Journal of *Marine Science and Engineering*

MDPI

Article

Using the Debiased Brier Skill Score to Evaluate S2S Tropical Cyclone Forecasting

Yuanben Li [1,2], Xiaochun Wang [1,2,*], Bingke Zhao [3], Ming Ying [3], Yimin Liu [4] and Frederic Vitart [5]

[1] State Key Laboratory of Climate System Prediction and Risk Management, Nanjing University of Information Science and Technology, Nanjing 210044, China; liyuanben-13@outlook.com
[2] School of Marine Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] Shanghai Typhoon Institute, Chinese Meteorological Administration, Shanghai 200030, China; zhaobk@typhoon.org.cn (B.Z.); yingm@typhoon.org.cn (M.Y.)
[4] Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100045, China; lym@lasg.iap.ac.cn
[5] European Center for Medium-Range Weather Forecasts, Reading RG2 9AX, UK; frederic.vitart@ecmwf.int
* Correspondence: xcwang@nuist.edu.cn

Abstract: To evaluate tropical cyclone forecasting on synoptic timescale, tracking and intensity are used. On subseasonal to seasonal (S2S) timescale, what aspects of tropical cyclones should be predicted and how to evaluate forecasting skills still remain open questions. Following our previous work, which proposed using daily tropical cyclone probability (DTCP) as a measure of tropical cyclone activity and the debiased Brier skill score (DBSS) to evaluate tropical cyclone forecasting on S2S timescale, the present research investigates the influence of several factors that may influence the use of DTCP and the DBSS framework. These factors are the forecast time window, tropical cyclone influence radius, evaluation region, forecast sample, and how the Brier score for the reference climate forecast is computed. The influence of these factors is discussed based on the output of the S2S prediction project database and comparison of the DBSS when the above factors are changed individually. Changes in the forecast time window, evaluation region, and tropical cyclone influence radius can change the DTCP. The larger the tropical cyclone influence radius and the longer the forecast time window, the larger the DTCP will be. However, the spatially averaged DBSS changes very little. Using estimated Brier score for reference climate forecast can cause variation due to limited forecast samples. It is recommended to use the theoretical value of the Brier score for reference climate forecasting, instead of its estimation.

Keywords: daily tropical cyclone probability; Brier score; debiased Brier skill score; tropical cyclone influence radius; forecast time window; reference climate forecast; evaluation region; forecast sample

check for updates

Academic Editor: Han Soo Lee

Received: 27 April 2025
Revised: 22 May 2025
Accepted: 22 May 2025
Published: 24 May 2025

Citation: Li, Y.; Wang, X.; Zhao, B.; Ying, M.; Liu, Y.; Vitart, F. Using the Debiased Brier Skill Score to Evaluate S2S Tropical Cyclone Forecasting. *J. Mar. Sci. Eng.* **2025**, *13*, 1035. https://doi.org/10.3390/jmse13061035

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The annual tropical cyclone (TC)-related coastal risk and associated destructive losses continue to rise [1–4]. Due to improvements in numerical modeling, forecasting skills for extreme events have been substantially enhanced [5], contributing to better TC monitoring.

Subseasonal-to-seasonal (S2S) prediction is generally defined as including forecasts spanning beyond 14 days but less than a season (14 days to 3 months). Previously, S2S forecasting was often referred to as a "predictability desert", representing a significant gap between weather and climate prediction [6]. This gap arises because the S2S scale is too long for the atmosphere to retain sufficient memory of initial conditions, unlike the

synoptic scale. Conversely, relative to seasonal scale, the S2S scale is too short to fully capture the coupling signals of boundary condition anomalies. This disparity is further reflected in the lack of operational forecasting products bridging the gap between weather and seasonal prediction.

However, recent technological advancements and growing demand for S2S predictions have renewed people's interest in this field [7]. An increasing number of research institutions are now focusing on S2S forecasting, and improvements in numerical models have uncovered new sources of predictability, significantly enhancing forecasting skills. Under the leadership of the World Meteorological Organization (WMO), multiple countries and research institutions have collaborated on S2S prediction across various domains, particularly for extreme events, such as in TC forecasting. Nevertheless, challenges remain, including inconsistencies in the TC variables used across different institutions.

S2S TC forecasting plays a crucial role between synoptic forecasting and seasonal forecasting. However, the transition of S2S TC forecasting from research to operational applications has been slow, partly due to the absence of standardized operational practice [8]. To advance research, improve forecasting skills, and expand the scope of operational S2S forecast, the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP) under the WMO jointly launched a ten-year (2013–2023) initiative known as the S2S Prediction Project [9,10]. The project aimed to bridge the gap between medium-range weather forecasting and seasonal prediction while improving the understanding of S2S predictability sources. Phase I of the S2S Prediction Project (2013–2017) focused on establishing the S2S database, which included hindcast data from 11 centers [11]. The project was subsequently extended into Phase II (2018–2023). A key component of Phase II was real-time prediction (RTP), which aimed to provide real-time S2S forecasts for broader geographical regions and applications [12]. Though the Phase II project was officially concluded at the end of 2023, many institutions still continue to provide their S2S forecast products.

Various TC variables and evaluation metrics have been used to evaluate S2S TC forecasting. In one early research project, the hindcast from the European Centre for Medium-Range Weather Forecasts (ECMWF) was used to assess weekly TC occurrences in different regions of the Southern Hemisphere, employing the Brier skill score (BSS) as the evaluation metric [13]. Subsequent work [14] introduced a series of TC forecast products from ECMWF, including "strike" probability and TC probability maps. The "strike" area was defined as within 120 km of the TC track. Additionally, probability maps for TC activity (including genesis) were provided for wind speeds exceeding 17 m/s and 33 m/s, calculated over a 7-day time window, with a TC influence radius of 300 km. A TC track verification method was developed using ECMWF forecasts in the North Atlantic (ATL), matching them with best-track datasets [15]. Although the model captured nearly all TCs during two TC seasons, substantial errors persisted, which underscores the inapplicability of using TC track for S2S prediction.

Probabilistic forecasts for TC activity (genesis and movement) using a 3-day sliding window and a 300 km TC influence radius were validated with BSS within 14 days [16]. The results demonstrated that models could provide skillful TC predictions beyond the second week, outperforming climatological forecasts. Monthly TC forecasts from the Geophysical Fluid Dynamics Laboratory High-Resolution Atmospheric Model were evaluated in ATL by TC frequency and Accumulated Cyclone Energy (ACE) [17]. However, the monthly resolution limited precise description of TC location.

Recent studies evaluating S2S TC forecasts have described TC activity within a "Box" (20° longitude × 15° latitude). Both the probability of TC occurrence and the frequency of genesis have been utilized to assess multi-model hindcasts from the S2S database on a

weekly basis [18]. Evaluations employed the ACE and BSS, using two different reference forecasts including monthly varying climatology. The Australian Bureau of Meteorology's seasonal forecasting systems (ACCESS-S2) were applied to predict weekly global TC activity [19]; the results showed that BSS varied across different time periods, and different calibration methods failed to improve the ACCESS skill. This indicates that differences in forecast samples lead to irreducible systematic variations that affect skill scores. Similarly, multi-week TC forecasts during the 2017–2019 Southern Hemisphere TC seasons were also evaluated using real-time predictions from ACCESS-S [20,21]. It was found that models with larger ensemble sizes, higher spatial resolution, and better initialization schemes exhibited better forecasting skill.

Daily tropical cyclone probability (DTCP) was proposed as a forecast variable and used to systematically evaluate 11 models in the S2S reforecast database for the Western North Pacific (WNP) via debiased Brier skill score (DBSS) and Taylor score [22]. The evaluation of S2S TC activity forecasts using NOAA's Global Ensemble Forecast System and ECMWF hindcasts for the ATL emphasized the impact of ensemble size on the model's skill. [23]. Genesis Potential Index (GPI) and DBSS were applied to assess NASA's GEOS-S2S-2 hindcasts for the ATL over a 30-day lead time [24]. The growing interest in S2S prediction led to the SubX subseasonal experiment [25], designed to integrate multi-model ensembles for S2S forecasting, mainly for the ATL region. SubX combines hindcast data from seven global models, leveraging their strengths to enhance prediction accuracy.

A recent review [26] summarized progress in S2S TC forecasting and highlighted multi-model TC forecast data from various databases. However, methods for evaluating TC forecasts remain diverse. As mentioned earlier, most S2S TC forecast products are probabilistic, but the application of probabilistic methods and the impact of different evaluation criteria on results require further exploration. Therefore, this study examines TC forecast data from four models in the S2S real-time database, testing and evaluating multiple factors influencing the models' forecasting skills. Clarifying the influence of these factors will enhance the societal application of TC forecasting. The WNP is the most active basin for TC globally [27]. On average, approximately 30 TCs occur annually in the WNP, accounting for over 30% of the global annual total of about 80 TCs [28] and contributing to roughly 40% of the worldwide ACE [29]. TC activity in the WNP exhibits remarkable complexity. Given these characteristics, this study used forecast products from the WNP to conduct tests with these factors changed individually, one by one. This paper is organized as follows. Materials and methods for the S2S TC forecasting skills evaluation test are described in Section 2, and the results of different test cases are reported in Section 3, followed by conclusions in Section 4.

2. Materials and Methods

2.1. Observational Tropical Cyclone Data

The International Best Track Archive for Climate Stewardship (IBTrACS) is an integrated dataset of historical TC activities [30]. The dataset combines all available information from buoys, offshore observation platforms, satellites, etc. IBTrACS unifies data from different institutions and different ocean basins, providing a global TC track database. The latest version of the dataset is 4.0, which includes a time interval of 3 h for all TC tracks. This study used data on the position of TC centers, maximum sustained wind speeds, and minimum pressure at cyclone centers. We calculated the observed DTCP from the TC center position information (longitude and latitude).

2.2. S2S Real-Time Database

Table 1 provides the relevant information regarding the real-time forecast models. The 4 models selected were those from the European Centre for Medium-Range Weather Forecasts (ECMWF), the Korea Meteorological Administration (KMA), the United States National Centers for Environmental Prediction (NCEP), and the United Kingdom Met Office (UKMO). These models started providing their real-time products in different years (Column 4), and they also have different ensemble members (Column 5). Moreover, the models have different forecast frequencies (Column 6) and numbers of forecast samples (Column 2). All these details could influence evaluation of the forecasting skill of specific models [26,31]. The forecast lead time also varies (Column 3). Thus, our evaluation test calculation was for forecasts with lead time up to 30 days.

Table 1. Details of 4 models that provided forecasts in the WMO S2S real-time prediction project. These models started to provide their forecasts in different years (Column 4) and include different forecasting samples (Column 2). The frequency of forecasts (Column 6) and numbers of ensemble members (Column 5) are also shown.

System	Forecast Samples	Lead Time (Day)	Period (Year)	Ensemble Size	Frequency
ECMWF	473	46	2015–2023	51	Twice/week
KMA	1656	60	2015–2023	4	Daily
NCEP	1656	44	2015–2023	16	Daily
UKMO	1452	60	2016–2023	4	Daily

Based on previous research [16,18,22], the ECMWF model is generally considered to have better forecasting skills; so, its real-time TC forecasting data are representative for conducting experiments. When testing the impact of forecast frequency, we additionally selected three daily forecast models to make the experiment robust. The same TC tracker was applied to all these S2S models [32]. The models' DTCP was computed based on the results of the TC tracker for each ensemble member and these values were used in our test. Because most TCs happen in summer or autumn, we used only the forecasting samples from May to October for the WNP. Though there are around a dozen institutions that provide their forecasts in real time, only these four were selected for the present research. The reason for this is that the ECMWF model tends to have high S2S TC forecasting accuracy and the models from KMA, NCEP and UKMO produce forecasts daily, providing a larger forecast sample than the other models from the S2S database that have lower frequency (weekly or bi-weekly).

2.3. Daily Tropical Cyclone Probability

When evaluating TC forecasting on the S2S timescale, several factors may influence the evaluation. This study utilized the DTCP proposed previously [22] to find out how these factors affected forecasting skill. DTCP is defined as the probability of TC occurrence within a specific TC influence radius of a location and 1 day forecast time window. We computed DTCP on a $1^\circ \times 1^\circ$ grid and tested the influence of factors such as TC influence radius, forecast time window, etc. The computation of DTCP from observation was based on the IBTrACS data being collected every three hours. When calculating the DTCP from the models, TC center positions predicted by each ensemble member were used to compute the DTCP for that ensemble member, and the final DTCP for the ensemble prediction system was the average of the DTCP fields of all ensemble members in the model. As an example, Figure 1 shows the DTCP predictions of Typhoons Noul and Dolphin made by the KMA model from 2 May to 7 May 2015. Due to the fact that the KMA ensemble

forecasting system has 4 ensemble members, there are 5 possible outcomes for calculating DTCP based on the model (DTCP values of 0, 0.25, 0.5, 0.75, and 1).

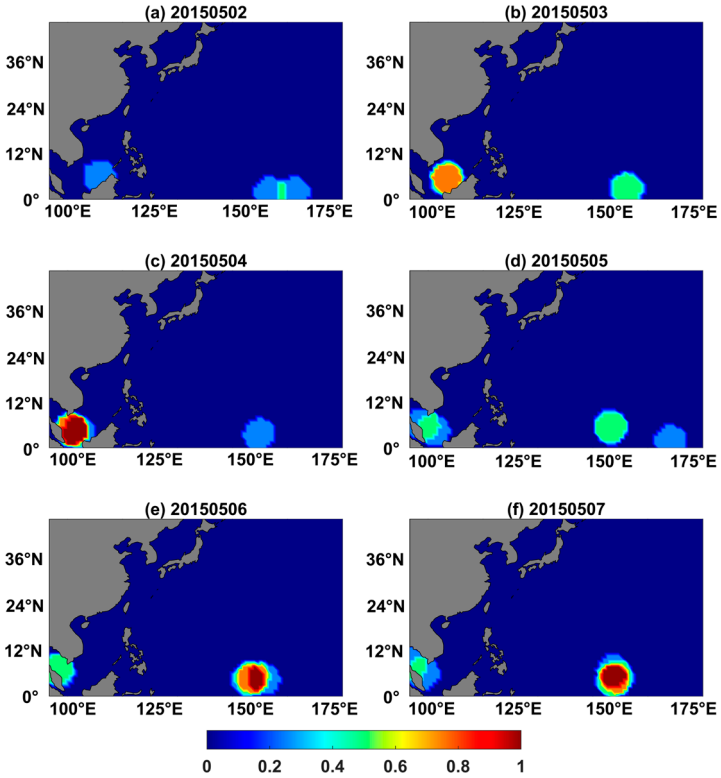


Figure 1. DTCP forecast issued on 1 May 2015 from the KMA model in S2S real-time prediction project from 2 May to 7 May 2015, with a 1-day forecast time window and 500 km TC influence radius.

2.4. Debiased Brier Skill Score

According to the definition of DTCP, the occurrence of a TC can be treated as a dichotomous event similar to precipitation, so it is possible to use the Brier skill score [33] to evaluate forecasting skills. A correction term is added to obtain the debiased Brier skill score, which corrects the negative bias caused by small ensemble members and makes it more suitable for ensemble forecasting systems [34]. The DBSS is defined as follows:

$$\mathrm{DBSS} = 1 - \frac{< \mathrm{BS} >}{< \mathrm{BS_{CL}} > + \mathrm{D}} \tag{1}$$

where <BS> is the Brier score, which is the mean squared error calculated from forecast and observation; <·> is the average over the forecast samples. The $< \mathrm{BS} >$ can be computed directly as follows:

$$< \mathrm{BS} > = \sum_{k=1}^{2} \sum_{j=1}^{N} \frac{\left(\mathrm{Y_{kj}} - \mathrm{O_{kj}}\right)^2}{\mathrm{N}}, \tag{2}$$

where k = 1, 2 is the number of categories (with and without TC in observation); $\mathrm{Y_{kj}}$ is the prediction for category k and sample j; $\mathrm{O_{kj}}$ is the observed DTCP for category k and sample j. The value of $< \mathrm{BS} >$ changes from 0 to 2, with 0 as the perfect prediction and 2 as the situation when the prediction is wrong every time.

$< \mathrm{BS_{CL}} >$ is the Brier score when a reference climate forecast is used; $< \mathrm{BS_{CL}} >$ can be calculated theoretically via the following formula:

$$< \mathrm{BS_{CL_theory}} > = 2\mathrm{p}\cdot(1 - \mathrm{p}). \tag{3}$$

On the other hand, $< BS_{CL} >$ can also be directly estimated based on forecast samples, according to the following formula:

$$< BS_{CL_{estimate}} >= \sum_{k=1}^{2} \sum_{j=1}^{N} \frac{\left(C_p - O_{kj}\right)^2}{N}. \quad (4)$$

where C_p is the climatological TC occurrence probability, either p or $1 - p$.

In Formula (3), $< BS_{CL_theory} >$ is computed from p, which is the climatological DTCP calculated from observation for the years when the model provided a forecast. It should be noted that in Formula (2), there is a summation over TC categories k, with and without TC activity, but in another formulation [23], there is no summation over TC category k. There is a difference in $< BS >$ and $< BS_{CL} >$ by a factor of 2 with and without summation over the category k. It should be noted that using summation over the category k is consistent with the derivation of the correction term D for ranked probability skill score [34].

In Formula (1), D is given as follows:

$$D = \frac{p \cdot (1 - p)}{M} \quad (5)$$

where M is the number of ensemble members. D can correct negative bias that occurs when M is small, and its effect decreases when M becomes large [34].

The meaning of the DBSS is to compare the Brier score from model forecast with the Brier score of reference climate forecast if D is ignored for convenience. When the forecast from a model is better than the reference climate forecast, the DBSS is greater than 0. The model forecast can provide further valuable information compared with a straightforward reference climate forecast. We removed the grid points with p below a certain threshold when computing the basin-averaged DBSS. Some grid points were removed before computing the basin-averaged DBSS to avoid exceptionally negative values at these points. Both the two terms in the denominator of Equation (1) have climatological DTCP p. If p is zero, the value of DBSS at that point is not defined. If p is small, an exceptionally large negative DBSS may be generated.

3. Results

Probabilistic forecasting variables have been widely applied for evaluation of S2S TC prediction [14,16,18,22]. However, multiple factors in the computation can influence the evaluation results, such as the TC influence radius and forecast time window in DTCP. The lack of standardized calculation methods for these probabilistic forecasts poses significant challenges for the operational application of S2S TC prediction.

Previous studies have commonly employed the Brier skill score (BSS) and debiased Brier skill score (DBSS) to assess probabilistic forecasting skill. Nevertheless, the computational approaches of the terms in the above equation are different. This section systematically focuses on DTCP and DBSS to examine the factors affecting evaluation of the models' forecasting skill, including different evaluation regions, varying forecast time windows, different TC influence radii, alternative methods for computing the Brier score of reference climate forecasts, and the application of the correction term D. Through comprehensive experiments, we analyzed the influence of these factors on DTCP and DBSS. In the initial framework [22], we used a definition of DTCP with a 1-day forecast time window and 500 km TC influence radius. Based on the four models' forecast outputs (mainly the ECMWF model) in Table 1, we investigated the effect of removing the grid points where climatological p was below certain criteria, such as 0.01, and when a different forecast time

window and TC influence radius were used. The influence of different methods to compute $< BS_{CL} >$ and whether to include correction term D are also analyzed.

3.1. Test of Evaluation Region for Basin Averaged DBSS

A mask was implemented in the calculation of basin-averaged DBSS to exclude grid points with observed climatological DTCP p below specified thresholds. This prevented the potential influence of exceptionally large negative DBSS values due to extremely low p values at certain grid points.

The selection of different thresholds inherently defines varying evaluation regions; higher thresholds correspond to oceanic regions with more TC activity. To systematically evaluate the sensitivity of DBSS to different evaluation regions, we examined four threshold values: 0 (Figure 2a), 0.01 (Figure 2b), 0.03 (Figure 2c), and 0.05 (Figure 2d). The spatial coverage showed progressive contraction with increasing thresholds. While the zero threshold excluded only limited equatorial areas (Figure 2a), the biggest threshold of 0.05 confined the evaluation region to the most active TC area including the East China Sea, South China Sea, and the ocean east of the Philippines and Japan (Figure 2d).

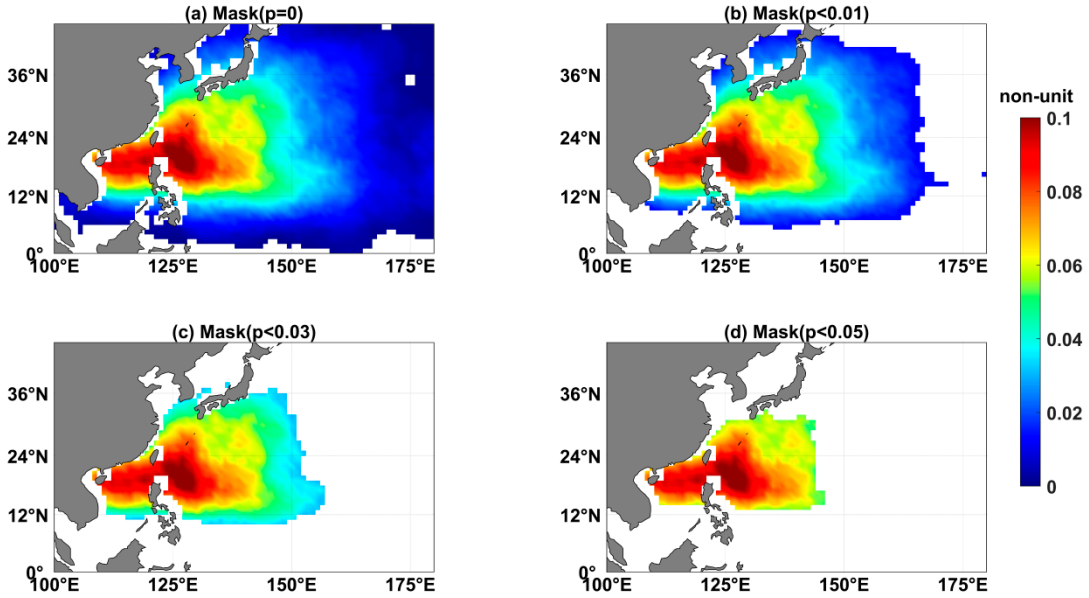


Figure 2. Observed climatological DTCP (p) for WNP under different evaluation region thresholds: (**a**) $p \geq 0$ (no masking), (**b**) $p \geq 0.01$, (**c**) $p \geq 0.03$, and (**d**) $p \geq 0.05$. Grid points with p below the threshold are masked (white). Data derived from IBTrACS (2015–2023).

Figure 3 demonstrates that the mean DBSS from the ECMWF model showed some variation for the different regions under evaluation. During the first 7 days, the mean DBSS gradually increased with rising threshold values. After the first week, the mean DBSS values for thresholds of 0.01, 0.03, and 0.05 exhibited negligible differences, while the zero-threshold DBSS displayed higher values with more pronounced fluctuations. This phenomenon may be attributed to excessively negative DBSS values with small p values, which ultimately affected the basin average. The elevated zero-threshold DBSS after the first week could potentially have resulted from inactive TC regions contributing more positive DBSS. The choice of evaluation region influenced the model's DBSS, but it produced no substantial variations except for extreme thresholds. The mean DBSS across different thresholds generally fell below zero around day 10.

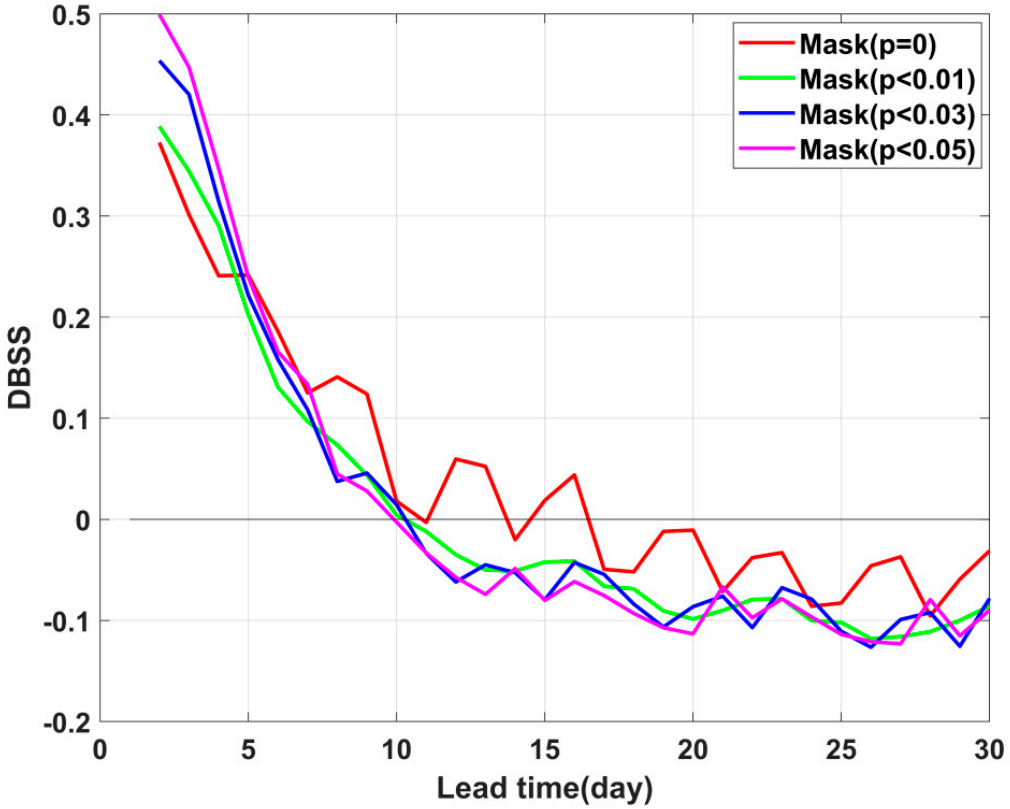


Figure 3. Regional mean DBSS for ECMWF model with different evaluation regions. The red, green, blue, and magenta lines represent mean DBSS with thresholds of 0, 0.01, 0.03, and 0.05, respectively (TC influence radius: 500 km; forecast time window: 1 day; Brier score for reference climate forecast is theoretical with correction effect of term D).

3.2. Test of Forecast Time Window

The choice of forecast time window can also influence the DBSS, and how to choose the time window is an interesting question. An excessively long time window is unfavorable for forecasting individual TCs [14,17], while a short time window can easily lead to inaccurate forecasting [15]. This experiment was conducted with forecast time windows of 1, 3, and 5 days, respectively. For the multi-day sliding forecast time window, the algorithm expands bidirectionally (both forward and backward in time). When processing the observational data, all TC positions within the extended 3-day (or 5-day) period were aggregated to compute the DTCP, though the word "daily" is not accurate when the forecast time window is 3 or 5 days. As illustrated in Figure 4, extending the forecast time window led to an increase in the observed climatological DTCP, as it included TC occurrences across multiple days. To compute the model's DTCP, the DTCP is first computed separately for each day, and the maximum DTCP across different days at the same grid point was selected as the final forecast DTCP.

It should be noted that when extending the forecast time window to 3 days, the first day (Day 1) and the last day (Day 30) cannot be expanded further backward or forward, respectively. Similarly, for a 5-day forecast time window, Days 1–2 and Days 29–30 cannot be extended beyond the available time range. Consequently, when applying a 3-day (or 5-day) forecast time window, the mean DBSS values for Day 1 and Day 30 (or Days 1–2 and Days 29–30) were excluded from the comparison.

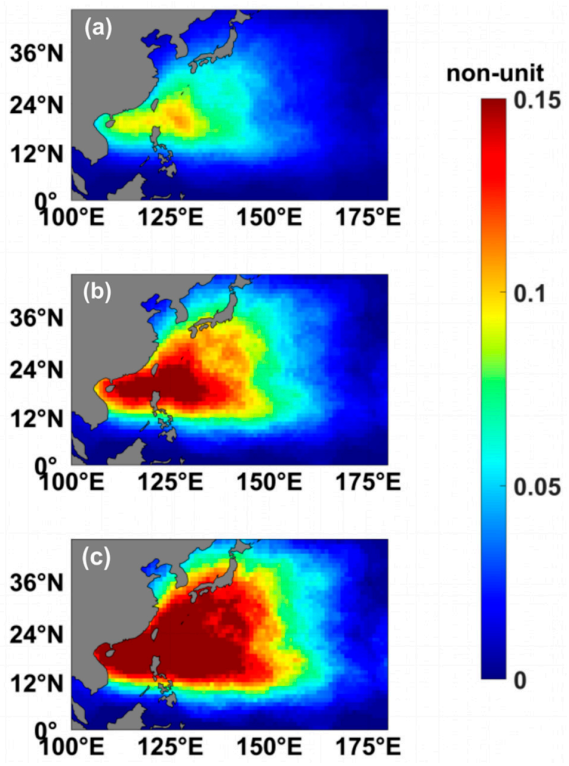


Figure 4. Spatial distribution of observed climatological DTCP using different forecast time windows: (**a**) 1 day, (**b**) 3 days, and (**c**) 5 days.

As shown in Figure 5, the mean DBSS of the ECMWF model increased with the extension of the forecast time window. The reason may have been that a longer forecast time window allowed for greater permissible errors in the model predictions, thereby leading to a higher mean DBSS. The increase of mean DBSS with a longer forecast time window is limited since the DBSS is the ratio of model forecast error and forecast error from a reference climate forecast (if the D term is ignored).

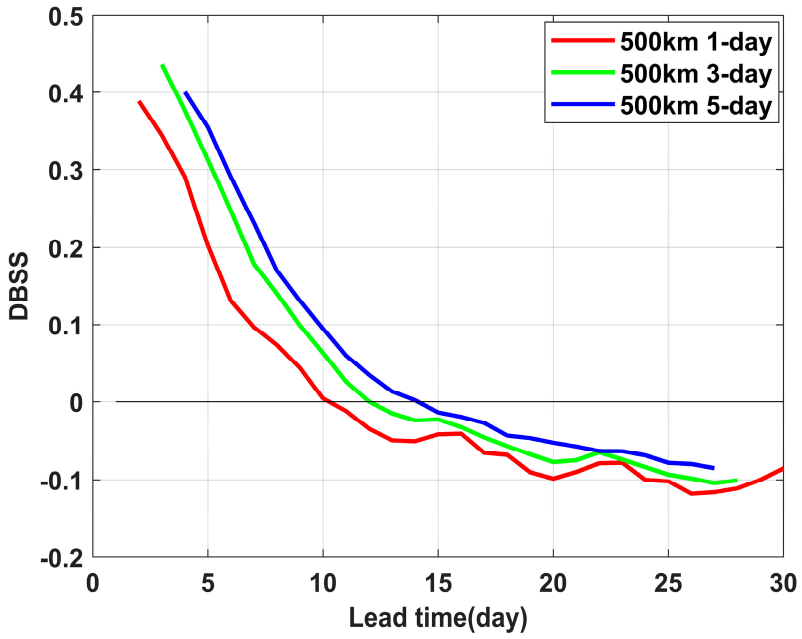


Figure 5. Mean DBSS for ECMWF model forecasts with 1-day, 3-day, and 5-day forecast time windows (TC influence radius: 500 km; evaluation region: $p \geq 0.01$; Brier score for reference climate forecast is theoretical with correction effect of term D).

When the forecast time window was extended to 3 days, the mean DBSS dropped below 0 by day 12. Similarly, with a 5-day forecast time window, the mean DBSS fell below 0 by day 14. This result demonstrates that the DBSS did not change dramatically and exhibited robustness in response to variations in the forecast time window.

3.3. Test of TC Influence Radius

Similar to the selection of time windows, choosing the appropriate TC influence radius is crucial for evaluation. When the spatial range of forecast indicators are excessively broad, it is impossible to accurately locate TCs (such as ACE [18] or 'box' evaluation [19–21]), and overly specific forecast variables are also not suitable for the S2S scale (such as TC tracks [15]). In the definition of DTCP, the area within the TC influence radius is considered to be affected by the TC. The size of the TC influence radius can influence the calculation of DTCP. We examined the impact on mean DBSS by setting the TC influence radius to 300 km and 700 km, and compared the results with a TC influence radius of 500 km. As illustrated in Figure 6, the increase in the TC influence radius led to a larger area being affected by TC, consequently resulting in higher observed climatological DTCP.

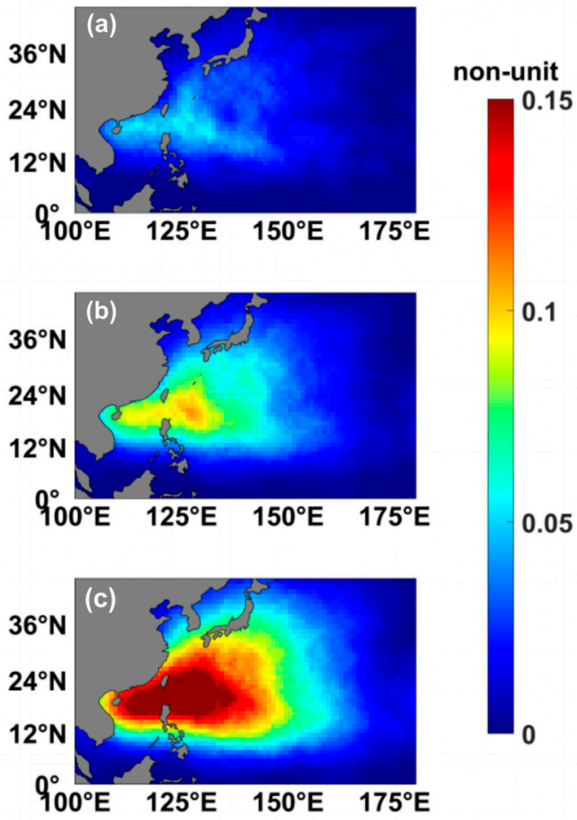


Figure 6. Spatial distribution of observed climatological DTCP using different TC influence radii: (**a**) 300 km, (**b**) 500 km, and (**c**) 700 km.

Figure 7 shows the mean DBSS of the ECMWF model using different TC influence radius. The smallest DBSS occured within the 300 km radius, dropping below zero by day 9. The 700 km radius yielded a higher DBSS that fell below 0 by day 12. For the intermediate 500 km radius, the mean DBSS became negative on day 10. Consistent with the findings from the influence of forecast time window, the differences in mean DBSS among various TC influence radii remained relatively small. This behavior was as expected, because a larger TC influence radius accommodates greater positional errors in model forecasts. However, the expanded radius also introduces more false alarms as it covers larger areas where the TC may not actually occur. Consequently, while increasing the

TC influence radius does moderately enhance the value of the DBSS, the enhancement is limited, demonstrating that the DBSS exhibits low sensitivity to variations in the TC influence radius.

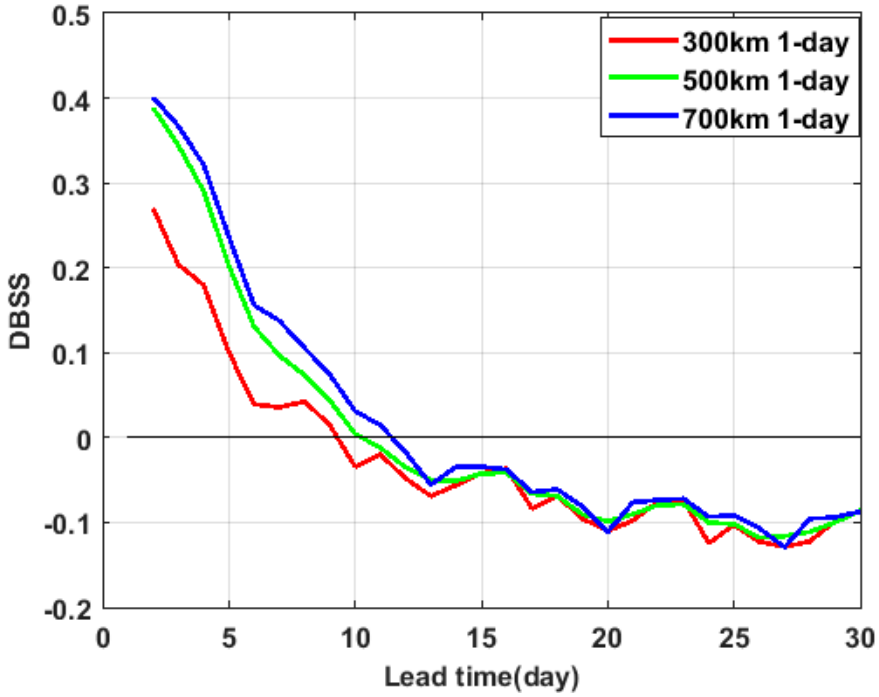


Figure 7. Mean DBSS for ECMWF model forecasts with TC influence radii of 300 km, 500 km, and 700 km (forecast time window: 1 day; evaluation region: $p \geq 0.01$; Brier score for reference climate forecast is theoretical with correction effect of term D).

3.4. Test of Brier Score for Reference Climate Forecast

Previous research conducted an analysis of the correction effect of term D, demonstrating that incorporating this term can effectively mitigate the negative bias in models with limited ensemble members [23,24]. However, in the application of Formula (1), $< BS >$ and $< BS_{CL} >$ were computed without distinguishing between event categories (i.e., TC occurrence and non-occurrence), which may have led to an overestimation of term D's contribution to the DBSS calculation. In Formulas (3) and (4), $< BS_{CL} >$ can be derived either through theoretical calculation or estimated directly from model forecast samples, potentially yielding different DBSS values depending on the computation method. To investigate this sensitivity, we designed four experimental calculations: (1) using $< BS_{CL_theory} >$ with D; (2) using $< BS_{CL_theory} >$ without D; (3) using $< BS_{CL_estimate} >$ with D; and (4) using $< BS_{CL_estimate} >$ without D. This comparison allowed evaluation of DBSS's sensitivity to different calculation methods of $< BS_{CL} >$ and the inclusion (or exclusion) of term D.

The experimental results presented in Figure 8 demonstrate that DBSSs computed using different $< BS_{CL} >$ methods remained nearly identical during the first 5 forecast days, but they began to diverge significantly after day 10. Notably, the mean DBSS based on $< BS_{CL_estimate} >$ approached zero after day 15, while that calculated using $< BS_{CL_theory} >$ dropped below zero after day 10.

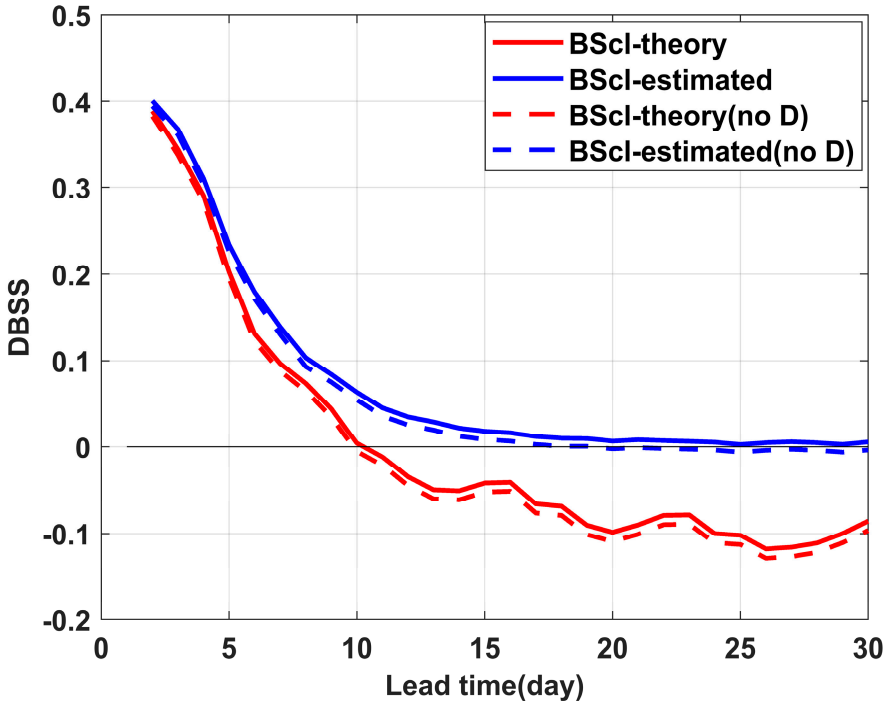


Figure 8. The ECMWF model's mean DBSS under four experimental configurations: (1) theoretical Brier score for reference climate forecast with D term (solid red), (2) theoretical Brier score without D term (dashed red), (3) estimated Brier score for reference climate forecast with D term (solid blue), and (4) estimated Brier score without D term (dashed blue). (TC influence radius: 500 km; forecast time window: 1 day; evaluation region: $p \geq 0.01$).

A distinct smoothing effect was observed in the mean DBSS curve when $< BS_{CL_estimate} >$ was employed, compared with the $< BS_{CL_theory} >$ results. This smoothing is likely to be because $< BS_{CL_estimate} >$ was derived from forecast samples where the sample size was inherently linked to the frequency of forecasts. This relationship between sample size and forecast frequency may have introduced fluctuations in the $< BS_{CL} >$ calculations. However, when both $< BS_{CL_estimate} >$ and $< BS >$ were computed based on model forecast frequencies, these fluctuations tended to cancel each other out. Further investigation into the effects of forecast frequency and sample size will be presented in a subsequent case study. When different models are involved in the evaluation, using an estimated Brier score for reference climate forecast is not appropriate, since the practice introduces a different Brier score for the reference climate forecast.

Figure 8 illustrates that including the D-term caused relatively minor variations in the mean DBSS. As indicated in Table 1, this limited impact can be attributed to the ECMWF model's substantial ensemble size of 51 members, which resulted in a considerably small magnitude of the D correction term computed from Formula (5). To further examine the impact of D in models with smaller ensemble sizes, we conducted experiments using the KMA model, which operates with only four ensemble members. Figure 9 demonstrates a pronounced enhancement in mean DBSS following the implementation of the D correction term in the KMA model. This analysis clearly indicates that the D term plays a crucial role in models with small numbers of ensemble members, effectively correcting negative skill score biases, while its contribution becomes less significant in models with large ensemble sizes such as the ECMWF model.

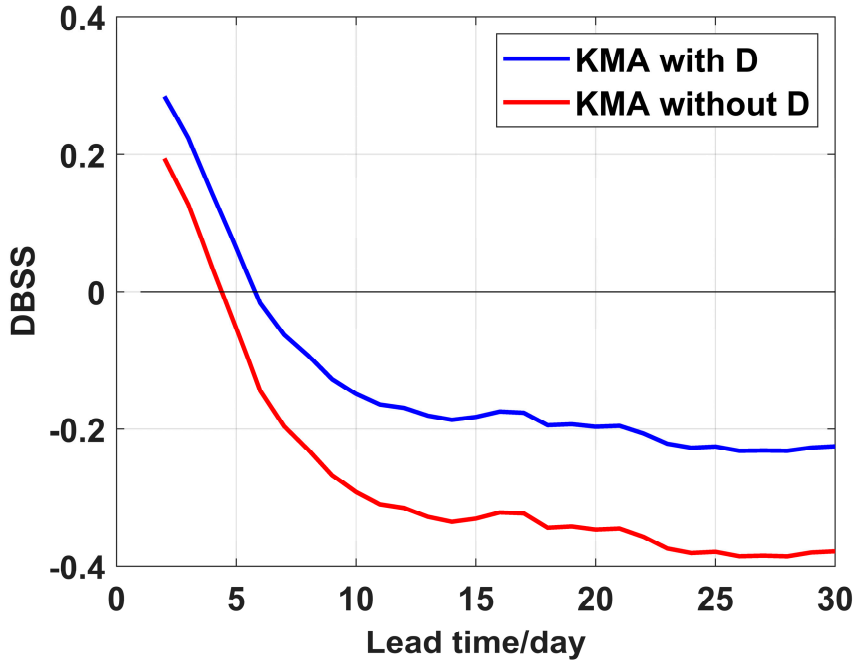


Figure 9. Comparison of the KMA model's mean DBSS with and without the correction term (TC influence radius: 500 km; forecast time window: 1 day; evaluation region: $p \geq 0.01$; Brier score for reference climate forecast is theoretical).

Previous studies [19,23,24,26] emphasized the fundamental importance of ensemble size in evaluating model forecast skill. The inclusion of the D correction term in the assessment of forecasting skill is particularly valuable, especially for operational models with few ensemble members.

3.5. Test of Forecast Frequency

In Section 3.4, we mentioned the impact of forecast sample size on DBSS. To further investigate its influence, we conducted experiments by modifying the frequency of forecasts to alter the sample size. We selected three daily forecast models (KMA, NCEP, and UKMO) from the S2S real-time database and artificially changed their forecast frequency from daily to weekly, thereby creating three hypothetical models.

Compared with the mean DBSS of the original daily forecast models, their weekly forecast counterparts revealed pronounced periodic oscillations in mean DBSS (Figure 10). This phenomenon arose from reduced forecast sample sizes leading to increased uncertainty. These results clearly demonstrate that variations in the sample size of forecasts significantly affect DBSS. Consequently, when evaluating models' forecasting skill, careful consideration of sample sizes is essential to ensure reliable assessment.

It is worth noting that although our experiments specifically focused on the DTCP and DBSS framework, the findings possess broader applicability to other TC probability forecasting evaluation frameworks. For instance, the 'Box' approach employed by earlier researchers [18–21] shares conceptual similarities with our TC influence radius definition, while the 3-day sliding window [16] closely resembles our forecast time window methodology. The quantitative findings of our research might be different for other models and other TC basins, as indicated by a research focusing on the predictability range of ensemble forecasts [35]. This study provides some guidance for evaluating S2S TC forecasts and establishes a foundation for future global assessments of S2S TC forecast skill.

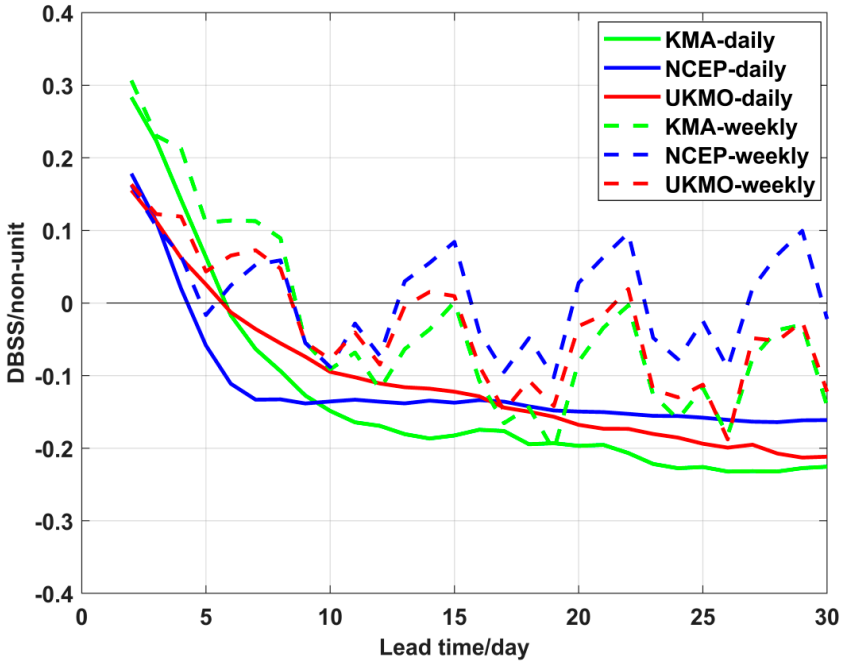


Figure 10. Comparison of mean DBSS for the three operational models (KMA, NCEP, and UKMO) under different forecast frequencies: daily (solid line) and weekly (dashed line) (TC influence radius: 500 km; forecast time window: 1 day; evaluation region: $p \geq 0.01$; Brier score for reference climate forecast is theoretical).

4. Discussion and Conclusions

Currently, there is no universally accepted standard for evaluating S2S TC forecasts. Several probabilistic variables have been proposed, including daily tropical cyclone probability (DTCP) and the debiased Brier skill score (DBSS) framework. In this study, we systematically examined the influence of several important factors using DTCP as the forecast variable and DBSS as the forecasting skill score metric. Through separate sensitivity tests, we demonstrated that while modifications to the evaluation region, forecast time window, and TC influence radius altered spatially averaged DBSS values, their overall impacts remained consistent and relatively minor. However, the use of estimated Brier score for reference climate forecasts introduces notable variability due to the limited sample sizes—a crucial consideration when evaluating models' forecasting skill.

Based on the results presented in Section 3, we recommend the following for multi-model evaluation on S2S scale. (1) The theoretical Brier score should be used for reference climate forecasts to ensure fair comparisons between different models. (2) The correction term D should be applied to mitigate negative biases while evaluating models with small ensemble sizes. (3) While expanding either the forecast time window or TC influence radius may improve forecasting skill scores, these adjustments simultaneously reduce spatial precision in TC localization. Therefore, these key parameters should be carefully selected based on specific operational requirements. These recommendations provide standardized evaluation criteria while maintaining the flexibility necessary for diverse operational applications.

The DTCP and DBSS framework evaluated here transcends the WNP focus of this study, offering a scalable template for evaluation of global S2S TC forecasting. By analyzing factors such as ensemble size, forecast frequency, and reference climatology, this work bridges gaps between research and operational implementation. Moreover, our results underscore the potential of S2S forecasts to enhance early warning systems and adaptive management strategies; for example, adjusting the forecast time window and TC influence

radius introduces greater flexibility into predictions. We also prove their robustness for operational applications.

As climate change amplifies TC risks, the ability to predict TC activity weeks in advance becomes indispensable for safeguarding lives and economies. Future research could extend this framework to other basins and explore couplings with oceanic and atmospheric drivers (e.g., MJO, ENSO) to further enhance predictions. This study refines the technical evaluation of S2S TC forecasts and by clarifying the impacts of evaluation choices, we pave the way for more credible, actionable, and globally relevant S2S predictions, ultimately contributing to resilience in the face of growing threats from tropical cyclones.

Author Contributions: Resources, X.W.; Data curation, B.Z., M.Y. and F.V.; Writing—original draft, Y.L. (Yuanben Li); Writing—review & editing, X.W., Y.L. (Yimin Liu) and F.V.; Funding acquisition, B.Z. and M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: X.W. and Y.L. (Yuanben Li) are supported by the National Key R&D Program of China (Grant Number 2024YFC2815705) and Shanghai Typhoon Research Foundation (TFJJ202205). Y.L. (Yimin Liu) is supported by the Natural Science Foundation of China (Fund Number 42288101). The retracking and archiving of S2S Prediction Project Database was conducted at ECMWF.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Emanuel, K. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **2005**, *436*, 686. [CrossRef] [PubMed]
2. Webster, P.; Holland, G.; Curry, J.; Chang, H. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science* **2005**, *309*, 1844–1846. [CrossRef] [PubMed]
3. Klotzbach, P.; Wood, K.M.; Schreck, C.; Bowen, S.G.; Patricola, C.; Bell, M. Trends in global tropical cyclone activity: 1990–2021. *Geophys. Res. Lett.* **2021**, *49*, e2021GL095774. [CrossRef]
4. Emanuel, K.; Chonabayashi, S.; Bakkensen, L.; Mendelsohn, R. The impact of climate change on global tropical cyclone damage. *Nat. Clim. Change* **2012**, *2*, 205–209.
5. Bauer, P.; Thorpe, A.; Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **2015**, *525*, 47–55. [CrossRef]
6. Robertson, A.W.; Vitart, F.; Camargo, S.J. Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *J. Geophys. Res. Atmos.* **2020**, *125*, e2018JD029375. [CrossRef]
7. Vitart, F.; Robertson, A.W. *Sub-Seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 1–15.
8. Camargo, S.J.; Camp, J.; Elsberry, R.L.; Tippett, M.K.; Wehner, M.F.; Wang, X. Tropical cyclone prediction on subseasonal time-scales. *Trop. Cyclone Res. Rev.* **2019**, *3*, 150–165. [CrossRef]
9. Vitart, F.; Ardilouze, C.; Bonet, A.; Brookshaw, A.; Chen, M.; Codorean, C.; Déqué, M.; Ferranti, L.; Fucile, E.; Fuentes, M.; et al. The subseasonal to seasonal (S2S) prediction project database. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 163–173. [CrossRef]
10. Vitart, F.; Robertson, A.W. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Clim. Atmos. Sci.* **2018**, *1*, 3. [CrossRef]
11. World Meteorological Organization (WMO). *WWRP/WCRP Sub-Seasonal to Seasonal Prediction Project (S2S) Phasse 1 Final Report*; Tropical Cyclone Programme Report; WMO: Geneva, Switzerland, 2018.
12. World Meteorological Organization (WMO). *WWRP/WCRP Sub-Seasonal to Seasonal Prediction Project (S2S) Phase 2 Final Report*; Tropical Cyclone Programme Report; WMO: Geneva, Switzerland, 2023.
13. Vitart, F.; Leroy, A.; Wheeler, M.C. A comparison of dynamical and statistical predictions of weekly tropical cyclone activity in the Southern Hemisphere. *Mon. Weather Rev.* **2010**, *138*, 3671–3682. [CrossRef]
14. Vitart, F.; Prates, F.; Bonet, A.; Doblas–Reyes, F.J.; Hagedorn, R.; Palmer, T.N. New tropical cyclone products on the web. *ECMWF Newsl. Meteorol.* **2011**, *130*, 17–23.
15. Tsai, H.C.; Elsberry, R.L.; Jordan, M.S.; Shieh, C.J. Objective verifications and false alarm analyses of western North Pacific tropical cyclone event forecasts by the ECMWF 32-day ensemble. *Asia-Pac. J. Atmos. Sci.* **2013**, *49*, 409–420. [CrossRef]
16. Yamaguchi, M.; Vitart, F.; Lang, S.T.K.; Weigel, A.P.; Liniger, M.A. Global distribution on the skill of tropical cyclone activity forecasts from short- to medium-range time scales. *Weather Forecast.* **2015**, *30*, 1695–1709. [CrossRef]

17. Gao, K.; Chen, J.; Harris, L.; Sun, Y.; Lin, S. Skillful prediction of monthly major hurricane activity in the North Atlantic with two-way nesting. *Geophys. Res. Lett.* **2019**, *46*, 9222–9230. [CrossRef]
18. Lee, C.-Y.; Camargo, S.J.; Vitart, F.; Tippett, M.K.; Wehner, M.F. Subseasonal predictions of tropical cyclone occurrence and ACE in the S2S dataset. *Weather Forecast.* **2020**, *35*, 921–938. [CrossRef]
19. Camp, J.; Gregory, P.; Marshall, A.G.; Wehner, M.F.; Tippett, M.K.; Vitart, F. Skilful multiweek predictions of tropical cyclone frequency in the Northern Hemisphere using ACCESS-S2. *Q. J. R. Meteorol. Soc.* **2024**, *150*, 2848–2868. [CrossRef]
20. Gregory, P.; Camp, J.; Bigelow, K.; Brown, A. Subseasonal predictability of the 2017–2018 Southern Hemisphere tropical cyclone season. *Atmos. Sci. Lett.* **2019**, *20*, e886. [CrossRef]
21. Gregory, P.; Vitart, F.; Brown, A.; Camp, J.; Wehner, M.F.; Tippett, M.K. Subseasonal forecasts of tropical cyclones in the southern hemisphere using a dynamical multimodel ensemble. *Weather Forecast.* **2020**, *35*, 1817–1829. [CrossRef]
22. Wang, X.; Waliser, D.E.; Jiang, X.; Tippett, M.K.; Wehner, M.F.; Vitart, F. Evaluating western North Pacific tropical cyclone forecast in the subseasonal to seasonal prediction project database. *Front. Earth Sci.* **2023**, *10*, 1064960. [CrossRef]
23. Switanek, M.B.; Hamill, T.M.; Long, L.N.; Tippett, M.K.; Wehner, M.F.; Vitart, F. Predicting Subseasonal Tropical Cyclone Activity Using NOAA and ECMWF Reforecasts. *Weather Forecast.* **2023**, *38*, 357–370. [CrossRef]
24. García-Franco, J.L.; Lee, C.; Camargo, S.J.; Tippett, M.K.; Wehner, M.F.; Vitart, F. Tropical Cyclones in the GEOS-S2S-2 Subseasonal Forecasts. *Weather Forecast.* **2024**, *39*, 1297–1318. [CrossRef]
25. Pegion, K.; Kirtman, B.P.; Becker, E.; Collins, D.C.; LaJoie, E.; Burgman, R.; Bell, R.; DelSole, T.; Min, D.; Zhu, Y.; et al. The subseasonal experiment(SubX): A multimodel subseasonal prediction experiment. *Bull. Am. Meteorol. Soc.* **2019**, *100*, 2043–2060. [CrossRef]
26. Schreck, C.J., III; Vitart, F.; Camargo, S.J.; Camp, J.; Darlow, J.; Elsberry, R.; Gottschalck, J.; Gregory, P.; Hansen, K.; Jackson, J.; et al. Advances in tropical cyclone prediction on subseasonal time scales during 2019–2022. *Trop. Cyclone Res. Rev.* **2023**, *12*, 136–150. [CrossRef]
27. Chen, L.S.; Duan, Y.H.; Song, L.L. *Typhoon Forecast and Disaster*; China Meteorological Press: Beijing, China, 2012; pp. 194–201.
28. Chan, J.C.L. Interannual and interdecadal variations of tropical cyclone activity over the western North Pacific. *Meteorol. Atmos. Phys.* **2005**, *89*, 143–152. [CrossRef]
29. Maue, R.N. Recent historically low global tropical cyclone activity. *Geophys. Res. Lett.* **2011**, *38*, L14803. [CrossRef]
30. Knapp, K.R.; Kruk, M.C.; Levinson, D.H.; Diamond, H.J.; Neumann, C.J. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data. *Bull. Am. Meteorol. Soc.* **2010**, *91*, 363–376. [CrossRef]
31. Domeisen, D.I.V.; Butler, A.H.; Charlton-Perez, A.J.; Ayarzagüena, B.; Baldwin, M.P.; Dunn-Sigouin, E.; Furtado, J.C.; Garfinkel, C.I.; Hitchcock, P.; Karpechko, A.Y.; et al. The role of the stratosphere in subseasonal to seasonal prediction Part II: Predictability arising from stratosphere–troposphere coupling. *J. Geophys. Res. Atmos.* **2019**, *125*, e2019JD030923.
32. Vitart, F.; Stockdale, T.N. Seasonal forecasting of tropical storms using coupled GCM integrations. *Mon. Weather Rev.* **2001**, *129*, 2521–2537. [CrossRef]
33. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **1950**, *78*, 1–3. [CrossRef]
34. Weigel, A.P.; Liniger, M.A.; Appenzeller, C. The discrete Brier and ranked probability skill scores. *Mon. Weather Rev.* **2007**, *135*, 118–124. [CrossRef]
35. Buizza, R.; Leutbecher, M. The forecast skill horizon. *Q. J. R. Meteorol. Soc.* **2015**, *141*, 3366–3382. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.