These large-scale patterns contribute to the predictability of the TC seasonal cycle (Camargo and Barnston 2009; Zhan et al. 2012). The main source of predictability for subseasonal TC variability, on the other hand, is the Madden Julian Oscillation (MJO). Models tend to be more skillful both when the MJO signal is strong during the initial forecast time (e.g., Belanger et al. 2010), and when the MJO is in phases that are favorable to TCs in the basin at the forecast verification time (e.g., Jiang et al. 2012). Tropical waves, such as Kelvin waves and African easterly waves, also influence TC genesis on subseasonal scales (e.g., Ventrice et al. 2011, 2012; Schreck 2015). The models' ability to forecast the large-scale environmental patterns associated with El Niño–Southern Oscillation, the Atlantic Meridional Mode (e.g., Belanger et al. 2010; Li et al. 2016), as well as extra-tropical-tropical interactions (Zhang and Wang 2019) influence subseasonal TC predictability as well.

The promising results mentioned above (Vitart et al. 2010; Camp et al. 2018; Gregory et al. 2019; Lee et al. 2018) are based on verifications that credit models for capturing the seasonal cycle and the subseasonal variability. That is to say, forecasts are evaluated against seasonally constant climatological forecasts as a reference. To understand if the S2S models have skill at predicting genesis anomalies, Lee et al. (2018) further used seasonally varying climatological forecasts as a reference (no credit for capturing the seasonal cycle), and showed that the ECMWF model is the only one that has skill in predicting genesis anomalies at 2–3 weeks lead-time in most TC basins. Vitart et al. (2010) also discuss the ECMWF model's prediction skill in southern hemisphere TC basins in comparison with seasonally varying climatological forecasts.

The present study is a continuation of Lee et al. (2018) which evaluated the S2S models' performance in predicting basin-wide TC formation. In contrast to Lee et al. (2018), we focus here on (1) the S2S models' performance in predicting regional TC occurrence (i.e., genesis and subsequent locations) and Accumulated Cyclone Energy (ACE); (2) applying the various calibration methods, including the one used in Camp et al. (2018), to the forecasts and discussing their impact; and (3) investigating the dependence of the prediction skill on the MJO as characterized by two MJO indices, namely the Real-Time Multivariate Index (RMM; Wheeler and Hendon 2004) and the Real-Time Outgoing-Longwave-Radiation (OLR) MJO index (ROMI; Kiladis et al. 2014). Data and methods for model evaluation are described in Section 2. The models' performance in storm occurrence is in Section 3, followed by discussion of the calibration schemes in Section 4. We report the dependence of model skill on MJO in Sections 5 and the models' performance in predicting ACE in Section 6, followed by Conclusions in Section 7.

2. Methods

a. The S2S dataset and observations

We consider the same S2S reforecasts as in Lee et al. (2018), based on coupled, global general circulation models run by six operational centers: the Australian Bureau of Meteorology (BoM), the China Meteorological Administration (CMA), the ECMWF, the Japan Meteorological Agency (JMA), the Météo-France/Centre National de Recherche Météorologiques (MetFr), and the National Centers for Environmental Prediction (NCEP). Basic characteristics of these six reforecasts are shown in Table 1 and further details of the S2S dataset are described in Vitart et al. (2017).

TCs in the S2S models are tracked daily using the methodology of Vitart and Stockdale (2001). The tracker defines a storm center at a local minimum sea-level pressure where (1) a local vorticity maximum ($> 3.5 \times 10^{-5}$ s^{-1}) at 850 hPa is nearby, (2) a local maximum in the vertically averaged temperature (warm core, $> 0.5\,^{o}$C) in between 250–500 hPa is within a distance (in any direction) equivalent to 2^{o} latitude, (3) the two locations detected from (1) and (2) are within a distance equivalent to 8^{o} latitude, and (4) a local maximum thickness between 1000–200 hPa can be identified within a distance equivalent to 2^{o} latitude. Additionally, a detected storm must last at least two days to be included in our analysis. The same criteria apply to TCs in all ocean basins.

Observations of tropical cyclone tracks are from the HURDAT2, produced by the National Hurricane Center (Landsea and Franklin 2013), and from the Joint Typhoon Warning Center (Chu et al. 2002). Both best-track datasets include 1-min maximum sustained wind, minimum sea level pressure (not used in this study), and storm location every 6 hours. Following the conventional definitions (Fig. 1), the TC basins are: Atlantic (ATL), northern Indian Ocean (NI), western North Pacific (WNP), eastern North Pacific (ENP), southern Indian Ocean (SIN, 0-90^{o}E), Australia (AUS, 90-160^{o}E), and southern Pacific (SPC, east of 160^{o}E). For each basin, we only use forecasts that are initialized during their respective TC seasons: May to November for ATL and WNP, May to October for ENP, April to June and September to November for NI, November to April for SIN and AUS and December to April for SPC.

b. Defining forecasts

Following Camp et al. (2018), we subdivide global TC basins into 20^{o} in longitude $\times$ 15^{o} in latitude boxes (centers are labeled by circles in Fig. 1). Each box overlaps with its neighboring boxes by 10^{o} and 7.5^{o} in the longitude and latitude direction, respectively. A grid on the border of the two basins belongs to the one on the east and/or on the north side. Thus, the 20^{o} $\times$ 15^{o} boxes centered at the equator belong to the Northern Hemisphere basins. Then, we define occurrence forecasts by the fraction of all the ensemble members that contain a TC (ensemble frequency) in individual grids for each of the six models. Similarly, we also define the accumulated cyclone energy forecast (ACE) by the fraction of ensemble members that have weekly ACE exceeding specified thresholds (Section 2d) over each box.

Forecasts are evaluated at daily time resolution with a weekly (7 days) window, starting from day 4. In other words, prediction skill at day 4 contains forecasts from day 1 to day 7, prediction skill at day 5 includes forecasts from day 2 to day 8, and so on. Sometimes we also use ‘week’ to describe the forecasts, such that "week 1 forecasts" refers to forecasts containing data from days 1 to 7, "week 2 forecasts" are forecasts from days 8 to 14, and so on. As an example, Figs. 2a and 2b show week-2 occurrence forecasts (in dots) and the gridded occurrence forecasts (in shading) from a ECMWF forecast initialized on Aug. 20, 2005. The observed storm occurrence and ACE are calculated following the same procedure as described above. For convenience, we refer to each of these $20^{\circ} \times 15^{\circ}$ boxes as a "region", and thus "regional" refers to the analyses done over individual boxes.

c. Defining the MJO

Two real-time MJO indices are considered. The first one is the RMM, which is calculated using intraseasonal zonal winds at 200 and 850 hPa and observed outgoing longwave radiation (OLR; Wheeler and Hendon 2004; Gottschalck et al. 2010; Vitart 2017). The second MJO index is ROMI, an OLR based index, calculated from observed intraseasonal OLR anomalies (Kiladis et al. 2014). Wang et al. (2018) showed that ROMI better represents northward propagation of the boreal summer intraseasonal oscillation than RMM.

d. Skill scores

1) Brier skill score

The Brier skill score (BSS) is used to assess the skill of a probabilistic forecast of TC occurrence relative to a climatological forecast. The Brier Score (BS) is defined as:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{1}$$

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{2}$$

where N is the total number of forecasts, o_i is the i$^{\text{th}}$ observation. p_i is the predicted probability of TC occurrence for the i$^{\text{th}}$ forecast, defined as:

$$p_i = \frac{1}{M}\sum_{j=1}^{M} P_{i,j}, \tag{3}$$

where M is the number of ensemble members, $P_{i,j}$ is the TC occurrence prediction from the j$^{\text{th}}$ ensemble member for the i$^{\text{th}}$ forecast. $P_{i,j}$ and o_i are 0 for no storm and 1 for 1 or more storm occurrences during the forecast period. Thus, the BS is the mean squared probability forecast error. When analyzing the models' performance over individual $20^{\text{o}} \times 15^{\text{o}}$ regions, N in Eq. 1 is the number of forecasts used. When evaluating models' performance in a basin, N is the product of the number of forecasts used and the number of regions in that basin. For example, for evaluating the ECMWF model in the Atlantic basin, N is 64554, which consists of 1218 forecasts across 53 regions. Note that the forecast number, 1218, is different to the one (2058) listed in Table 1, because we only use data during the Atlantic hurricane season.

The BS_{ref} is similar to the BS, but for a reference forecast based on the observed climatology. The observed climatology is calculated using observations over the same period and at the same temporal resolution as the S2S model data. In this study, two climatologies are used. The first one is the seasonally varying climatology at monthly time resolution. The second one is a constant, seasonal-mean climatology. When a model is skillful compared to the climatology, the BSS is positive. For convenience, we refer to the BSS for the monthly–varying climatology as BSS_{m}, and the BSS for the seasonal mean, constant climatology as BSS_{c} hereafter. BSS_{c} can be interpreted as the model skill in predicting the absolute TC occurrence, including seasonality. On the other hand, BSS_{m} evaluates the model's ability to predict the anomalies in TC activity that deviate from the seasonal cycle. The values of BSS_{m} are lower than those of BSS_{c} because its reference forecast (monthly-varying mean) is more informative.

2) Ranked probability skill score

To verify ACE predictions (Section 6), we use the ranked probability skill score (RPSS). RPSS is a squared-error score for categorical forecasts. The cumulative forecasts and observations (P_c and O_c and the ranked probability score (RPS) are denoted as:

$$P_c = \sum_{j=1}^{c} p_j, c = 1, ..., C \tag{4}$$

$$O_c = \sum_{j=1}^{c} o_j, c = 1, ..., C \tag{5}$$

$$\mathrm{RPS} = \sum_{c=1}^{C} (P_c - O_c)^2 \tag{6}$$

where C is the number of forecast categories and p_j is the forecast probability of the storm intensity falling in the j$^{\mathrm{th}}$ category. The observed probability o_j is 1 if the observations fall in the j$^{\mathrm{th}}$ category and 0 otherwise. The RPS is the sum of the squared differences between the cumulative probabilities P_c and O_c. RPS is oriented so that smaller values indicate better forecasts. A correct forecast with no uncertainty has an RPS of 0. Similar to the BSS, the RPSS compares the average RPS to that of a reference forecast:

$$\mathrm{RPSS} = 1 - \frac{\sum_{i=1}^{N} \mathrm{RPS_i}}{\sum_{i=1}^{N} \mathrm{RPS_{ref_i}}}. \tag{7}$$

We again have two reference forecasts: the first uses the seasonal-mean climatology, the second uses the monthly-varying seasonal climatology. They are referred to as $\mathrm{RPSS_c}$ and $\mathrm{RPSS_m}$, respectively. The RPSS is sensitive to the definitions of the forecast categories. Because TCs are rare events, more than 95% of the observations have ACE of 0, and the categories should not be equally spaced, Here, we define 6 categories, and the first category is for ACE = 0. The other 5 categories correspond to the 0, 20, 40, 60, 80 quantiles of the observed distribution of non-zero ACE.

3. TC occurrence prediction

TC occurrence predictions are evaluated here from both regional and basin-wide perspectives. From a basin-wide perspective, the ECMWF model is skillful in predicting TC occurrence (BSS_c) at all TC basins up to four weeks in advance (Fig. 3). The BoM and MetFr models also have positive BSS_c at weeks 1–4 in most TC basins. The JMA model is skillful up to 10 days in all TC basins except the NI. In terms of predicting seasonal anomalies (BSS_m), the ECMWF model is skillful up to 2–3 weeks in the WNP, ENP, SIN, and SPC, and 1–2 weeks in the ATL and AUS. Other S2S models have limited skill: the BoM model has positive BSS_m in the SIN and SPC at week 1–2, the MetFr model is skillful in the SIN and AUS at week 1, and the JMA model is skillful in the SIN and SPC at week 1. The CMA and NCEP models do not have skill in predicting TC occurrence globally. The basin-wide prediction skill scores shown in Fig. 3 do not always reflect the models' performance on the regional scale. For example, while the ECMWF model is skillful in predicting TC occurrence at weeks 1–2 globally, Fig. 4a shows that the model has negative BSS_c in parts of AUS (Timor Sea, Arafura Sea, Banda Sea). Similarly, ECMWF model has no skill in predicting TC activity over the Arabian Sea at week 2, but it has an overall positive BSS_c in NI. In contrast, the model is not skillful in predicting TC occurrence anomaly in the NI, but is skillful in the Bay of Bengal (Fig. 4b).

The TC occurrence prediction skill scores in the S2S models are qualitatively consistent with those for genesis prediction shown in Lee et al. (2018); both suggest that the ECMWF is the most skillful model and can predict storm activity anomalies with respect to monthly climatology up to 2–3 weeks in advance. This similarity is not surprising as the prevailing circulation associated with the genesis location may influence the subsequent track pattern. Still, it is interesting to know how a model's occurrence prediction skill is limited by its genesis prediction skill. To address this question, we conduct an additional BSS analysis using the forecasted storms forming within 500 km and $\pm$ 3 days of the observed TC genesis locations. We keep cases in which the observed genesis is captured by at least one ensemble member. In other words, we are looking at BSS conditioned on the genesis having occurred correctly in at least one of the ensemble members in the forecast ($\mathrm{BSS_{m|TC}}$). One can also think of $\mathrm{BSS_{m|TC}}$ as a measure of occurrence forecast skill only with the genesis element removed.

Using the ECMWF forecasts, Fig. 5 shows that the positive $\mathrm{BSS_{m|TC}}$ values (gray lines) can last much longer than the positive $\mathrm{BSS_m}$ values (black lines). In the NI and the three southern basins $\mathrm{BSS_{m|TC}}$ is positive from week 1–4 while $\mathrm{BSS_m}$ is only positive up to week 2. The increase in the prediction skill is smaller (a few days to one week) in the WNP, ENP, and ATL. It is well known that TCs are steered by their ambient steering flow (Dong and Neumann 1986) and storm motion forecasts depend upon skillful prediction of the environmental wind field (Galarneau and Davis 2012). While S2S models' performance on steering flow has not yet been examined in the literature (to the best of our knowledge), the difference between $\mathrm{BSS_m}$ and $\mathrm{BSS_{m|TC}}$ values implies that the ECMWF model may be able to predict the steering flow weeks in advance. An interpretation of Fig. 5 is that the biggest challenge for subseasonal storm occurrence predictions is to forecast genesis well. Vitart and Robertson (2018) also mentioned that if a model can predict genesis correctly, there is a potential for skillful prediction of the subsequent track even at long lead times, at least for long-lived storms. In practice, however, we will not be able to identify which genesis (and subsequent track) predictions are reliable in advance.

4. Calibration

Next, we discuss whether the occurrence prediction skills, particularly as measured by the $\mathrm{BSS_m}$, can be further improved through a post-processing calibration. Three techniques are explored here:

removing the mean genesis bias, removing the mean occurrence bias, and the linear regression method. In principle, the calibration parameters should be developed using a subset of the entire data set, known as the "training" or "in-sample" data, and evaluated with the remainder of the data set, known as the "testing" or the "out-of-smaple" data. Here, we apply a calibration method to the whole dataset and examine the impact of the method in the in-sample dataset. If the results are promising, we will test the method by separating the dataset into in-sample and out-of-sample groups. As shown in this section, we only conduct out-of-sample data evaluation for the linear regression method.

a. Removing the mean genesis bias

The $\mathrm{BSS}_{\mathrm{m|TC}}$ results suggest that there is potential to improve the models' occurrence prediction skill by removing the mean genesis bias – that is, by correcting the mean forecast genesis rate to match the observed one:

$$p_{i|gen} = p_i \times r_{gen} \tag{8}$$

$$r_{gen} = \frac{\sum_{i=1}^{N} o_{i,gen}}{\sum_{i=1}^{N} p_{i,gen}}. \tag{9}$$

Here, the genesis rate is defined as the number of genesis events per day, and the mean genesis bias is the ratio (r_{gen}) between the observed genesis rate ($\sum_{i=1}^{N} o_{i,gen}$) and model simulations ($\sum_{i=1}^{N} p_{i,gen}$) over each region. This ratio is multiplied by the forecast occurrence probability to get the calibrated occurrence probability, $p_{i|gen}$. r_{gen} is a function of lead times and regions. The modified forecasts are then used for calculating the Brier Skill Score for anomalies ($\mathrm{BSS}_{m|\mathrm{gen}}$):

$$\mathrm{BS}_{m|\mathrm{gen}} = \frac{1}{N}\sum_{i=1}^{N}(p_{i|gen} - o_i)^2 \tag{10}$$

$$\mathrm{BSS}_{m|\mathrm{gen}} = 1 - \frac{\mathrm{BS}_{m|\mathrm{gen}}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{11}$$

Eq. 11 is the BSS conditioned on the same genesis rate. Compared to the $\mathrm{BSS_m}$ (black lines in Fig. 5), $\mathrm{BSS_{m|gen}}$ (green dashed lines in Fig. 5) has positive skill in NI and AUS for almost a week longer. In other words, in these two basins the mean genesis biases reduces the ECMWF model occurrence prediction skill by one week. $\mathrm{BSS_{m|gen}}$ and $\mathrm{BSS_m}$ are closer in the WNP, SIN, and SPC than in other basins. In the ENP and ATL, $\mathrm{BSS_{m|gen}}$ values are even smaller than $\mathrm{BSS_m}$.

b. Removing the mean occurrence bias

Another common approach for calibrating occurrence forecasts is to remove the mean occurrence biases (e.g., Vitart et al. 2010; Camp et al. 2018). Similar to Eq. 8, the calibrated probability ($p_{i|mean}$) is derived by multiplying the forecast probability by a ratio, but now it is the ratio (r_{mean}) of mean observed probability and the mean forecast probability:

$$r_{mean} = \frac{\sum_{i=1}^{N} o_i}{\sum_{i=1}^{N} p_i}. \tag{12}$$

r_{mean} is also a function of lead times and regions. We follow Camp et al. (2018) and restrict r_{mean} to values between 0.5 and 2. For example, a r_{mean} value of 3 is changed to 2, and a r_{mean} value 0.02 is changed to to 0.5. This restriction is done to avoid unreasonably large $p_{i|mean}$ at areas where the sample size (of TCs) in the forecasts is too small, and to avoid forcing the model to predict very small or 0 probability values at regions where the observed sample TC size is small. As mentioned in the Introduction, removing the mean occurrence biases increases the ACCESS–S1's occurrence prediction skill from week 2 to week 5 (Camp et al. 2018; Gregory et al. 2019). Spatial maps of $\mathrm{BSS}_{m|\mathrm{mean}}$ from ECMWF week 2 forecasts are used to show the impact of this calibration method. The ECMWF week 2 $\mathrm{BSS_{m|mean}}$ has positive values in the NI, ENP, SIN, AUS, and SPC (Fig. 6a). When compared to $\mathrm{BSS_m}$ (Fig. 4b), the calibrated score ($\mathrm{BSS_{m|mean}}$) increases the prediction skill in the Bay of Bengal, western SIN, AUS, and SPC (Fig. 6b). On the basin-wide scale, $\mathrm{BSS_{m|mean}}$

(green solid lines in Fig. 5) improves the skill of predicting NI, SIN, and AUS storms at all lead times ($\mathrm{BSS_m}$) but degrades the skill of predicting WNP, ENP, and ATL storms. In the SPC, it has positive impact on $\mathrm{BSS_m}$ before day 10 lead time but negative impact afterwards.

The results above show that removing the mean occurrence bias does not always have a positive impact on the forecast. This is consistent with Camargo et al. (2019) who showed that this calibration method improves ACCESS–S1 southern hemisphere skill scores for long-leads in 2017-18 but degrades the skill in 2018-19. Because this calibration method has been used in several studies, we conduct further analysis to understand how it works. First of all, we decompose Eqs. 1 and 2 following Murphy and Winkler (1992); Murphy (1988):

$$\begin{aligned}
\mathrm{BS} &= \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \\
&= \left\langle (p-\overline{p}+\overline{p}-o-\overline{o}+\overline{o})^2 \right\rangle \\
&= \left\langle [(p-\overline{p})-(o-\overline{o})+(\overline{p}-\overline{o})]^2 \right\rangle \\
&= \left\langle (p-\overline{p})^2 \right\rangle + \left\langle (o-\overline{o})^2 \right\rangle + \left\langle (\overline{p}-\overline{o})^2 \right\rangle - \left\langle 2(p-\overline{p})(o-\overline{o}) \right\rangle \\
&= \sigma_p^2 + \sigma_o^2 + (\mu_f - \mu_o)^2 - 2\sigma_p\sigma_o\gamma_{p,o},
\end{aligned} \tag{13}$$

where σ^2 is the variance, μ is the mean, γ is the correlation coefficient, and $\overline{()}$ and $\langle\rangle$ represent averaging over N forecasts. The skill score BSS can then be rewritten as:

$$\begin{aligned}
BSS &= 1 - \frac{BS}{BS_{ref}} \\
&= 1 - \frac{\sigma_p^2 + \sigma_o^2 + (\mu_f - \mu_o)^2 - 2\sigma_p\sigma_o\gamma_{p,o}}{\sigma_o^2} \\
&= 2\frac{\sigma_p}{\sigma_o}\gamma_{p,o} - (\frac{\sigma_p}{\sigma_o})^2 - (\frac{\mu_f - \mu_o}{\sigma_o})^2 \\
&= \gamma_{p,o}^2 - (\gamma_{p,o} - \frac{\sigma_p}{\sigma_o})^2 - (\frac{\mu_f - \mu_o}{\sigma_o})^2,
\end{aligned} \tag{14}$$

in which the three terms on the right-hand-side represent the potential skill (correlations), conditional bias, and unconditional bias (Bradley et al. 2008). To gain higher values of BSS (better prediction skill), a calibration scheme needs to increase the correlation between forecasts and observations, and/or reduce the conditional and unconditional biases. Removing the mean occurrence biases reduces the unconditional bias to zero. However, it also changes the value of σ_p and therefore does not guarantee a smaller conditional bias. Consequently, Eq. 8 could potentially result in lower values of BSS.

When will $\mathrm{BSS}_{\mathrm{m|mean}}$ guarantee higher values of $\mathrm{BSS}_{\mathrm{m}}$? To obtain the necessary conditions for increasing BSS values, we compare BS and $\mathrm{BS}_{\mathrm{m|mean}}$ ($\mathrm{BS}_{\mathrm{m|mean}}$ should be smaller than BS) and obtain the following:

$$r_{mean} \leq \frac{2\overline{po}}{\overline{p^2}} - 1; \text{if } r_{mean} >= 1 \tag{15}$$

$$r_{mean} > \frac{2\overline{po}}{\overline{p^2}} - 1; \text{if } r_{mean} < 1. \tag{16}$$

When a model has a positive mean bias, the ratio r_{mean} between the mean observed probability and the mean modeled probability has to be smaller than the threshold $\frac{2\overline{po}}{\overline{p^2}} - 1$. On the other hand, when the model is biased low, r_{mean} needs to be larger than the threshold. Figures 7a and 7b show the spatial distributions of r_{mean} and the threshold. The colorbars in both figures are designed such that for the calibration method to have positive impact, the regions that are red (blue) in Fig. 7a need to be redder (bluer) in Fig. 7b. The comparison is shown in Fig. 7c in which regions where the ECMWF TC occurrence prediction skill can be improved by the calibration method are labeled in red and those where it cannot are labeled in blue. The red and blue areas in Fig. 7c are similar to the reddish and bluish areas in Fig. 6b. Figure 7c also suggests that removing the mean occurrence bias seems to work better when the model mean occurrence forecast is biased low (gray dots in Fig. 7). While not shown here, the blue-red pattern shown in Fig. 7c is model dependent. The impact of the restriction of r (0.5 to 2) on the calibrated forecast skill score is not investigated here but is an interesting question that should be further explored.

c. Linear regression method

Removing the mean occurrence biases does not always work because it corrects only the mean probabilistic forecast error, but not the mean squared probability forecast error, which is what BSS measures. While one can argue that it is better to use mean error as an evaluation metric instead, BSS is a conventional metric for evaluating the performance of probabilistic forecasts. Therefore, we explore a linear regression-based technique (van den Dool et al. 2016) that minimizes the mean square error. In this approach, the calibrated probabilistic forecast is:

$$p_{i|linear} = a \times p_i + b, \tag{17}$$

where a ($a = \gamma_{p,o}\frac{\sigma_o}{\sigma_p}$) is the regression coefficient and b is the intercept. It is noted that $p_{i|linear}$ may be negative or greater than 1 despite the forecast probability being defined between 0 and 1. In this study, we set all the negative $p_{i|linear}$ to 0; and 1 if it is greater than 1. For the in-sample data, Eq. 17 can remove the unconditional biases and minimize the conditional biases. The resulting Brier Skill Score is therefore the potential skill, $\gamma^2_{p,o}$. Figure 6c shows that the week 2 $\mathrm{BSS_{m|linear}}$ for ECMWF model is positive everywhere except the North Atlantic; the ECMWF's week 2 forecasts of TC occurrence anomaly in the North Atlantic are negatively correlated to observations. The differences between $\mathrm{BSS_{m|linear}}$ and the $\mathrm{BSS_m}$ (Fig. 6d), as expected, show that Eq. 17 improves the ECMWF model's prediction skill globally. At the basin scale, $\mathrm{BSS_{m|linear}}$ also outperforms $\mathrm{BSS_m}$ (comparing the pink lines to the black lines in Fig. 5).

We further examine the impact of applying Eq. 17 to out-of-sample data. To do so, the first two-third of ECMWF forecasts (from 1995 to 2009) are used as training data and the remaining one-third (from 2010 to 2015) are the testing data. When applied to out-of-sample data, Eq. 17 does not guarantee higher prediction skill scores (Fig. 6e and 6f). This is especially true in regions where the training data are insufficient to capture the statistics of model's forecast errors, and thus the derived a and b do not minimize the mean square error of the testing data. In central North Pacific and part of North Atlantic, $BSS_{m|linear,\ out}$ is smaller than BSS_m. At the basin scale, $BSS_{m|linear,out}$ (red lines in Fig. 5) still improves the ECMWF week 2 occurrence prediction skill. The improvement is small in the WNP and SIN, though. The basin-wide $BSS_{m|linear,out}$ for all models are shown in Fig. 8. Compared to Fig. 3, applying Eq. 17 seems to improve the S2S models' occurrence prediction skill in all basins. The improvement is especially evident in the SIN where all the six S2S models are skillful at week 1 with ECMWF, BoM, MetFr and JMA having skill at week 2. A more sophisticated way to minimize the mean square error is to use logistic regression, which will be explored in the future.

The three calibration techniques used here suggest that calibrating subseasonal, probabilistic TC predictions is not straightforward. A method that works for in-sample data may not work for out-of-sample data, especially regional scales. Further effort is necessary to develop a comprehensive calibration method.

5. Dependence of occurrence prediction skill on the MJO

As discussed in the Introduction, the predictability of subseasonal TC activity is commonly related to the MJO phase and amplitude (e.g., Belanger et al. 2010; Jiang et al. 2012). To systematically assess the dependence of the S2S models' prediction skill on the MJO, we compare the lag relationships of TC occurrence and Brier Skill Scores to the MJO phases defined by RMM and ROMI (Section 2c). To make sure the relationships are not contaminated by the calibration methods, we use the original BSS_c and BSS_m here.

We start by examining the observed MJO–TC genesis relationship from these two indices using the candy-plot analysis (Lee et al. 2018), a two–dimensional histogram of genesis probability as a function of MJO phases and basins. In Figure 9, the TC basins are arranged so that the convectively active MJO phases (with black circles) are aligned diagonally. The probability of genesis in convectively active (favorable) MJO phases is higher (red colors) than in suppressed phases (blue colors). The ROMI-candy diagram shows more dark red and dark blue circles than does the RMM-candy diagram, indicating that ROMI is sharper and better represents the MJO's modulating influence on TC genesis. The favorable MJO phases defined by ROMI are shifted to the east by one phase in the WNP, SPC, and ENP, compared to those defined by RMM. The lag–analysis between TC occurrence and MJO (Fig. 10) shows the eastward shift of the favorable MJO phases from RMM to ROMI as well. This shift may be related to the fact that RMM mostly represents the MJO circulation (Straub 2012; Ventrice et al. 2013), while ROMI represents the MJO convection (Kiladis et al. 2014). Another possibility is the existence of a shift in the geographic locations of the MJO phases associated defined using ROMI compared with those defined using RMM. However, Kiladis et al. (2014) showed that the maximum correlation between OMI (the non-realtime version of ROMI) and RMM occurs at lags -2 to 4 days, and thus these two indices do represent MJO phases with similar (while not exactly the same) geographic location.

While not perfect, the candy-plot analyses (Fig. 11) suggest that the S2S models capture the shifts of the favorable MJO phases. Except in the JMA model, the pattern correlations between simulated and observed MJO–TC relationships are higher when MJO is defined by RMM than by ROMI. This is an indication that S2S models better simulate the influence of MJO wind signal on TC frequency than they simulate the influence of the MJO convection signal. The CMA and MetFr models are the two extreme cases because their simulations of the ROMI defined MJO–TC relationship yields correlations with observations that are only 11% and 5%, while in the case of RMM the correlation coefficients are 41% and 42%, respectively.

Next, we analyze the contribution of the MJO to S2S models' prediction skill by grouping the forecasts by MJO phase. Using $\mathrm{BSS_c}$ as an example, first we calculate the difference of $BSS_{c|mjo}$, i.e. the BSS_c conditioned on the MJO phase, and BSS_c: $\delta BSS_{c|mjo} = BSS_{c|mjo} - BSS_c$. Positive $\delta BSS_{c|mjo}$ means that forecasts initialized at the MJO phase mjo contribute positively to BSS_c, which is calculated using the full dataset. Then, we use lag analysis to examine the MJO–$\mathrm{BSS_c}$ relationship.

Figure 12 shows that the positive $\delta\mathrm{BSS_c}$ (red shading) is in phase with the positive TC activity anomalies (black contour) in the ECMWF simulations, when the MJO is defined by ROMI. Similar results are found when MJO is defined by RMM (not shown). In other words, the ECMWF model has better skill in predicting total TC occurrence during favorable MJO phases than unfavorable ones. The pattern correlation coefficients between the relationships of MJO–TC and MJO–$\mathrm{BSS_c}$ in the seven TC basins from the six S2S models are shown in Table 2. In most cases, the S2S models have positive correlation coefficients, meaning that they likely have better skill in predicting total TC occurrence during favorable MJO phases. Exceptions include the BoM model in the ENP and ATL when the MJO is defined by RMM, and the CMA model in the ENP and ATL when the MJO is defined by ROMI. The relationships between MJO–TC and MJO–$\mathrm{BSS_c}$ are significant only in a few TC basins in the JMA and NCEP models. In contrast, the relationships between MJO–$\mathrm{BSS_m}$ and MJO–TC in the ECMWF model are not as strongly in phase (Fig. 13). For the ECMWF model, the pattern correlation coefficients are still positive in most TC basins (Table 3) except in the ENP and SPC when the MJO is defined by ROMI. In the BoM model, the MJO–$\mathrm{BSS_m}$ relationship is negatively related to the MJO–TC relationship, indicating that the BoM model has better skill in predicting the anomaly of TC occurrence during the suppressed phases than the active ones.

While the impacts of the MJO phase on the prediction skill (whether $\mathrm{BSS_c}$ or $\mathrm{BSS_m}$) vary by basin and by model, Tables 2 and 3 suggest that favorable MJO phases are associated with better forecasting skills for predicting total TC occurrence. Favorable MJO phases are associated with better $\mathrm{BSS_m}$ in the ECMWF and CMA models in most TC basins but not in other models. It is not clear to us why there is no general relationship between favorable MJO and BSS_m, since the MJO is associated with subseseasonal TC variability. Causal connections between the MJO phases and BSS_c and BSS_m are left for future research.

6. ACE prediction

Next, we briefly discuss S2S models' performance in predicting ACE. As mentioned in Section 2, the ACE forecasts are analyzed using $RPSS_c$ and $RPSS_m$ (Section 2d). Due to insufficient horizontal grid spacing, most S2S models are unable to simulate either the TC's core structure or the occurrence of the most intense TCs. In the case of the ECMWF model, another reason for low intensity values is that TC occurrence was derived using a 1.5^o grid, which corresponds to a lower resolution than the original model grid (0.5^o). The strongest TC winds generated by the S2S models are around 50 kt (Lee et al. 2018), except for the BoM model (60-70 kt) which has 2^o horizontal resolution. The BoM model, however, might be reaching higher values of wind speed than expected, as a 2^o horizontal resolution model should not be able to generate storms with such strong winds (Davis 2018).

To correct the low-intensity bias in the S2S models, we apply quantile matching, similar to that in Camargo and Barnston (2009). One can also categorize the predicted and observed ACE into 6 categories using their respective thresholds. Here we adjust the forecast intensities before calculating ACE, so that the observed thresholds are used for all models. Results from the $RPSS_c$ analyses (Fig. 14) suggest that the ECMWF model is skillful in predicting regional TC intensity in all basins at all leads. BoM and MetFr models are skillful in most TC basins. The prediction skill scores of the NCEP and CMA models are the lowest among the six S2S models, though CMA has positive $RPSS_c$ values up to 4 weeks in the SIN. ECMWF has skill in predicting ACE anomaly ($RPSS_m$). In the WNP and SIN, the model is skillful up to 2 weeks, while in other basins only at week 1. In the same way that a model's occurrence prediction skill is influenced by its ability in capturing the genesis, the S2S models' skill predicting ACE is influenced by its ability in capturing observed genesis and occurrence. Isolating such impacts is left for a future study, as is the calibration of ACE.

7. Conclusions

The subseasonal (week 1–4) prediction skills of probabilistic forecasts of TC occurrence (genesis with subsequent daily position) and accumulated cyclone energy (ACE), at both basin and regional spatial scales, are examined using reforecasts from the BoM, CMA, ECMWF, JMA, MetFr, and NCEP in the S2S dataset. We use Brier Skill Score (BSS) for evaluating the TC occurrence predictions, and the Ranked Probabilistic Skill Score (RPSS) for ACE. Both quantities are evaluated over 15° in latitude $\times$ 20° in longitude regions (Fig. 1). The forecasts are defined as skillful when they outperform the climatological forecasts, defined by either the seasonal mean constant climatology (BSS_c and $RPSS_c$) or the monthly-varying climatology (BSS_m and $RPSS_m$). Thus, BSS_c and $RPSS_c$ evaluate the models' ability to forecast the observed TC activity, including its seasonality, while BSS_m and $RPSS_m$ considers only the TC activity deviation from that seasonality. Additionally, we investigate how the occurrence prediction skill is affected by imperfect genesis predictions and how various calibration schemes impact a model's prediction skill. We also systematically examine the dependence of S2S models' prediction skills on MJO phase.

Among the six models examined here, the ECMWF model has the best performance (Fig. 3). It is skillful in predicting TC occurrence up to 4 weeks in all TC basins, except in the NI where the model is skillful up to week 3. The model is also skillful in predicting TC occurrence anomaly 2–3 weeks in advance. Following the ECMWF are the MetFr and BoM models, which are skillful in predicting TC activity 4 weeks in advance in most TC basins. They are not skillful in predicting the TC occurrence anomaly, however. The JMA model is skillful in predicting storm occurrence 2 weeks in advance, while the CMA and NCEP models have no skill in predicting either TC occurrence or anomalies at all TC basins and leads. The prediction skills of the CMA and NCEP models may be limited by their small ensemble sizes as discussed in Lee et al. (2018). In addition to the different ensemble sizes, the S2S data periods are also different, which may also affect the S2S models' performance. By examining the BSS conditioned on the same TC (no genesis errors), we showed that the most challenging task in subseasonal occurrence predictions is to forecast genesis correctly (Fig. 5). In the case of the ECMWF model, correct genesis predictions can improve prediction skills (for TC occurrence anomaly) from 2 to 4 weeks. The S2S models' performance for ACE prediction (Fig. 14) follows their performance for the occurrence predictions, since the storm frequency largely influences ACE. The ECMWF, MetFr, and BoM model skillfully predict ACE up to 3–4 weeks. The ECMWF model is the only one that is skillful in predicting the ACE anomaly 2 weeks in advance, however.

Calibration of the mean probabilistic forecast error has been used for improving TC occurrence prediction, e.g. Camp et al. (2018) and Gregory et al. (2019). Here we showed that while calibrating the mean bias can reduce the unconditional bias component of the BSS, it does not always lead to a reduction of conditional bias (Fig. 6 and Eqs. 13 and 14). As a result, this calibration method may lead to lower BSS values (or worse skill). To know whether a calibration of the mean probabilistic forecast error benefits the BSS evaluation, one can compare the ratio between the mean forecast probability ($\overline{p}$) and the mean observed probability ($\overline{o}$) to the threshold $\frac{2\overline{po}}{\overline{p^2}} - 1$ (Eqs. 15, and 16). The prediction skill of models with large mean bias, such as CMA and NCEP, can be significantly improved with this calibration method. To calibrate the mean square probabilistic forecast error, the metric that BSS measures, we used the linear regression approach proposed by van den Dool et al. (2016). For the in-sample dataset, the linear regression method improves the

S2S model prediction skill globally. For the out-of-sample datasets, this method can improve the models' skill everywhere, except in areas where the sample TC size is too small.

Next, the dependence of the S2S models' TC forecast skill on MJO is examined using both RMM and ROMI. The S2S models' prediction skill in TC occurrence (including the seasonality) is positively related to the favorable MJO phases (Table 2). The relationship between MJO phases and the models' prediction skill for TC occurrence deviation from the seasonality varies by models and basin (Table 3). This finding is consistent with our previous work on genesis anomaly prediction (Lee et al. 2018), which showed that there is no clear relationship between MJO and genesis prediction skill. An unexpected result is that the ROMI–defined favorable MJO phases have an eastward shift when compared to those defined by RMM (Fig. 9). To the best of our knowledge, there has not yet been a satisfying answer in the literature to explain why this is the case.

Based on our findings and those in Lee et al. (2018), the ECMWF model is the most skillful ensemble prediction system for subseasonal TC genesis, occurrence and ACE forecasts in the S2S dataset, followed by BoM and MetFr. The forecast skill in predicting the anomaly of TC activity from the seasonal climatology remains low, however, even in these models. Genesis prediction is the key bottleneck causing this low prediction skill. Our results highlight the importance of improving our fundamental understanding of TC genesis in order to obtain more skillful subseasonal TC predictions. Calibrating subseasonal probabilistic TC predictions is not easy, but a comprehensive calibration method can largely increase models' prediction skills and should be further explored in the future. It should be mentioned that this research and Lee et al. (2018) present the prediction skill directly derived from the reforecasts in the S2S dataset. Our results may not reflect the latest prediction skill of the operational centers mentioned here because they may have further improved since the collections of the S2S dataset. Also, reforecasts in the S2S dataset have small ensemble sizes, except for BoM, and both BSS and RPSS punish small ensemble sizes. Such a negative impact maybe even more significant for NCEP and CMA because both models have only four members in the S2S datasets. Variants of the RPSS and BSS (Weigel et al. 2007), which take into account the ensemble size, may be used in the future to examine model skill if the ensemble size was infinite.

Data Availability Statement. S2S data and S2S TC tracks are available to research community at `http://s2sprediction.net`. Best-track data for Northern Atlantic, and Eastern Pacific are available at `https://www.nhc.noaa.gov/data/#hurdat` and those for Southern Hemisphere, Northern Indian Ocean and Western North Pacific are archived at `https://www.metoc.navy.mil/jtwc/jtwc.html?best-tracks`.

Acknowledgments. We thank the three anonymous reviewers for their thorough reviews. The research was supported by NOAA S2S projects NA16OAR4310079 and NA16OAR4310076.

References

Belanger, J. I., J. A. Curry, and P. J. Webster, 2010: Predictability of North Atlantic tropical cyclone activity on intraseasonal time scales. *Mon. Wea. Rev.*, **138**, 4362–4374.

Bradley, A., S. Schwartz, and T. Hashino, 2008: Sampling uncertainty and confidence intervals for the brier score and brier skill score. *Wea. Forecasting*, **23 (5)**, 992–1006.

Camargo, S. J., and A. G. Barnston, 2009: Experimental dynamical seasonal forecasts of tropical cyclone activity at IRI. *Wea. Forecasting*, **24**, 472–491.

Camargo, S. J., and Coauthors, 2019: Tropical cyclone prediction on subseasonal time-scales. *Trop. Cyclone Res. Rev.*, **8**, 156–165.

Camp, J., and Coauthors, 2018: Skilful multi-week tropical cyclone prediction in ACCESSS1 and the role of the MJO. *Q.J.R. Meteorol. Soc..*, Accepted Author Manuscript.

Chen, J.-H., and S.-J. Lin, 2013: Seasonal predictions of tropical cyclones using a 25–km-resolution general circulation model. *J. Climate*, **(2)**, 380–398.

Chu, J.-H., C. R. Sampson, A. Lavine, and E. Fukada, 2002: *The Joint Typhoon Warning Center tropical cyclone best-tracks, 1945-2000*. 22pp, Naval Research Laboratory Tech. rep. NRL/MR/7540-02-16.

Davis, C. A., 2018: Resolving tropical cyclone intensity in models. *Geophys. Res. Lett.*, **45**, 2082–2087.

DeMaria, M., C. R. Sampson, J. A. Knaff, and K. D. Musgrave, 2014: Is tropical cyclone intensity guidance improving? *Bull. Amer. Meteor. Soc.*, **95**, 387–398.

Dong, K., and C. J. Neumann, 1986: The relationship between tropical cyclone motion and environmental geostrophic flows. *Mon. Wea. Rev.*, **114**, 115–122.

Elsberry, R. L., M. S. Jordan, and F. Vitart, 2011: Evaluation of the ECMWF 32-day ensemble predictions during the 2009 season of the western North Pacific tropical cyclone events on intraseasonal timescales. *Asia-Pacific J. Atmos. Sci.*, **47**, 305–318.

Galarneau, T. J., and C. A. Davis, 2012: Diagnosing forecast errors in tropical cyclone motion. *Mon. Wea. Rev.*, **141**, 405–430.

Gao, K., J.-H. Chen, L. Harris, Y. Sun, and S.-J. Lin, 2019: Skillful prediction of monthly major hurricane activity in the north atlantic with two-way nesting. *Geophys. Res. Lett.*, **0 (0)**.

Gottschalck, J., and Coauthors, 2010: A framework for assessing operational Madden–Julian Oscillation forecasts: A CLIVAR MJO working group project. *Bull. Amer. Meteor. Soc.*, **91**, 1247–1258.

Gregory, P. A., J. Camp, K. Bigelow, and A. Brown, 2019: Sub-seasonal predictability of the 2017—2018 Southern Hemisphere tropical cyclone season. *Atmos Sci Lett*, **20 (4)**, e886, doi: 10.1002/asl.886.

Jiang, X., M. Zhao, and D. E. Waliser, 2012: Modulation of tropical cyclones over the Eastern Pacific by the intraseasonal variability simulated in an AGCM. *J. Climate*, **25**, 6524–6538.

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.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592.

Lee, C.-Y., S. J. Camargo, F. Vitart, A. H. Sobel, and M. K. Tippett, 2018: Subseasonal tropical cyclone genesis prediction and MJO in the S2S dataset. *Wea. Forecasting*, **33 (4)**, 967–988, doi:10.1175/WAF-D-17-0165.1.

Li, W. W., Z. Wang, and M. S. Peng, 2016: Evaluating tropical cyclone forecasts from the NCEP Global Ensemble Forecasting System (GEFS) Reforecast Version 2. *Wea. Forecasting*, **31**, 895–916.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7 (4)**, 435–455.

Schreck, C. J., 2015: Kelvin waves and tropical cyclogenesis: A global survey. *Mon. Wea. Rev.*, **143**, 3996–4011.

Straub, K. H., 2012: Mjo initiation in the Real-Time Multivariate MJO index. *J. Climate*, **(4)**, 1130–1151.

Tsai, H.-C., R. L. Elsberry, M. S. Jordan, and F. Vitart, 2013: Objective verifications and false alarm analyses of western North Pacific tropical cyclone event forecasts by the ECMWF 32-day ensemble. *Asia-Pacific J. Atmos. Sci.*, **49**, 409–420.

van den Dool, H., E. Becker, L.-C. Chen, and Q. Zhang, 2016: The probability anomaly correlation and calibration of probabilistic forecasts. *Wea. Forecasting*, **32 (1)**, 199–206.

Ventrice, M. J., C. D. Thorncroft, and M. A. Janiga, 2011: Atlantic tropical cyclogenesis: A three-way interaction between an African easterly wave, diurnally varying convection, and a convectively coupled atmospheric Kelvin wave. *Mon. Wea. Rev.*, **140**, 1108–1124.

Ventrice, M. J., C. D. Thorncroft, and C. J. Schreck, 2012: Impacts of convectively coupled Kelvin waves on environmental conditions for Atlantic tropical cyclogenesis. *Mon. Wea. Rev.*, **140**, 2198–2214.

Ventrice, M. J., M. C. Wheeler, H. H. Hendon, C. J. Schreck, C. D. Thorncroft, and G. N. Kiladis, 2013: A modified multivariate Madden–Julian Oscillation index using velocity potential. *Mon. Wea. Rev.*, **141**, 4197–4210.

Vitart, F., 2017: Madden-Julian Oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2210–2220.

Vitart, F., A. Leroy, and M. C. Wheeler, 2010: A comparison of dynamical and statistical predictions of weekly tropical cyclone activity in the southern hemisphere. *Mon. Wea. Rev.*, **138**, 3671–3682.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmospheric Science*, **1 (1)**, 3.

Vitart, F., and T. N. Stockdale, 2001: Seasonal forecasting of tropical storms using coupled GCM integrations. *Mon. Wea. Rev.*, **129**, 2521–2537.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S)prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173.

Wang, S., D. Ma, A. H. Sobel, and M. K. Tippett, 2018: Propagation characteristics of BSISO indices. *Geophys. Res. Lett.*, **45**, 9934–9943.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124.

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.

Yamaguchi, M., F. Vitart, S. T. K. Lang, L. Magnusson, R. L. Elsberry, G. Elliot, M. Kyouda, and T. Nakazawa, 2015: Global distribution of the skill of tropical cyclone activity forecasts on short- to medium-range time scales. *Wea. Forecasting*, **30**, 1695–1709.

Zhan, R., Y. Wang, and M. Ying, 2012: Seasonal forecasts of tropical cyclone activity over the Western North Pacific: A review. *Tropical Cyclone Research and Review*, **1 (3)**, 307 – 324.

Zhang, G., and Z. Wang, 2019: North Atlantic Rossby wave breaking during the hurricane season: Association with tropical and extratropical variability. *J. Climate*, **32**, 3777–3801.

LIST OF TABLES

TABLE 1. Characteristics of the six S2S reforecasts used here. (Adapted from Lee et al. 2018)

Model	forecast time	Resolution	Period	Ens. size	Frequency & Sample size
BoM	0–64 days	2^o, L17	1981–2013	33	$\sim$5 days & 2160
CMA	0–61 days	1^o, L40	1994–2014	4	daily & 7665
ECMWF	0–46 days	0.25^o for first 10 days 0.5^o after day 10, L91	1994–2014	11	$\sim$4 days & 2058
JMA	0–33 days	0.5^o, L60	1981–2010	5	$\sim$10 days & 1079
MetFr	0–61 days	$\sim 0.7^o$, L91	1993–2014	15	$\sim$15 days & 528
NCEP	0–44 days	$\sim 1^o$, L64	1999–2010	4	daily & 4380

TABLE 2. Pattern correlation coefficients between the lag-plots of TC occurrence anomaly (%) and MJO and those of $BSS_{c|mjo}$- BSS_c and MJO. Positive (negative) values correspond to favorable (suppressed) MJO phases having a positive (negative) impact onto BSS_c. Correlations significant at the 95% level (p-value < 0.05) are shown in bold.

	BSS_c v.s. RMM					
basins/models	BoM	CMA	ECMWF	JMA	MetFr	NCEP
ni	**0.15**	**0.38**	**0.58**	-0.02	**0.23**	**0.27**
wnp	**0.29**	**0.30**	**0.66**	**0.32**	**0.27**	**0.53**
enp	**-0.25**	**0.29**	**0.23**	**0.52**	**0.32**	-0.08
atl	**-0.22**	0.09	**0.17**	**0.27**	-0.01	-0.03
sin	**0.61**	**0.58**	**0.64**	0.05	**0.44**	**0.57**
aus	**0.38**	**0.46**	**0.46**	**0.17**	**0.22**	**0.35**
spc	**0.31**	**0.74**	**0.37**	0.08	**0.26**	**0.45**
	BSS_c v.s. ROMI					
basins/models	BoM	CMA	ECMWF	JMA	MetFr	NCEP
ni	**0.47**	**0.63**	**0.38**	-0.04	**0.16**	0.07
wnp	**0.55**	**0.45**	**0.33**	0.09	**0.32**	**0.37**
enp	**0.13**	**-0.16**	**0.27**	**0.26**	0.01	-0.10
atl	0.09	**-0.31**	**0.43**	**0.22**	**0.13**	-0.00
sin	**0.68**	**0.26**	**0.34**	-0.04	**0.23**	-0.07
aus	**0.57**	**0.51**	**0.51**	-0.02	**0.28**	**0.23**
spc	**0.25**	**0.35**	**0.33**	**-0.18**	**0.29**	**0.63**

TABLE 3. Similar to Table 2, but for BSS_m

	BSS_m v.s. RMM					
basins/models	BoM	CMA	ECMWF	JMA	MetFr	NCEP
ni	-0.12	**0.25**	**0.42**	-0.06	**0.13**	**0.17**
wnp	**-0.26**	**0.20**	**0.13**	-0.10	**-0.21**	0.09
enp	**-0.37**	**0.29**	-0.07	**0.21**	-0.05	**-0.16**
atl	**-0.49**	0.11	**0.28**	0.01	**-0.23**	0.05
sin	**0.36**	**0.17**	**0.35**	-0.06	0.12	**0.45**
aus	**-0.44**	**0.28**	**0.24**	0.02	-0.03	-0.05
spc	**-0.41**	**0.74**	-0.10	**0.15**	0.00	**0.34**
	BSS_m v.s. ROMI					
basins/models	BoM	CMA	ECMWF	JMA	MetFr	NCEP
ni	0.05	**0.53**	**0.14**	**-0.21**	0.11	-0.02
wnp	**-0.26**	**0.46**	-0.10	**-0.20**	0.05	**0.18**
enp	**-0.26**	-0.03	**-0.36**	0.07	**-0.12**	0.06
atl	**-0.17**	**-0.41**	**0.14**	0.07	**-0.26**	0.04
sin	**0.28**	**-0.23**	**0.15**	**-0.42**	0.10	**-0.24**
aus	**-0.46**	**0.27**	**0.28**	**-0.19**	0.01	**-0.34**
spc	**-0.59**	**0.35**	**-0.26**	**-0.23**	**0.25**	**0.52**

LIST OF FIGURES

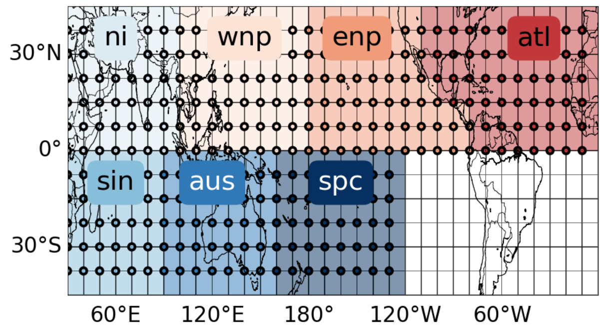


FIG. 1. The verification areas for seven TC basins. The verification is conducted over regions of 20^{o} in longitude $\times$ 15^{o} in latitude, and there are a total of 303 regions (11$\times$33 grids minus southern Atlantic and eastern South Pacific). The regions overlap by 10^{o} in longitude and 7.5 ^{o} in latitude.

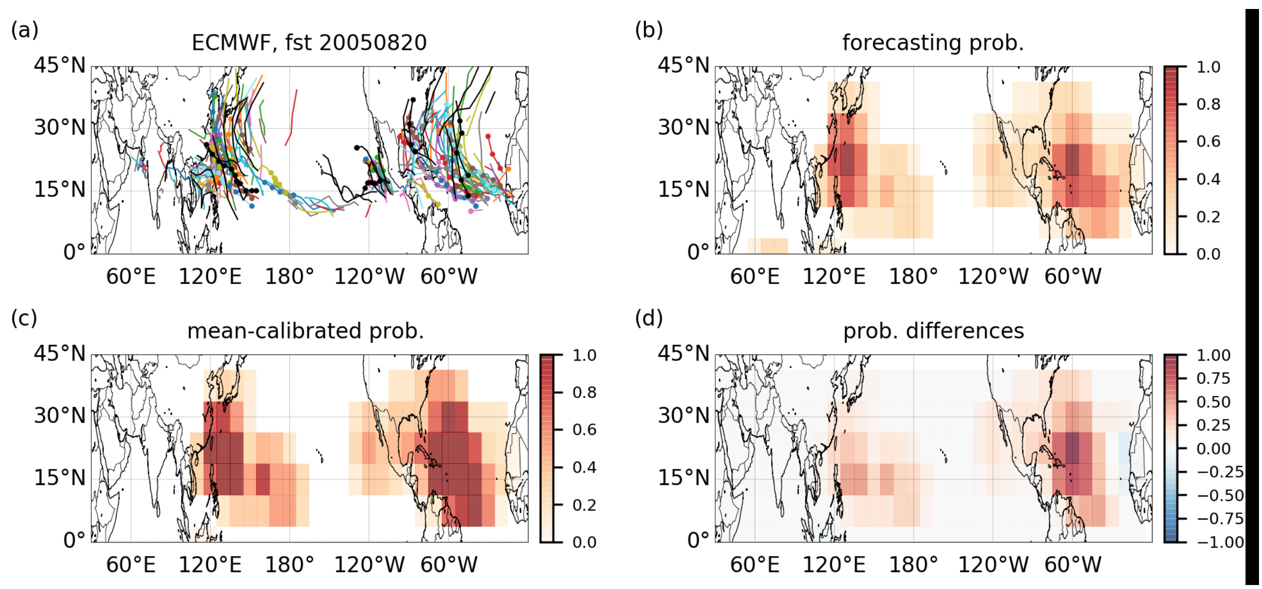


FIG. 2. (a) All TC tracks (colored lines) predicted from an ECMWF forecast initialized at August 20, 2005. There are 11 ensemble members for the ECMWF model and one color per ensemble member. Forecast storm centers (occurrence) at lead times 8 – 14 days (week 2) are marked by colored circles. The corresponding observed TC tracks and storm centers are marked in black lines and circles. (b) Week 2 forecast probability of storm occurrence (Eq. 3). (c) Week 2 forecast after calibration (Eq. 8). (d) Difference between (b) and (c).

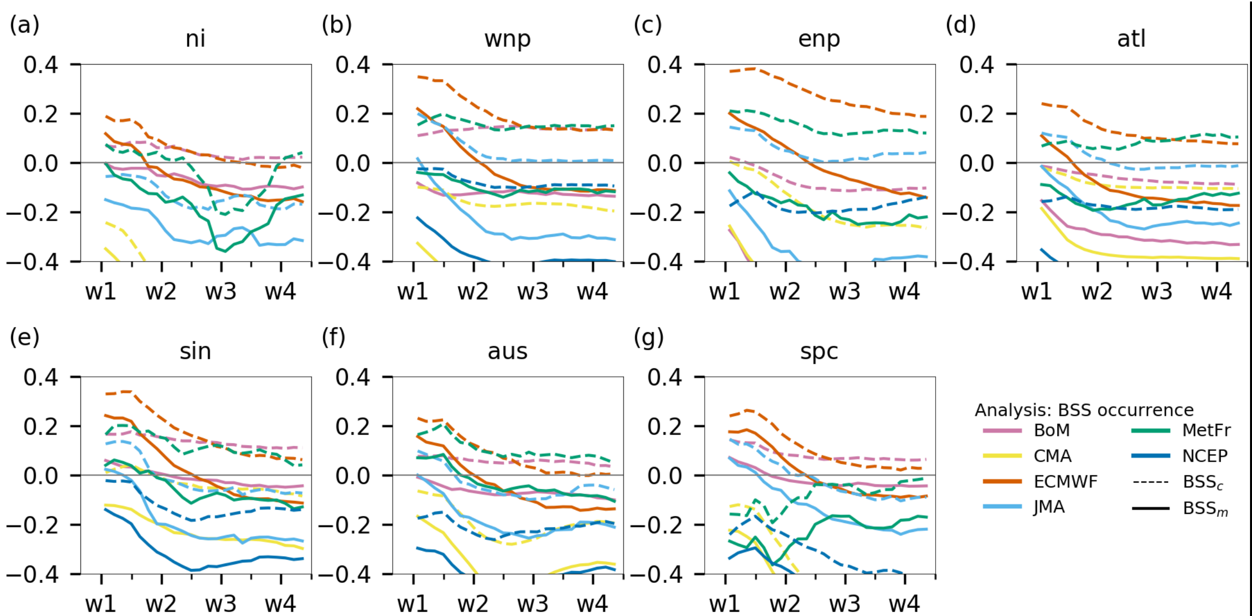


FIG. 3. Basin-wide BSS_c (dashed lines) and BSS_m (solid lines) for TC occurrence prediction in the S2S models.

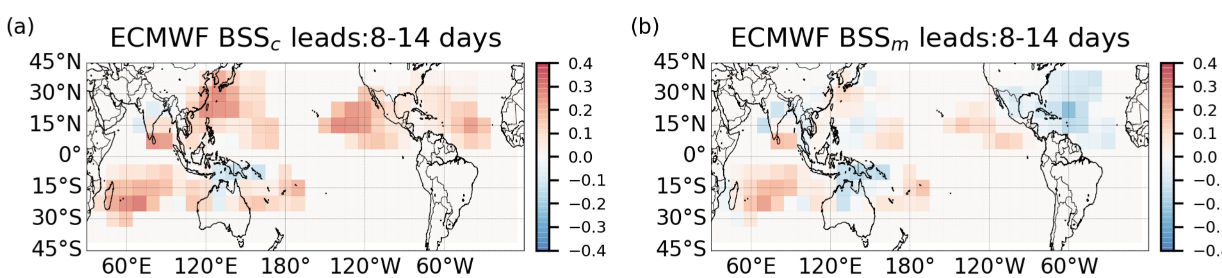


FIG. 4. Global map of ECMWF week 2 TC occurrence skill scores for (a) $\mathrm{BSS_c}$ (seasonal mean constant climatology), (b) $\mathrm{BSS_m}$ (seasonal monthly varying climatology).

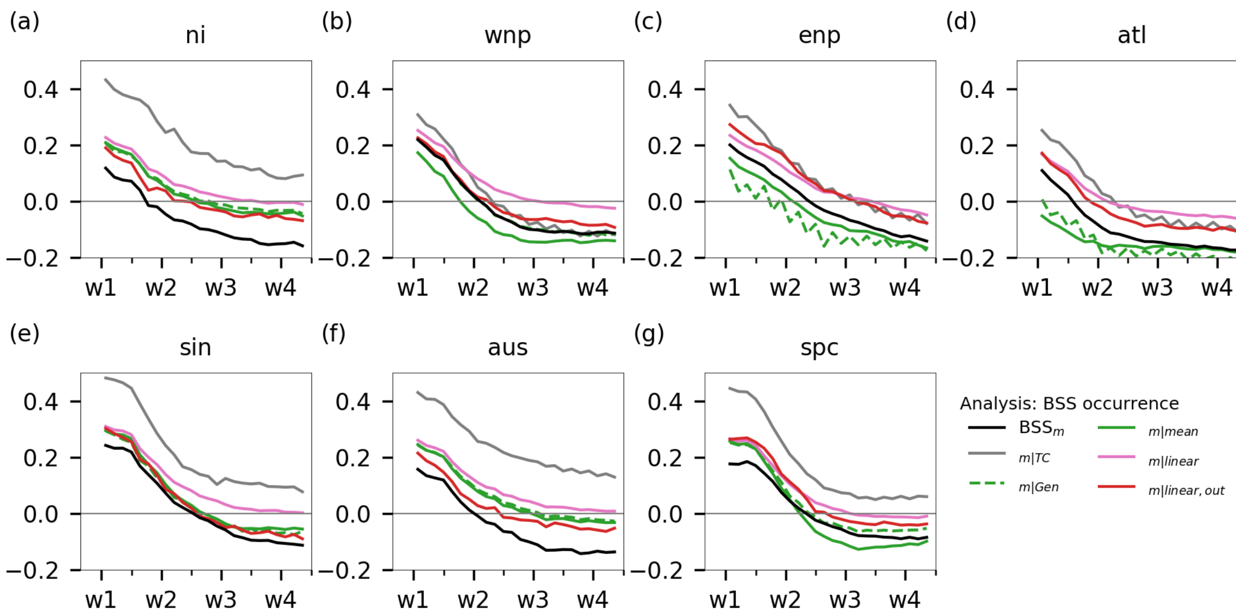


FIG. 5. Basin-wide ECMWF $\mathrm{BSS_m}$ (black lines), $\mathrm{BSS_{m|TC}}$ (gray lines), $\mathrm{BSS_{m|gen}}$ (green dashed lines), $\mathrm{BSS_{m|mean}}$ (green solid lines), and $\mathrm{BSS_{m|linear}}$ (pink lines) calculated with the whole forecast data. $\mathrm{BSS_{m|linear,out}}$ (red lines) are similar to $\mathrm{BSS_{m|linear}}$ but use the out-of-sample data. See Sections 3 and 4 for details.

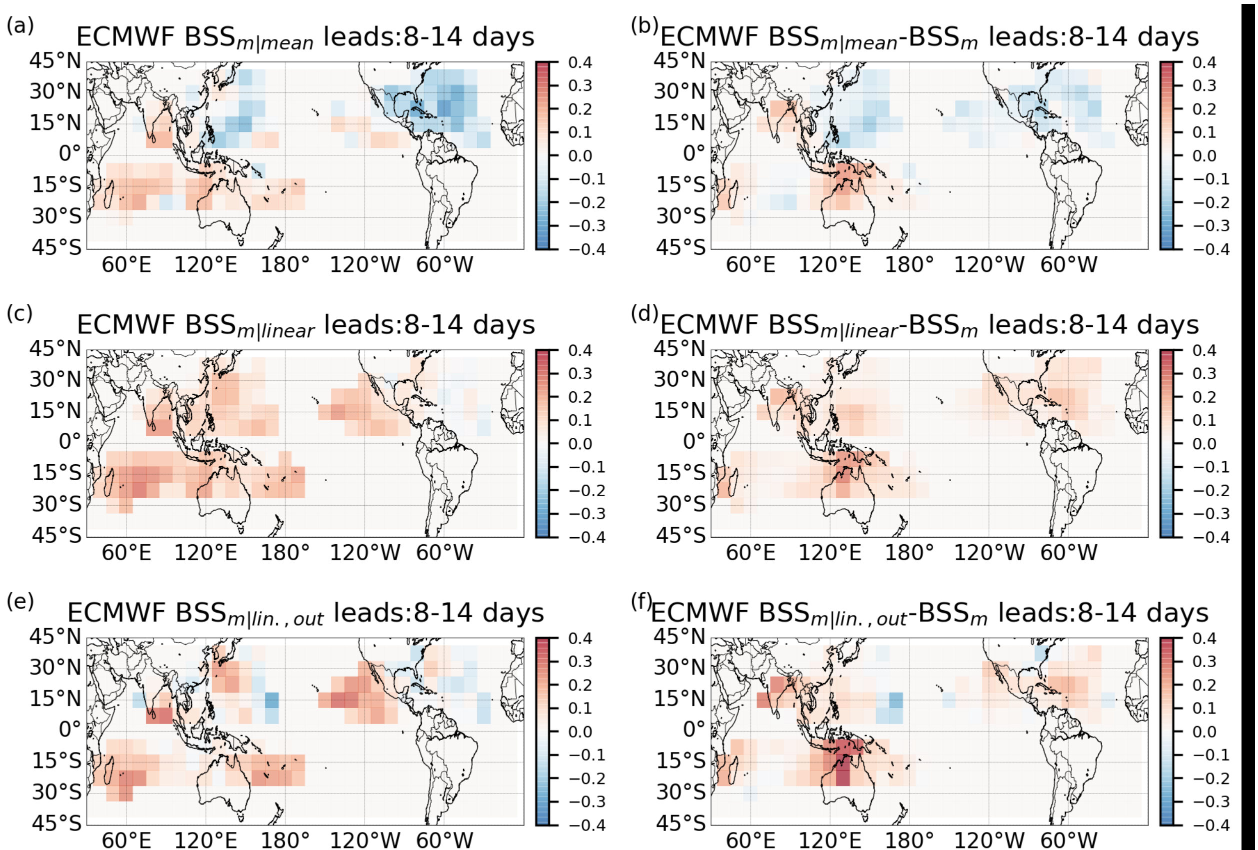


FIG. 6. Global map of calibrated ECMWF week 2 TC occurrence skill score for (a) $\mathrm{BSS}_{\mathrm{m|mean}}$, (c) $\mathrm{BSS}_{\mathrm{m|linear}}$, and (e) $\mathrm{BSS}_{\mathrm{m|linear,\ out}}$. (b) and (d) are the differences between (a) and (c) to the $\mathrm{BSS}_{\mathrm{m}}$, respectively, in Fig. 4b. (f) is the difference between $\mathrm{BSS}_{\mathrm{m|linear,\ out}}$ and the corresponding $\mathrm{BSS}_{\mathrm{m}}$ from the same out-of-sample period (not shown).

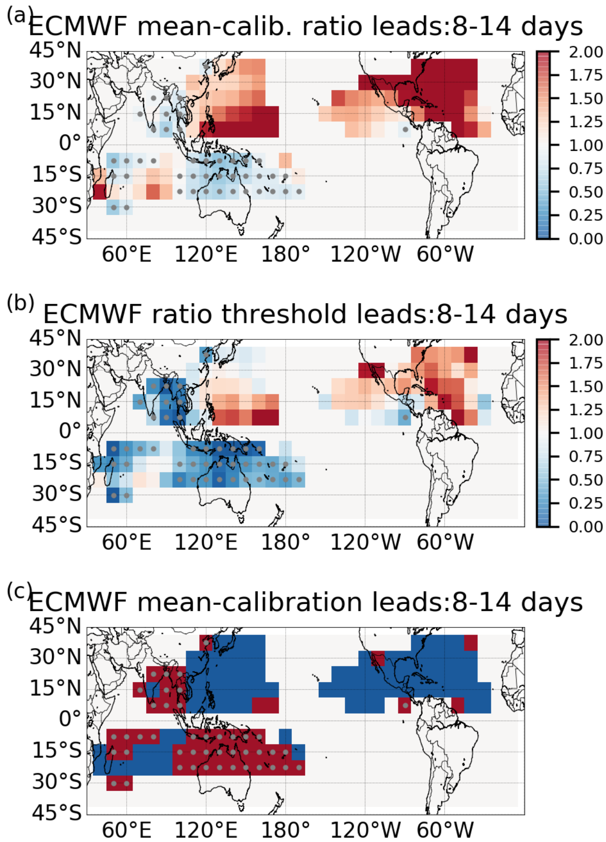


FIG. 7. (a) Week 2 ECMWF forecasts' ratio between the mean forecast probability and observed probability. (b) Global maps of $\frac{2\overline{po}}{\overline{p^2}} - 1$ (c) Areas where the calibration scheme has a positive (negative) impact are marked in red (blue). Regions where the ECMWF model has low biases (the values in (a) is smaller than 1) are labeled by gray dots in all three figures. (see Section 4 for details).

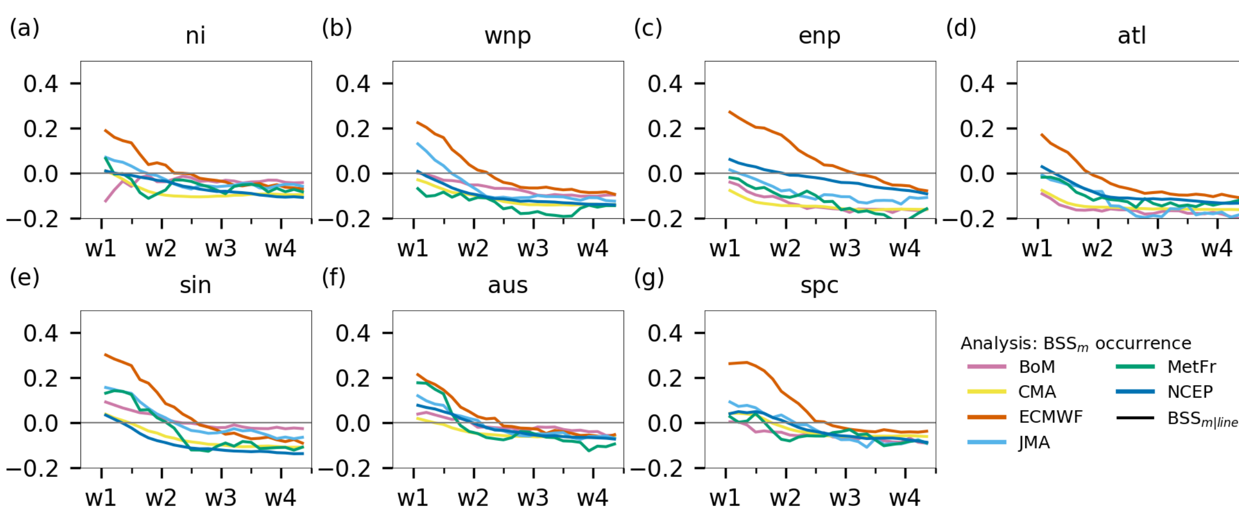


FIG. 8. Basin-wide $BSS_{m|linear,out}$ of TC occurrence prediction in the S2S models.

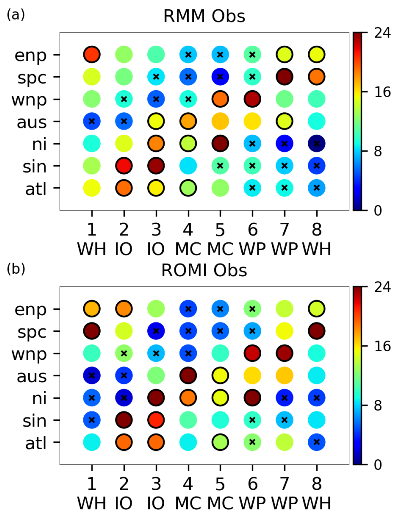
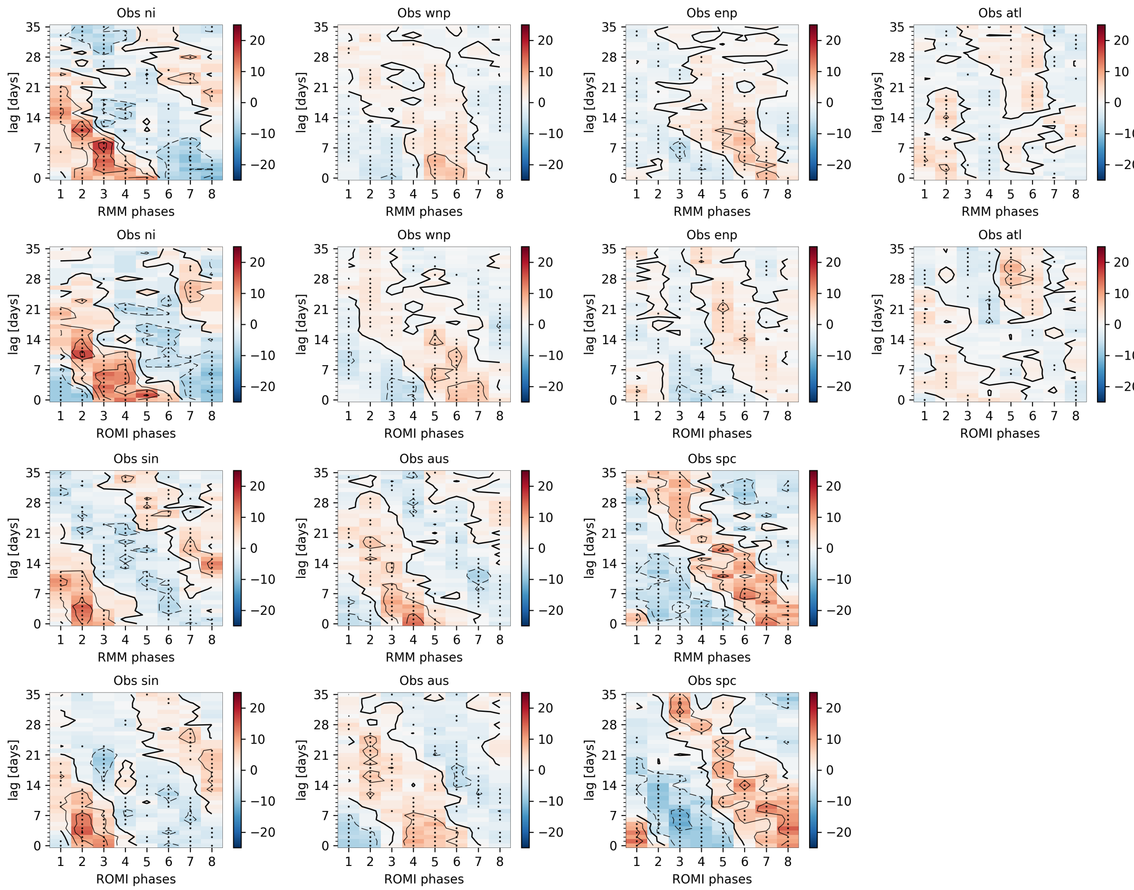


FIG. 9. Candy plots for the MJO–TC relationship in the observations. The color of each candy indicates the PDF (%) of TC frequency in the corresponding MJO phase in the basin. The sum of the circles across the MJO phases in each basin is 100%. The black circle at the edge indicates that the value is above the 90$^{\text{th}}$ percentile while the cross symbol (X) at the center means the value is below the 10$^{\text{th}}$ percentile. (a) uses RMM to define MJO phases while (b) uses ROMI. We use only the data from MJO events with a magnitude larger than 1.



713 FIG. 10. Observed lag-plot of TC occurrence anomaly (%) based on RMM and ROMI. Gray dots show where the anomaly is statistically significant. Data are normalized by the number of the MJO days in each phase.

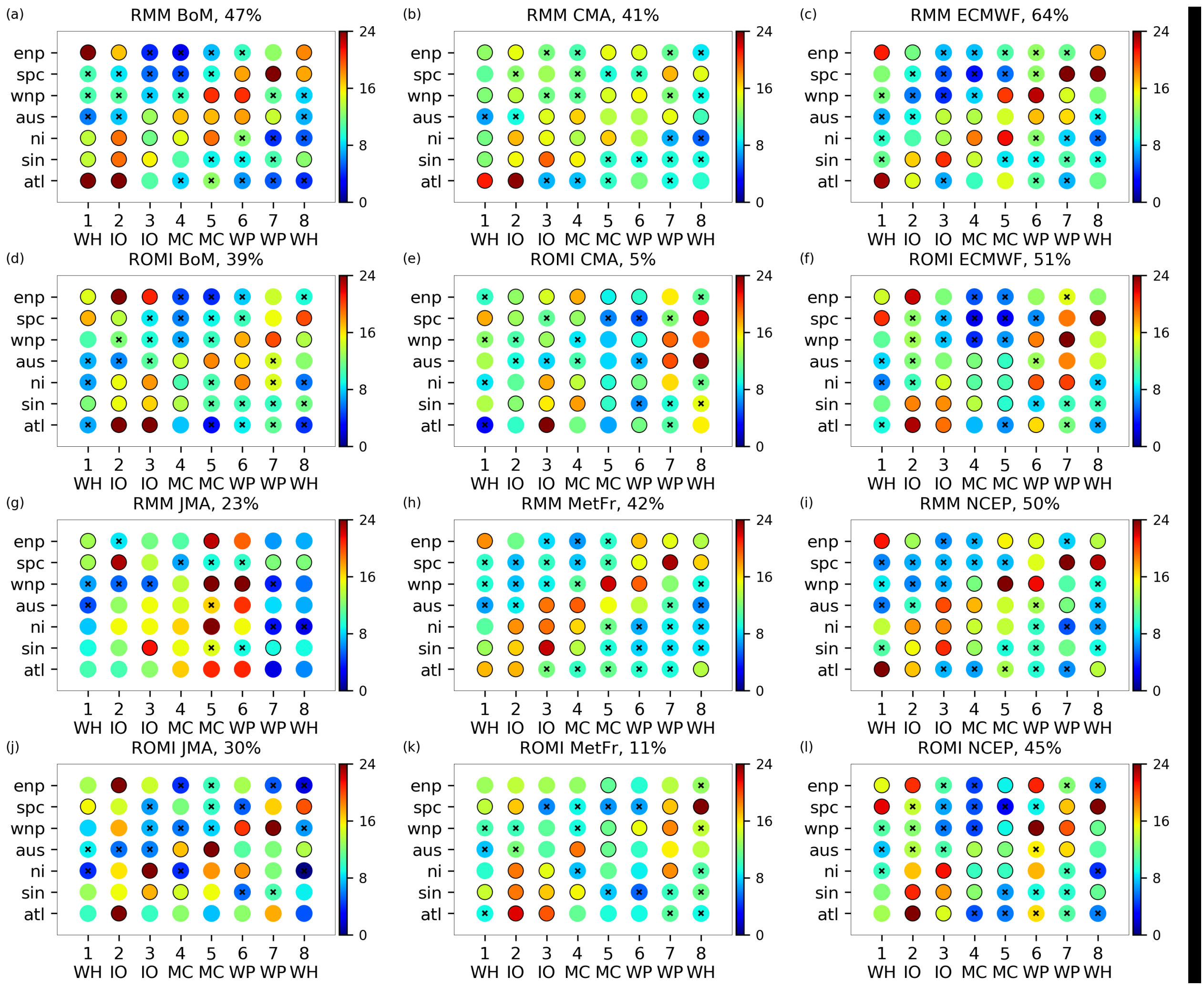


FIG. 11. Similar to Fig. 9 but for week 2 forecasts of the S2S models.The % in the title of each figure shows the pattern correlation between model simulations and observations from Fig. 9

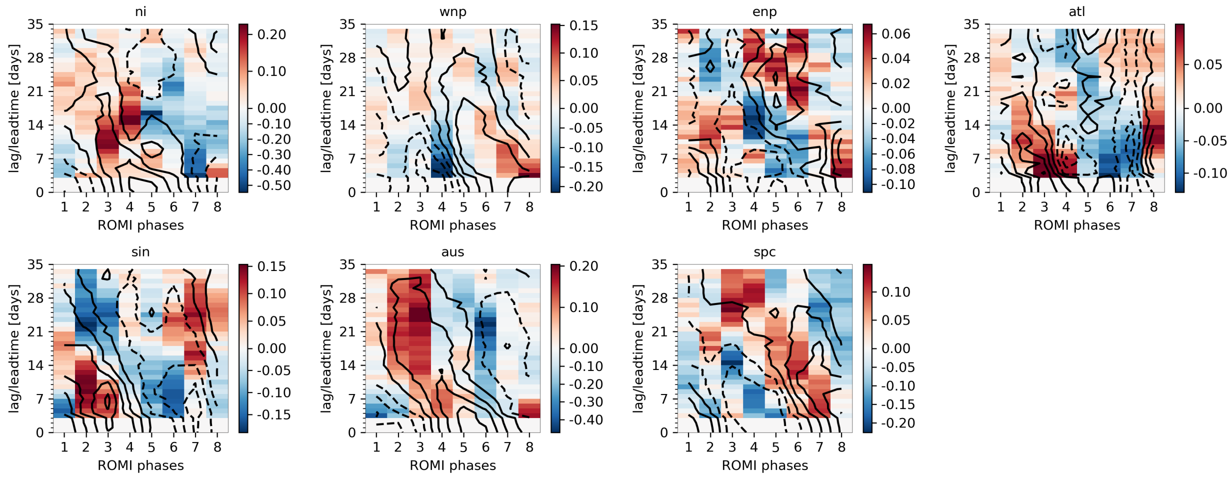


FIG. 12. ECMWF lag-plot of BSS_c anomaly ($BSS_{c|mjo}$-BSS_c) based on the ROMI index. $BSS_{c|mjo}$ is the BSS_c using only forecasts at specified MJO phases. Note that the color scheme is centered at 0, and thus reddish (bluish) color indicates positive (negative) contribution from MJO favorable (suppressed) phases. We only use data for MJO events with magnitudes larger than 1. The contours show the simulated MJO–TC relationships, similar to those shown in Fig. 10.

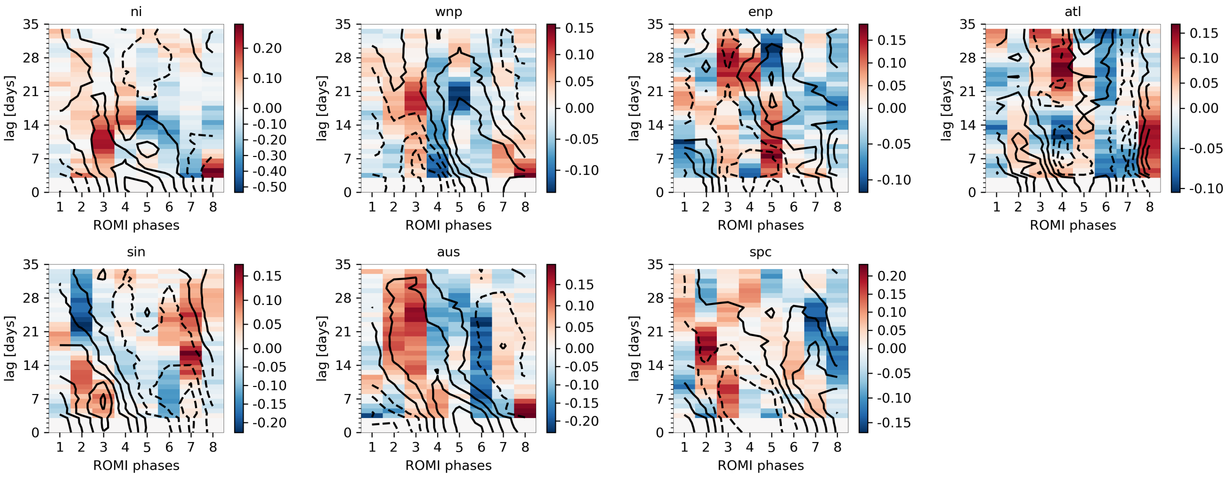


FIG. 13. Similar to Fig. 12 but for BSS_m.

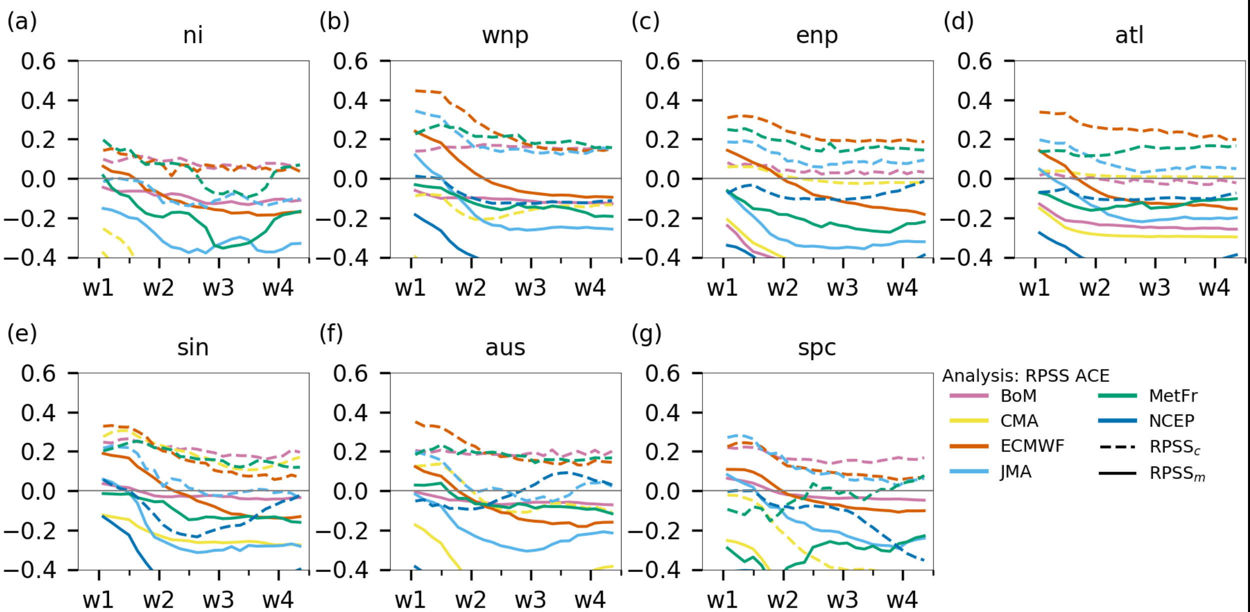


FIG. 14. RPSS$_c$ and RPSS$_m$ for ACE predictions in the S2S models.