Skill of numerical weather prediction ensembles in forecasting genesis of South Asian monsoon low pressure systems

D. L. Suhas[1] and William R. Boos[1,2]

[1] *Department of Earth and Planetary Science, University of California, Berkeley*

[2] *Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory*

Corresponding author: D. L. Suhas, suhasdl@berkeley.edu

ABSTRACT: Synoptic-scale vortices known as monsoon low pressure systems (LPS) frequently produce intense precipitation and hydrological disasters in South Asia, so accurately forecasting LPS genesis is crucial for improving disaster preparedness and response. However, the accuracy of LPS genesis forecasts by numerical weather prediction models has remained unknown. Here, we evaluate the skill of two global ensemble models—the U.S. Global Ensemble Forecast System (GEFS) and the Ensemble Prediction System of the European Centre for Medium-Range Weather Forecasts (ECMWF)—in predicting LPS genesis during the years 2021–2022. The GEFS successfully predicted about half the observed LPS genesis events one to two days in advance; the ECMWF model captured an additional 10% of observed genesis events. Both models had a False Alarm Ratio (FAR) around 50% for one- to two-day lead times. In both ensembles, the control run typically exhibited a higher probability of detection (POD) of observed events and a lower FAR compared to the perturbed ensemble members. However, a consensus forecast, in which genesis is predicted when at least 20% of ensemble members forecast LPS formation, had POD values surpassing that of the control run for all lead times. Moreover, probabilistic predictions of genesis over the Bay of Bengal, where most LPS form, were skillful, with the fraction of ensemble members predicting LPS formation with five days of lead time approximating the observed frequency of genesis, without any adjustment or bias-correction.

1. Introduction

Monsoon low pressure systems (LPS) are transient, synoptic-scale cyclonic vortices that deliver heavy precipitation to South Asia in boreal summer (Sikka 1977; Hurley and Boos 2015). With an outer diameter of about 2000 km, these vortices typically form over the northern Bay of Bengal then propagate northwestward for 3–6 days (Godbole 1977; Hurley and Boos 2015). The precipitation that falls in these storms is concentrated southwest of the vortex center (**??**Boos et al. 2015) and produces a large fraction of South Asia's hydrological disasters (Suhas et al. 2023), as well as much of the region's total summer rainfall (Godbole 1977; Yoon and Chen 2005; Hunt and Fletcher 2019).

It seems natural to turn to numerical weather prediction (NWP) models for forecasts of monsoon LPS because these models have demonstrated effectiveness in forecasting synoptic-scale disturbances, including low-latitude vortices such as tropical cyclones (TCs; Cangialosi et al. 2020; Goerss et al. 2004; Yamaguchi et al. 2015). Past studies have assessed the skill of NWP forecasts of TC positions and intensities, and have also extensively evaluated the skill of NWP forecasts of TC genesis across various ocean basins and lead times (Chan and Kwok 1999; Briegel and Frank 1997; Cheung and Elsberry 2002; Halperin et al. 2013, 2016; Liang et al. 2021). The reliability of TC genesis predictions made by global NWP models has improved over the years (Halperin et al. 2013, 2016), although variations in skill exist for different basins (Halperin et al. 2016; Liang et al. 2021). Additionally, studies have explored various factors that could influence the skill of NWP forecasts of TC genesis (McTaggart-Cowan et al. 2013; Wang et al. 2018; Halperin et al. 2020).

Ensemble forecasts may be particularly useful for the phenomenon of vortex genesis, which has been described with some success as a random Poisson process that depends on the larger-scale atmospheric state (Tippett et al. 2011). Studies have accordingly investigated TC genesis forecasts made using ensembles of NWP simulations (Majumdar and Torn 2014; Yamaguchi and Koide 2017; Nakano et al. 2015; Zhang et al. 2023). These ensemble systems typically represent uncertainties by incorporating variations in initial conditions or parameterizations, enabling probabilistic TC genesis forecasts and potentially extending predictive skill by a week or longer (Elsberry et al. 2014; Komaromi and Majumdar 2015; Lee et al. 2018). This probabilistic guidance and uncertainty information has proven useful for operational forecasting (Titley et al. 2019), with forecast reliability improving for TC genesis when a larger number of ensemble members predict TC formation at a particular location and time (Yamaguchi and Koide 2017; Jaiswal et al. 2016). Recently, Zhang et al. (2023) demonstrated the effectiveness of using ensemble models and even multi-model ensembles to enhance the skill of TC genesis forecasts.

For monsoon LPS, short-term NWP forecasts of vortex tracks and precipitation have been suggested to be useful for early warning systems and improving disaster preparedness (Suhas et al. 2023). However, in contrast to TCs, only a few studies have examined the forecast capabilities of NWP models for monsoon LPS (Sarkar et al. 2021; Deoras et al. 2021), and none have focused on the genesis of LPS. For instance, Sarkar et al. (2021) examined the skill of a version of the Global Forecasting System (GFS) used operationally by the India Meteorological Department (IMD) in simulating the dynamics of LPS, but did not focus on genesis. Deoras et al. (2021) investigated biases in track position, intensity, and precipitation using 11 subseasonal to seasonal (S2S) ensemble models, but did not evaluate the skill of LPS genesis in these models. In related work that introduces the automated operational tracking of LPS **?**, we examine the skill of deterministic GFS and ECMWF model forecasts of LPS behavior, including genesis, over South Asia; however, that work solely examines the control run of these NWP models and did not assess the skill of the ensemble members or probabilistic predictions by model ensembles.

Here we evaluate the skill of NWP ensembles in forecasting monsoon LPS genesis. Specifically, we assess LPS genesis as represented by two operational global ensemble models: the Ensemble Prediction System from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Global Ensemble Forecast System (GEFS) from the U.S. National Centers for Environmental Prediction (NCEP). The datasets and tracking methodology are described in the next section. The results are then presented in Section 3 and the conclusions are summarised in Section 4.

2. Data and methods

a. Datasets

We use global ensembles from the GEFS and ECMWF models for the boreal summer monsoon season, spanning June to September in the years 2021 to 2022. The GEFS version 12, the latest ensemble prediction system developed by NCEP, consists of 31 members using the finite-volume cubed-sphere (FV3) dynamical core at C384L64 resolution (horizontal grid spacing of ~25 km with 64 vertical levels). It is initialized four times a day, with the control run using the hybrid GFS analysis as the initial conditions; other ensemble members use the analysis perturbed by the 6-h EnKF forecast ensemble (Zhou et al. 2022). Further details regarding the model can be found in Zhou et al. (2022).

The ECMWF system comprises 51 ensemble members, each with slightly modified initial conditions and model physics. These models are initialized twice daily at 00 and 12 UTC (Molteni et al. 1996; Palmer 2019). During the period considered in this study, the model used a horizontal resolution of O640 (an octahedral grid, nominally 18 km grid spacing in the horizontal) with 137 vertical levels. A recent model upgrade at the end of June 2023 (IFS cycle 48r1) enhanced the horizontal resolution to O1280 (approximately 9 km), but we did not analyze this here because less than one full summer of output was available at the time of writing.

We obtained forecasts from both model ensembles from the archive of The International Grand Global Ensemble (TIGGE; Bougeault et al. 2010; Swinbank et al. 2016), at horizontal grid spacing of $0.5^\circ \times 0.5^\circ$ and 6-hourly frequency. To ensure consistency between the models, we only consider forecasts initialized at 00 and 12 hours UTC. We limit our analysis to the years 2021 and 2022 mainly due to storage limitations. For instance, a single summer (June to September) of ECMWF ensemble output requires approximately 60 TB of storage space. Additionally, this period of analysis avoids any impact of the change in the number of GEFS ensemble members that occurred in late September 2020. Since our focus is on short-term forecasts, we further restrict the forecasts to a maximum of 10 days after model initialization. Throughout this period, there were no significant changes to the models or ensemble perturbation methods.

The genesis skill of the forecasts is evaluated using LPS tracks from the ERA5 reanalysis as the reference. ERA5 is the fifth-generation reanalysis from ECMWF, which assimilates a much larger number of reprocessed observational datasets and novel observations compared to earlier reanalyses. It is available at hourly frequency with horizontal grid spacing of 0.25° (Hersbach et al. 2020). LPS tracks identified in ERA5 have good agreement with those obtained by expert hand analysis and with tracks identified in other reanalysis products (Vishnu et al. 2020). Additionally, to ensure the robustness of our results with respect to the reference dataset, we also evaluate the models' genesis skill using the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) as the reference, providing the corresponding results in the appendix.

b. Track identification

We identify tracks of monsoon LPS from the streamfunction of 850 hPa horizontal winds, using an objective tracking algorithm developed by Vishnu et al. (2020). This algorithm is based on TempestExtremes, an automated Lagrangian feature tracker (Ullrich et al. 2021), with thresholds and parameters optimized to identify LPS tracks in five reanalyses. The tracking process involves two steps. First, candidate vortex centers are identified where the 850 hPa streamfunction exhibits a local minimum, with the closed contour that can be drawn around this minimum increasing by at least $12.5 \times 10^5\ \mathrm{m^2\,s^{-1}}$, relative to the minimum, over a 10° great circle distance from the minimum. These candidate points are then stitched together to form LPS tracks, retaining only tracks that persist for a minimum of 24 hours and have no gaps between candidate points beyond 12 hours. Since we focus on the South Asian summer monsoon, we only consider tracks passing through the region 5°N-30°N, 60°E-100°E during June to September. Additional details regarding the tracking algorithm and its optimized parameters can be found in Vishnu et al. (2020).

c. Genesis verification

We consider a model forecast successful, defined as a “hit”, if it predicts the formation of an LPS within 72 hours of the LPS genesis in the reference dataset (e.g., ERA5), and if the first four points of the model track that coincide in time with the reference track lie within 4° of the respective reference track points (Froude 2010; Hodges and Emerton 2015; Zhang et al. 2023). A similar criterion was employed by Deoras et al. (2021) to assess LPS tracks in S2S forecasts. In cases where a single reference track is matched by multiple LPS tracks of an ensemble member, at a given model initialization time, the model track with the genesis closest in time to the reference LPS genesis is selected (Zhang et al. 2023). Furthermore, if a model’s LPS track matches multiple reference tracks, we choose the reference track with the minimum mean spatial separation from the corresponding model track (with the mean taken over the common track time points).

Model forecast skill is commonly assessed using two key metrics: the Probability of Detection (POD) and the False Alarm Ratio (FAR). They are defined as:

$$\mathrm{POD} = \frac{\mathrm{hit}}{\mathrm{hit} + \mathrm{miss}}, \tag{1}$$

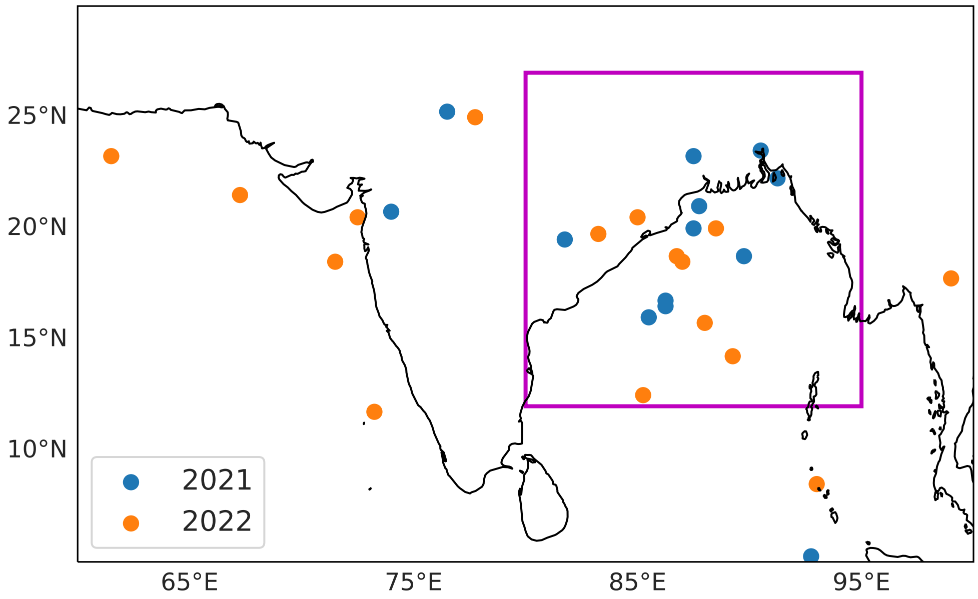

Fig. 1. The locations of LPS genesis in ERA5 over South Asia during June to September in 2021 and 2022. Climatologically (1979-2022), about 70% of South Asian LPS formed within the magenta box, which spans 12°N-27°N, 80°E-95°E.

$$\mathrm{FAR} = \frac{\text{false alarm}}{\text{hit} + \text{false alarm}}. \tag{2}$$

Here, a "false alarm" occurs when a forecast LPS genesis does not match an observed genesis event, while a "miss" indicates situations where the model failed to predict an observed genesis event (Halperin et al. 2013). For a perfect forecast, the POD would be unity, indicating that all events were correctly detected, while the FAR would be zero, indicating no false alarms (Halperin et al. 2016). FAR values are typically plotted as a function of forecast lead time, defined as time elapsed since the model's initialization, and we follow that practice here. As the POD is defined with respect to observed events, we express that metric as a function of time prior to observed genesis (Zhang et al. 2023).

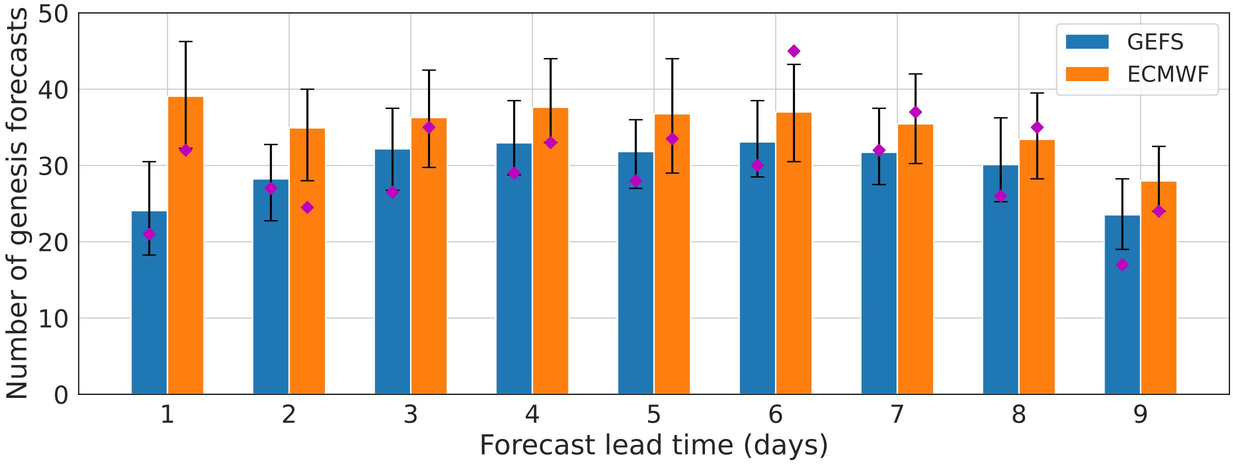

FIG. 2. The average number of LPS genesis events forecast by each ensemble member, from the GEFS and ECMWF models for the period of June to September in 2021-2022, shown as a function of forecast lead time. Vertical bars represent the mean of all ensemble members, while error bars indicate the range between the 5^{th} and 95^{th} percentiles of the ensemble members. The mean values from the control runs are represented by purple diamonds. The counts shown here include both model forecasts of genesis that match ERA5 genesis and those that do not match. The genesis counts have been averaged over the twice-daily model initializations.

3. Results

a. Total LPS counts

For the period 2021-2022, 31 LPS genesis events were identified in the ERA5 reanalysis data, with 15 forming in 2021 and 16 in 2022 (Figure 1). As is typical for South Asian monsoon LPS, most of these genesis events occurred in the northwestern Bay of Bengal and immediately adjacent land regions, although several LPS formed over the Arabian Sea, the southern Bay of Bengal, and other South Asian land regions. As a first step in comparing the NWP ensembles with the reanalysis, we count the number of genesis events occurring in each 24-hour period of forecast lead time (i.e., time since model initialization, as described in Section c). The control forecast and the perturbed ensemble members predict about 20-40 genesis events on average (encompassing the observed count of 31), with more events in the ECMWF model and no clear linear dependence on forecast lead time in either model (Figure 2). This confirms that neither model is excessively more or less cyclogenetic than observations, and neither has a strong trend over forecast lead time in its tendency to form LPS. However, the proximity of the model genesis counts to the observed number of 31 does not indicate that the forecasts were accurate because the forecasts include successful predictions, false alarms, and instances where the models failed to predict LPS genesis. Consequently, even if the total number of forecast genesis events remains relatively constant over forecast lead times (as shown in Figure 2), the models' performance may deteriorate at longer forecast lead times, e.g., with more false alarms and missed genesis events. Interestingly, the average control runs produced fewer genesis events than the ensemble members at most forecast lead times; a similar behavior has been documented for TC genesis forecasts, which Zhang et al. (2023) speculate arises from the perturbations added to the ensemble members.

b. Genesis skill of individual forecasts

Before evaluating the skill of the perturbed ensemble members, we first assess the skill of the control runs. The control run of the GEFS model successfully forecasts about half the observed LPS genesis events, up to 2 days before their actual occurrence (Figure 3a). As expected, the POD (probability of detection, defined in Section c) decreases as the forecast is made progressively earlier in time from the observed genesis, with POD values dropping from around 0.55 about a day prior to the event to only about 0.2 around 8 days before the event. A similar trend occurs for the ECMWF control run, although its POD values are, on average, about 10% higher than those of the GEFS at nearly all lead times.

In addition to having a high probability of forecasting observed LPS genesis, skillful models should also have a low number of false alarms. In the GEFS control run, the FAR is around 0.5 at a forecast lead time of 2 days and increases for longer model forecast lead times, reaching 0.8 for an 8-day forecast lead time (Figure 3b). The ECMWF control run shows comparable FAR values to the GEFS control at most forecast lead times, although the ECMWF may have lower FAR values for forecast lead times of 4–8 days.

We now evaluate the skill of the perturbed ensemble members in forecasting genesis. In both models, the mean POD across all these ensemble members is consistently lower than that of the control run, which exhibits about 5%-10% greater skill in forecasting genesis events (Figure 3a; this difference is statistically significant at a 95% confidence level, as determined by a one-sample t-test to assess whether the POD averaged across the ensemble members differs from the POD of the control run. This holds true across nearly all times prior to observed genesis. Despite generally

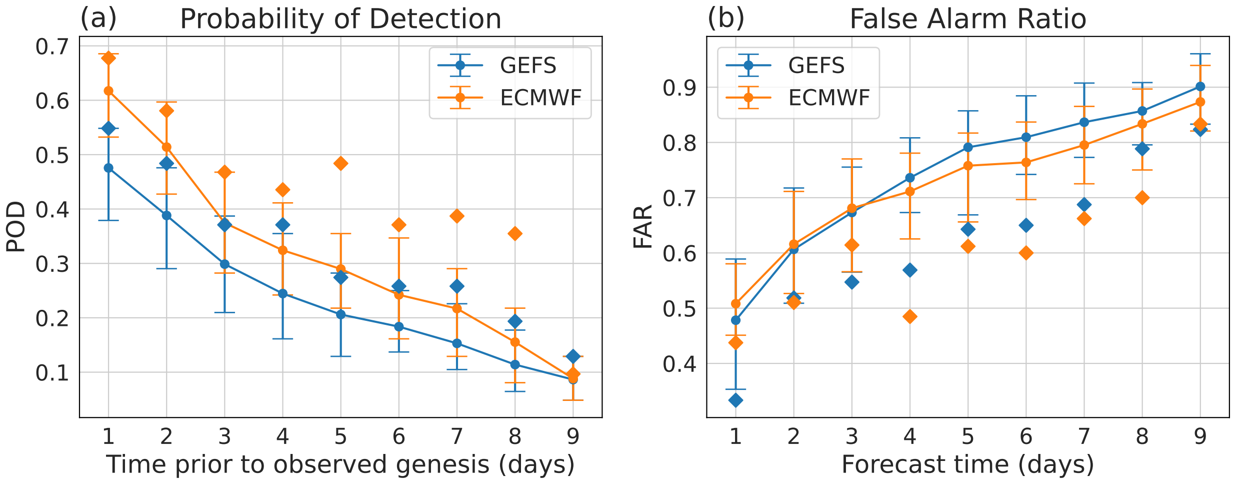

Fig. 3. (a) Probability of Detection (POD) and (b) False Alarm Ratio (FAR) for the GEFS and ECMWF ensemble models during June to September of 2021-2022. POD is shown a function of the time prior to observed genesis, and FAR is shown as a function of model forecast lead time. Solid lines represent the mean of all ensemble members, with error bars indicating the range between the 5^{th} and 95^{th} percentiles of the ensemble members. Values from the control run are depicted as diamonds.

forecasting fewer genesis events than the perturbed ensemble members (Figure 2), the control runs outperform 95% of the ensemble members in detecting observed LPS in both models, at most times prior to observed genesis. As was true for the control runs, the ensemble mean POD of the ECMWF model is about 5%-10% higher than that of the GEFS model, across nearly all times.

Alongside superior POD values, the control runs of both models exhibit lower FARs compared to the perturbed ensemble members, a difference that is again statistically significant (at all forecast lead times; Figure 3b). For the ensemble means, the FAR is comparable across both models, particularly during the initial 3 days of forecast lead time. Beyond this period, the ECMWF model has slightly lower FAR values.

c. Consensus ensemble forecasts

Although individual members of the perturbed ensemble exhibit lower skill, on average, than the control forecast, the ensemble as a whole may provide useful predictive skill. We test this idea by constructing a consensus ensemble forecast, issuing a prediction of genesis when at least 20% of the individual ensemble members predict LPS genesis. This follows similar practices used

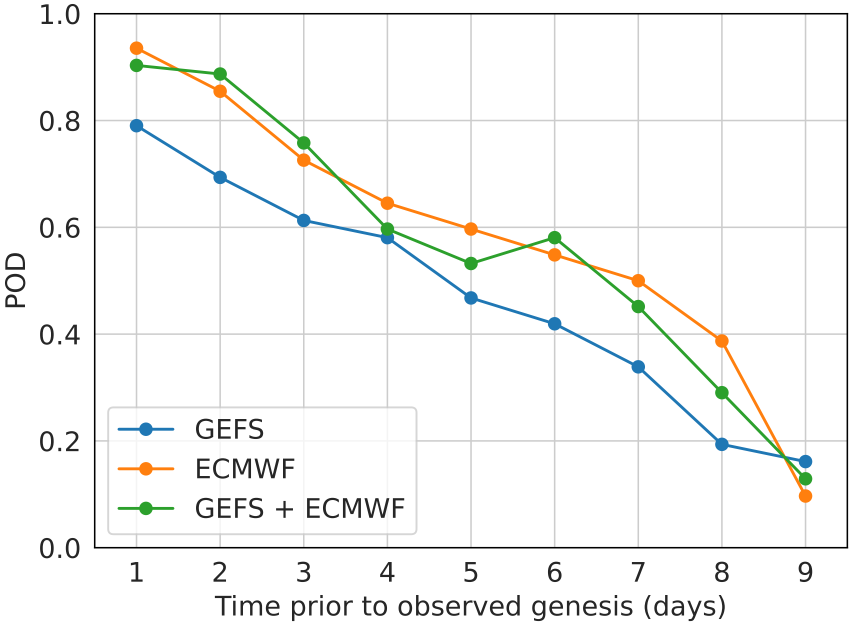

Fig. 4. The Probability of Detection (POD) for consensus ensemble forecasts by the GEFS and ECMWF models during the period 2021-2022. In these consensus forecasts, a prediction of genesis is issued a forecast when at least 20% of the individual members of an ensemble predict the formation of an LPS. The POD of the multi-model super ensemble is shown in green; in that method, a forecast of genesis is issued when at least 20% of individual members among the combined GEFS and ECMWF ensemble predict LPS genesis.

for TC genesis, although rather than using a minimum absolute count of 5 or 8 members as the required threshold for a genesis prediction (Froude 2010; Hodges and Emerton 2015), we opted for a percentage-based threshold to avoid biasing the results toward the model with a larger number of ensemble members. Since this sort of consensus ensemble forecast cannot be easily adapted to give false alarm ratios (Zhang et al. 2023), we only show its POD (Figure 4). For both the GEFS and ECMWF models, the POD of this consensus ensemble exceeds that of the control run. For instance, three days before observed genesis, the consensus ensemble POD for the GEFS model reaches 0.6, compared to the POD of 0.38 for the control run and 0.3 for the mean ensemble member (Figure 3a). Similarly, at a 3-day lead time, the ECMWF model demonstrates a consensus ensemble POD that exceeds its control run by almost 0.3 and its mean ensemble member by almost 0.4. One expects the POD for these consensus ensemble forecasts to increase as one decreases the threshold fraction of ensemble members required for a genesis forecast; no useful skill would be achieved if one attained a POD of 1.0 using a model with an extremely large ensemble size and an extremely small threshold fraction. However, the POD curves shown in Figure 4 constitute a quantitative metric that can be used by forecasters when interpreting ensemble genesis forecasts at particular lead times.

We additionally evaluate whether a multi-model consensus ensemble forecast of genesis might be useful, as has been proposed for TC genesis (Zhang et al. 2023). For monsoon LPS, the multi-model super ensemble consists of the ECMWF and GEFS ensembles, with a super ensemble genesis forecast issued if the same fraction (20%) of ensemble members individually predict genesis. The POD of this super ensemble consensus forecast is approximately the same as the ECMWF model alone, with the ECMWF model having a higher POD than the GEFS (Figure 4). The recent study that proposed this practice for TC genesis (Zhang et al. 2023) used a fixed 5-member threshold for defining consensus forecasts of TC genesis, which advantages the multi-model super ensemble because of its larger number of ensemble members. We also see the super ensemble having a higher POD when using a fixed 5-member threshold (not shown), but that advantage disappears when a percentage-based threshold is used.

d. Ensemble-based probabilistic genesis forecasts over the Bay of Bengal

While monsoon LPS have widespread impacts across South Asia, about 70% of these storms form over the northwestern Bay of Bengal and its adjacent land areas (Figure 1). This leads us to investigate the performance of probabilistic ensemble-based forecasts of LPS genesis in that geographic region. We choose as a domain a 15° × 15° box that encompasses the main genesis region (magenta box in Figure 1), and focus on evaluating the models' ability to forecast LPS formation anywhere in this region within a 5-day window after model initialization. In particular, we aim to ascertain whether the fraction of ensemble members predicting genesis can be leveraged to produce an accurate forecast of the likelihood of LPS formation. In previous work focused on TC genesis, the fraction of ensemble members simulating genesis was not found to be a reliable indicator for the probability of genesis (Majumdar and Torn 2014; Tsai et al. 2011; Zhang et al. 2023). However, there might be higher utility of such probabilistic forecasts for monsoon LPS genesis, given their distinct dynamics and the specific task of detecting genesis within a 5-day window in the confined region centered over the northwestern Bay of Bengal.

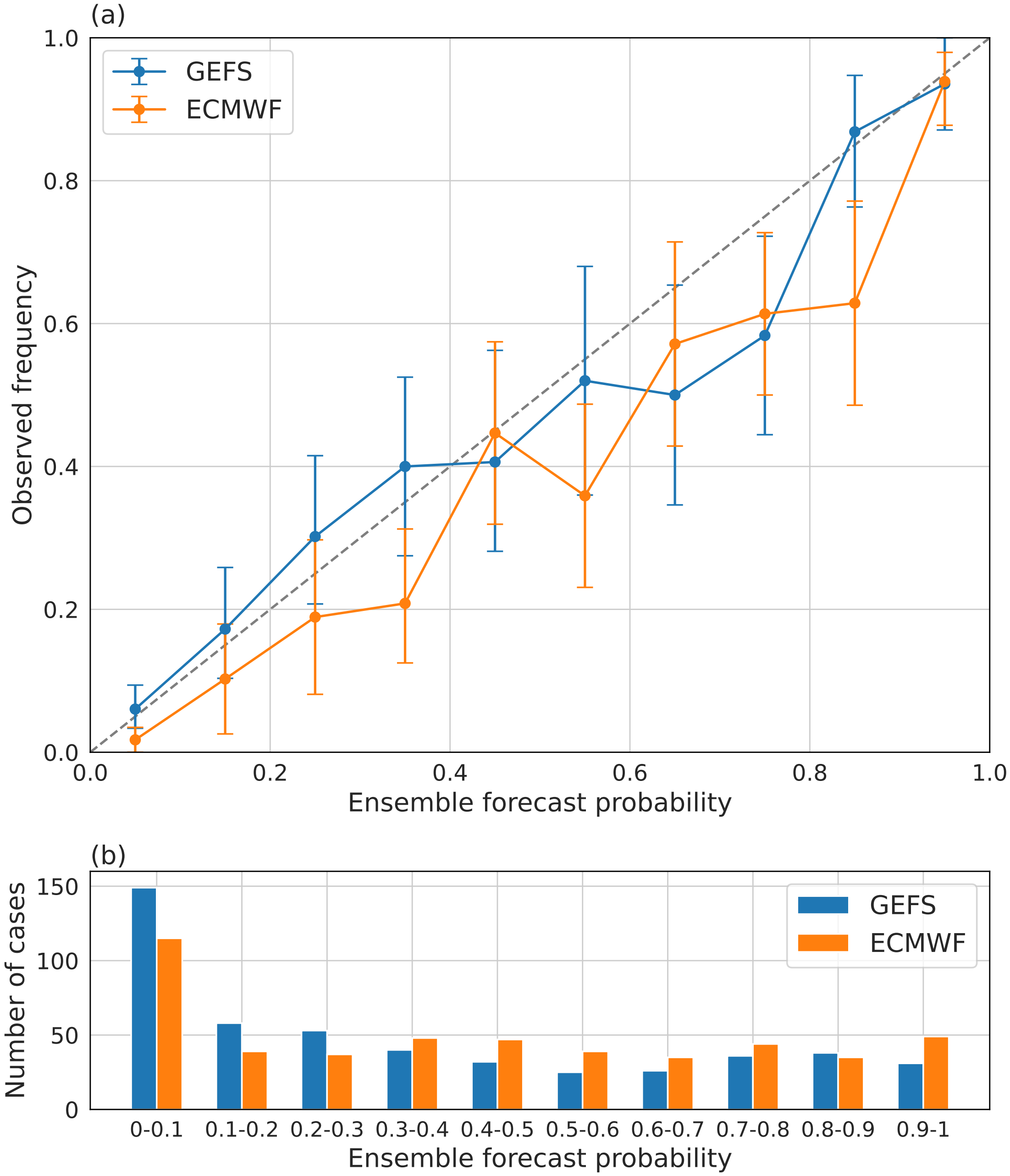

FIG. 5. (a) Reliability diagram showing the ensemble forecast probability of LPS formation during the 5 days after model initialization, compared to the observed frequency of LPS occurrence. This analysis is performed for genesis events in the region delineated by the magenta box in Figure 1. A perfectly calibrated forecast would align along the dashed grey diagonal line. Error bars represent the 90% confidence interval obtained from bootstrap resampling with 10,000 replicates. (b) Sample sizes within the bins used in (a); the reliability diagram is constructed by binning the model predictions into 10 equally spaced bins, covering the range of fractions of ensemble members forecasting genesis, from 0 to 1.

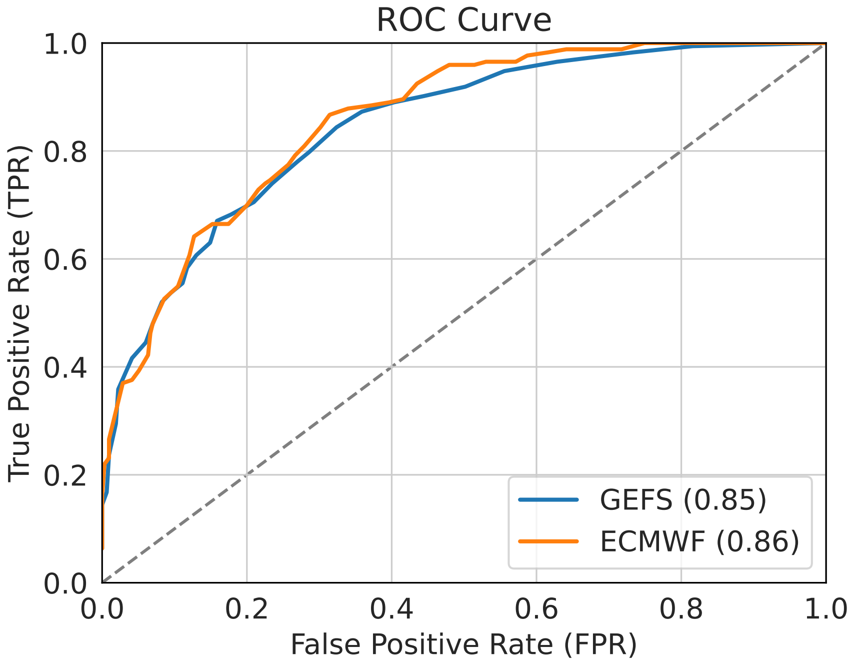

FIG. 6. Receiver Operating Characteristics (ROC) curve for the probability of LPS formation during the 5 days after model initialization, in the region outlined by the magenta box in Figure 1. True Positive Rate (TPR) measures the proportion of actual genesis events correctly forecast by the model, while the False Positive Rate (FPR) measures the proportion of actual negative cases incorrectly forecast as a genesis event by the model. A model with no predictive skill would have its ROC curve coinciding with the grey dashed line, resulting in an Area Under the Curve (AUC) of 0.5. An AUC greater than 0.5 is often taken to indicate skill, with a maximum value of 1 representing a perfect forecast. The AUC values for the GEFS and ECMWF models are enclosed in parentheses within the legend.

To assess the performance of our model ensembles, we use a reliability plot (Figure 5a) to gauge how effectively the forecast probabilities match observed outcomes. In this context, a model's forecast probability represents the proportion of ensemble members predicting the genesis of an LPS within the first 5 days after model initialization, in our 15° × 15° region centered on the northwestern Bay of Bengal. In an ideal scenario where the forecast is perfectly calibrated, the curve would align precisely along the diagonal (dashed grey line in Figure 5a). When the curve falls below this diagonal, it indicates a tendency to overestimate the likelihood of genesis, while if it rises above it suggests an inclination to underestimate. Both model ensembles produce forecasts that lie fairly close to the ideal calibration line; all mean values for the ECMWF ensemble lie below the line, but the offset is modest, especially compared to the uncertainties. This indicates that the raw fraction of the ensemble members which forecast genesis can be interpreted, without bias correction, as a probability of observed genesis. While a substantial number of cases fall within the lowest probability bin, corresponding to a negligible likelihood of LPS formation, more than half the time there is a larger than 10% probability of genesis (Figure 5b). An important caveat is that this reliability plot compares the predicted likelihood of genesis with the observed likelihood of genesis in a particular region, and does not assess whether specific observed genesis events were correctly forecast by the models. It is thus a diagnostic of forecast skill that complements the event-based POD and FAR metrics shown in Figures 3 and 4.

To further evaluate the performance of ensemble-based forecasts of genesis in our 15° × 15° box centered over the Bay of Bengal, we employ a Receiver Operating Characteristic (ROC) curve (Figure 6). This provides an assessment of the models' performance in making a binary classification (i.e., genesis or no genesis in the magenta box of Figure 1) across various classification thresholds, with the threshold in this case being the ensemble forecast probability required to issue a genesis prediction. When the threshold is set too low, for instance, issuing a positive forecast for genesis when only one or more ensemble member predicts it, the ensemble model can effectively detect many of the observed events but also generates a large number of false positives. This scenario would produce a data point in the upper right portion of the ROC space. Conversely, using a very high threshold may reduce false alarms but at the expense of missing many observed genesis events, resulting in a data point in the lower left part of the ROC space. A model with no predictive skill would have its ROC curve coinciding with the grey dashed line, resulting in an Area Under the Curve (AUC) of 0.5. An AUC greater than 0.5 has been taken to indicate skillful forecasts, with a maximum value of 1 representing a perfect forecast. The AUC for both ensemble models is around 0.85, suggesting that both exhibit similar skill in predicting the formation of LPS within 5 days of model initialization in the main development region centered over the Bay of Bengal.

4. Summary and discussions

Monsoon LPS are the main rain-bearing weather system of much South Asian land (Godbole 1977; Yoon and Chen 2005), producing a large fraction of that region's heavy precipitation events

(Ajayamohan et al. 2010; Fletcher et al. 2018) and disasters (Suhas et al. 2023). Accurate and timely forecasts of LPS genesis are thus essential. However, few studies have assessed the skill of NWP models in forecasting LPS behavior, with none, to our knowledge, specifically examining genesis prediction. In a related study (cite our BAMS paper), we evaluated the performance of control runs of the GEFS and ECMWF models in forecasting LPS, including their genesis. Here, we extend that analysis by assessing the genesis prediction skill of the perturbed ensemble members and the possible utility of ensemble statistics.

A key result is that the control runs in both models produced more skillful forecasts than the average ensemble member, with POD and FAR values that surpassed even the 95th percentile of the ensemble member values of those metrics. Two days before observed LPS genesis events, the GEFS control run successfully forecast roughly half of the events. The ECMWF model forecast about 10% more of these observed events. About 50% of the forecast LPS in the control runs of both models were false alarms for forecasts with 2-day lead time. The average perturbed ensemble member had a POD value about 10 points lower, and an FAR roughly 10 points higher for lead times up to 3 days, with these differences increasing even more at lead times of 4–7 days.

While individual ensemble members exhibited lower average skill than the control run, the ensemble as a whole provided useful information complementary to that of the control. Specifically, we tested a consensus method that predicts genesis when a minimum of 20% of ensemble members forecast an LPS, and the resulting consensus POD values consistently surpassed those of the control run and the ensemble mean across all forecast lead times (false alarms are assessed in the context of the ROC metric discussed below). This follows a similar approach taken for TC genesis Zhang et al. (2023), although unlike the TC case, we did not find a multi-model super ensemble consensus forecast to have superior skill for LPS genesis.

Since the majority of South Asian LPS form in a small area over the Bay of Bengal and adjacent land, we evaluated the ensemble models' probabilistic predictions of genesis within this region. We found that the fraction of ensemble members forecasting genesis could be interpreted, without correction, as a likelihood of genesis. Specifically, probabilistic forecasts from both ensemble models fell close to the 1:1 line on a reliability diagram for forecasts over the five days after model initialization. This differs substantially from the behavior seen in past studies of TC genesis forecasts, which under- and over-predict the likelihood of genesis in such probabilistic forecasts Majumdar and Torn (2014); Zhang et al. (2023) .

All of these results suggest that NWP models can be useful in forecasting the formation of monsoon LPS over South Asia. They demonstrate the utility of ensemble models for probabilistic forecasts of genesis, while highlighting the greater skill of the control runs for deterministic genesis predictions. There are, of course, important caveats. The analysis was limited to a two-year period, 2021-2022. This constraint was primarily imposed by data storage limitations, but it was also necessary to accommodate the increased in ensemble members from 20 to 30 in the GEFS model in late 2020. Further evaluation with longer periods and additional models will help establish the robustness of these findings. Additionally, this study did not examine the various factors that affect forecast skill in the ensemble models; if that skill proves to be state-dependent, for example, additional information could enhance the utility of forecasts beyond that found here.

APPENDIX

A1. Evaluation of genesis skill using MERRA-2

This study evaluated the performance of NWP models in forecasting LPS genesis by comparison with genesis events from the ERA5 reanalysis. It seems possible that use of ERA5 as the reference dataset may give an advantage to the ECMWF model. To address this possibility, we repeated some analyses using the MERRA-2 reanalysis as an alternative reference dataset. Over South Asia, 45 LPS genesis events were identified in the MERRA-2 dataset during June to September of 2021-2022, which is more than the 31 genesis events in ERA5. A similarly higher number of LPS genesis events was also seen in MERRA-2 by Vishnu et al. (2020) (see their Figure 6), while the ERA5 count was closer to the subjectively analyzed data track datasets compiled by Sikka (2006). The higher number of LPS events in the MERRA-2 dataset reduces the probability of detection in both NWP models (compare Figure A1a with Figure 3a). Nevertheless, the ECMWF model still exhibits about 10% higher skill in successfully predicting LPS genesis. False alarms also exhibit a similar behavior as when compared to ERA5, with the GEFS model having lower FAR values for the first day of lead time, while the ECMWF model performs better beyond that. Overall, the POD

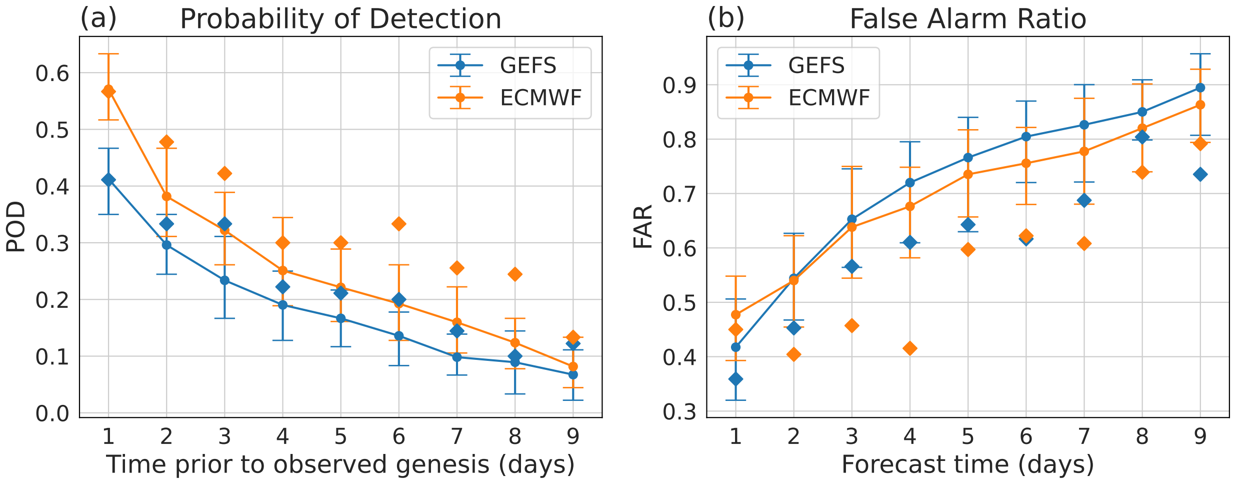

Fig. A1. As in Figure 3, but using MERRA-2 instead of ERA5 as a reference. (a) Probability of Detection (POD) and (b) False Alarm Ratio (FAR) for the GEFS and ECMWF ensemble models during June to September of 2021-2022, using LPS genesis data from MERRA-2 as a reference. POD is shown a a function of the time prior to observed genesis, and FAR is shown as a function of model forecast lead time. Solid lines represent the mean of all ensemble members, with error bars indicating the range between the 5^{th} and 95^{th} percentiles of the ensemble members. Values from the control run are depicted as diamonds.

and FAR distributions obtained using MERRA-2 as a reference confirm the general conclusions discussed in the main text.

References

Ajayamohan, R., W. J. Merryfield, and V. V. Kharin, 2010: Increasing trend of synoptic activity and its relationship with extreme rain events over central India. *Journal of Climate*, **23 (4)**, 1004–1013.

Boos, W., J. Hurley, and V. Murthy, 2015: Adiabatic westward drift of Indian monsoon depressions. *Quarterly Journal of the Royal Meteorological Society*, **141 (689)**, 1035–1048.

Bougeault, P., and Coauthors, 2010: The thorpex interactive grand global ensemble. *Bulletin of the American Meteorological Society*, **91 (8)**, 1059–1072.

Briegel, L. M., and W. M. Frank, 1997: Large-scale influences on tropical cyclogenesis in the western north pacific. *Monthly Weather Review*, **125 (7)**, 1397–1413.

Cangialosi, J. P., E. Blake, M. DeMaria, A. Penny, A. Latto, E. Rappaport, and V. Tallapragada, 2020: Recent progress in tropical cyclone intensity forecasting at the national hurricane center. *Weather and Forecasting*, **35 (5)**, 1913–1922.

Chan, J. C., and R. H. Kwok, 1999: Tropical cyclone genesis in a global numerical weather prediction model. *Monthly weather review*, **127 (5)**, 611–624.

Cheung, K. K., and R. L. Elsberry, 2002: Tropical cyclone formations over the western north pacific in the navy operational global atmospheric prediction system forecasts. *Weather and forecasting*, **17 (4)**, 800–820.

Deoras, A., K. M. Hunt, and A. G. Turner, 2021: Comparison of the prediction of indian monsoon low pressure systems by subseasonal-to-seasonal prediction models. *Weather and Forecasting*, **36 (3)**, 859–877.

Elsberry, R. L., H.-C. Tsai, and M. S. Jordan, 2014: Extended-range forecasts of atlantic tropical cyclone events during 2012 using the ecmwf 32-day ensemble predictions. *Weather and forecasting*, **29 (2)**, 271–288.

Fletcher, J. K., D. J. Parker, K. M. Hunt, G. Vishwanathan, and M. Govindankutty, 2018: The interaction of Indian monsoon depressions with northwesterly midlevel dry intrusions. *Monthly Weather Review*, **146 (3)**, 679–693.

Froude, L. S., 2010: Tigge: Comparison of the prediction of northern hemisphere extratropical cyclones by different ensemble prediction systems. *Weather and Forecasting*, **25 (3)**, 819–836.

Godbole, R. V., 1977: The composite structure of the monsoon depression. *Tellus*, **29 (1)**, 25–40.

Goerss, J. S., C. R. Sampson, and J. M. Gross, 2004: A history of western north pacific tropical cyclone track forecast skill. *Weather and Forecasting*, **19 (3)**, 633–638.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, and J. H. Cossuth, 2016: Verification of tropical cyclone genesis forecasts from global numerical models: Comparisons between the north atlantic and eastern north pacific basins. *Weather and Forecasting*, **31 (3)**, 947–955.

Halperin, D. J., H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, 2013: An evaluation of tropical cyclone genesis forecasts from global numerical models. *Weather and Forecasting*, **28 (6)**, 1423–1445.

Halperin, D. J., A. B. Penny, and R. E. Hart, 2020: A comparison of tropical cyclone genesis forecast verification from three global forecast system (gfs) operational configurations. *Weather and Forecasting*, **35 (5)**, 1801–1815.

Hersbach, H., and Coauthors, 2020: The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146 (730)**, 1999–2049.

Hodges, K. I., and R. Emerton, 2015: The prediction of northern hemisphere tropical cyclone extended life cycles by the ecmwf ensemble and deterministic prediction systems. part i: Tropical cyclone stage. *Monthly Weather Review*, **143 (12)**, 5091–5114.

Hunt, K. M., and J. K. Fletcher, 2019: The relationship between Indian monsoon rainfall and low-pressure systems. *Climate Dynamics*, **53 (3)**, 1859–1871.

Hurley, J. V., and W. R. Boos, 2015: A global climatology of monsoon low-pressure systems. *Quarterly Journal of the Royal Meteorological Society*, **141 (689)**, 1049–1064.

Jaiswal, N., C. Kishtawal, S. Bhomia, and P. Pal, 2016: Multi-model ensemble-based probabilistic prediction of tropical cyclogenesis using tigge model forecasts. *Meteorology and Atmospheric Physics*, **128**, 601–611.

Komaromi, W. A., and S. J. Majumdar, 2015: Ensemble-based error and predictability metrics associated with tropical cyclogenesis. part ii: Wave-relative framework. *Monthly Weather Review*, **143 (5)**, 1665–1686.

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. *Weather and Forecasting*, **33 (4)**, 967–988.

Liang, M., J. C. Chan, J. Xu, and M. Yamaguchi, 2021: Numerical prediction of tropical cyclogenesis part i: Evaluation of model performance. *Quarterly Journal of the Royal Meteorological Society*, **147 (736)**, 1626–1641.

Majumdar, S. J., and R. D. Torn, 2014: Probabilistic verification of global and mesoscale ensemble forecasts of tropical cyclogenesis. *Weather and Forecasting*, **29 (5)**, 1181–1198.

McTaggart-Cowan, R., T. J. Galarneau Jr, L. F. Bosart, R. W. Moore, and O. Martius, 2013: A global climatology of baroclinically influenced tropical cyclogenesis. *Monthly weather review*, **141 (6)**, 1963–1989.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, **122 (529)**, 73–119.

Nakano, M., M. Sawada, T. Nasuno, and M. Satoh, 2015: Intraseasonal variability and tropical cyclogenesis in the western north pacific simulated by a global nonhydrostatic atmospheric model. *Geophysical Research Letters*, **42 (2)**, 565–571.

Palmer, T., 2019: The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, **145**, 12–24.

Sarkar, S., P. Mukhopadhyay, S. Dutta, R. Phani Murali Krishna, R. D. Kanase, V. Prasad, and M. S. Deshpande, 2021: Gfs model fidelity in capturing the transition of low-pressure area to monsoon depression. *Quarterly Journal of the Royal Meteorological Society*, **147 (738)**, 2625–2637.

Sikka, D., 1977: Some aspects of the life history, structure and movement of monsoon depressions. *Monsoon dynamics*, Springer, 1501–1529.

Sikka, D., 2006: *A study on the monsoon low pressure systems over the Indian region and their relationship with drought and excess monsoon seasonal rainfall*. Center for Ocean-Land-Atmosphere Studies.

Suhas, D. L., N. Ramesh, R. M. Kripa, and W. R. Boos, 2023: Influence of monsoon low pressure systems on south asian disasters and implications for disaster prediction. *npj Climate and Atmospheric Science*, **6 (1)**, 48.

Swinbank, R., and Coauthors, 2016: The tigge project and its achievements. *Bulletin of the American Meteorological Society*, **97 (1)**, 49–67.

Tippett, M. K., S. J. Camargo, and A. H. Sobel, 2011: A poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. *Journal of Climate*, **24 (9)**, 2335–2357.

Titley, H., M. Yamaguchi, and L. Magnusson, 2019: Current and potential use of ensemble forecasts in operational tc forecasting: results from a global forecaster survey. *Tropical Cyclone Research and Review*, **8 (3)**, 166–180.

Tsai, H.-C., K.-C. Lu, R. L. Elsberry, M.-M. Lu, and C.-H. Sui, 2011: Tropical cyclone–like vortices detection in the ncep 16-day ensemble system over the western north pacific in 2008: Application and forecast evaluation. *Weather and forecasting*, **26 (1)**, 77–93.

Ullrich, P. A., C. M. Zarzycki, E. E. McClenny, M. C. Pinheiro, A. M. Stansfield, and K. A. Reed, 2021: Tempestextremes v2. 1: a community framework for feature detection, tracking, and analysis in large datasets. *Geoscientific Model Development*, **14 (8)**, 5023–5048.

Vishnu, S., W. Boos, P. Ullrich, and T. O'Brien, 2020: Assessing historical variability of south asian monsoon lows and depressions with an optimized tracking algorithm. *Journal of Geophysical Research: Atmospheres*.

Wang, Z., W. Li, M. S. Peng, X. Jiang, R. McTaggart-Cowan, and C. A. Davis, 2018: Predictive skill and predictability of north atlantic tropical cyclogenesis in different synoptic flow regimes. *Journal of the Atmospheric Sciences*, **75 (1)**, 361–378.

Yamaguchi, M., and N. Koide, 2017: Tropical cyclone genesis guidance using the early stage dvorak analysis and global ensembles. *Weather and Forecasting*, **32 (6)**, 2133–2141.

Yamaguchi, M., F. Vitart, S. T. Lang, L. Magnusson, R. L. Elsberry, G. Elliott, M. Kyouda, and T. Nakazawa, 2015: Global distribution of the skill of tropical cyclone activity forecasts on short-to medium-range time scales. *Weather and Forecasting*, **30 (6)**, 1695–1709.

Yoon, J.-H., and T.-C. Chen, 2005: Water vapor budget of the Indian monsoon depression. *Tellus A: Dynamic Meteorology and Oceanography*, **57 (5)**, 770–782.

Zhang, X., J. Fang, and Z. Yu, 2023: The forecast skill of tropical cyclone genesis in two global ensembles. *Weather and Forecasting*, **38 (1)**, 83–97.

Zhou, X., and Coauthors, 2022: The development of the ncep global ensemble forecast system version 12. *Weather and Forecasting*, **37 (6)**, 1069–1084.