JAMES | Journal of Advances in Modeling Earth Systems®

AGU ADVANCING EARTH AND SPACE SCIENCE

RESEARCH ARTICLE
10.1029/2022MS003381

Key Points:

- A realistic context of model uncertainty is introduced to evaluate the physical parameterization schemes for tropical cyclone (TC) forecasts
- The performance of scheme combination is more comprehensive in the uncertainty context than in the deterministic context
- The well-performed combination is selected for both TC track and intensity forecasts in the model uncertainty context

Correspondence to:
Z.-M. Tan,
zmtan@nju.edu.cn

Citation:
Wang, X., & Tan, Z.-M. (2023). On the combination of physical parameterization schemes for tropical cyclone track and intensity forecasts in the context of uncertainty. *Journal of Advances in Modeling Earth Systems*, *15*, e2022MS003381. https://doi.org/10.1029/2022MS003381

Received 26 AUG 2022
Accepted 21 MAR 2023

Author Contributions:
Conceptualization: Xuan Wang, Zhe-Min Tan
Data curation: Xuan Wang
Formal analysis: Xuan Wang, Zhe-Min Tan
Funding acquisition: Zhe-Min Tan
Investigation: Xuan Wang
Methodology: Xuan Wang, Zhe-Min Tan
Project Administration: Zhe-Min Tan
Resources: Zhe-Min Tan
Software: Xuan Wang
Validation: Xuan Wang
Visualization: Xuan Wang
Writing – original draft: Xuan Wang

© 2023 The Authors. Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

On the Combination of Physical Parameterization Schemes for Tropical Cyclone Track and Intensity Forecasts in the Context of Uncertainty

Xuan Wang[1] and **Zhe-Min Tan**[1]

[1]School of Atmospheric Sciences, Nanjing University, Nanjing, China

Abstract The selection of physical parameterization schemes for tropical cyclone (TC) forecasts has required a substantial amount of effort. In general, the evaluation of physical parameterization schemes and their combined performance was based solely on the deterministic forecast, which had inherent limitations in representing the overall performance of physical parameterization schemes due to the model uncertainty. This study introduces an uncertainty-informed framework of evaluating and selecting the combination of physical parameterization schemes for TC forecasts, based on the ensemble forecast that could include the model uncertainty roles. The performance ranking of the scheme combination based on the ensemble mean error is found to be distinct from that based on the deterministic forecast error. Moreover, differences in both ensemble mean errors and ensemble spreads for various scheme combinations highlight the importance of considering two metrics concurrently, that is, via the quality of the forecast distribution as a whole, to assess the forecast performance. Consequently, the ensemble Continuous Ranked Probability Score (eCRPS) is used to quantify the performance of the scheme combinations, and it is demonstrated that the performance is more comprehensive than that in the deterministic context. Finally, the well-performed scheme combination for the forecasts of six intense TCs is chosen from the evaluated schemes in the context of model uncertainty, based on the overall quality of TC track and intensity forecast distributions.

Plain Language Summary In order to improve the accuracy of TC track and intensity forecasts, it is crucial to select the appropriate physical parameterization schemes for the forecasts. In general, the performance of the physical scheme was quantified by comparing the observation with a single forecast value. Taking into account the non-negligible uncertainty sources in the forecast that contribute to the final errors, the forecast value with the chosen physical scheme will be a distribution rather than a point. Through a type of ensemble perturbation, this study restores some model uncertainty information and evaluates the pre-selected combinations of physical parameterization schemes in the ensemble forecasts. The performances of the scheme combinations based on the ensemble mean error differ from those based on the single forecast error. In addition to the differences in ensemble mean errors, there are also differences in the forecast distributions of various scheme combinations. Thus, in the context of uncertainty, the performances of the scheme combinations are quantified by the overall quality of the forecast distribution and shown to be more comprehensive than in the deterministic context. Finally, the well-performed scheme combination for both TC track and intensity forecasts of six intense TCs is quantitatively selected from the evaluated schemes.

1. Introduction

The heavy devastations caused by tropical cyclones (TCs) on life and property make the improvement of TC forecasts crucial for public safety. However, the accurate prediction of TC track and intensity still face great challenges. There are numerous uncertainty sources in TC forecasts, such as the initial error caused by the inaccuracy of input analysis data, the model error either from the structural bias in the numerical scheme and interpolation, or from the parameterization errors due to deficiencies in various physical parameterization schemes. In order to reduce physical parameterization errors and improve the representation of TC conditions, much efforts have been made in the development and assessment of physical parameterization schemes, especially for those general physics such as cumulus convection, microphysics and boundary layer processes. For example, a new convection trigger in the Kain-Fritsch (KF) cumulus scheme was proposed for improving TC forecasts with weak synoptic forcing (Ma & Tan, 2009). Shi and Wang (2022) showed that the multiscale KF scheme performed better than those conventional cumulus schemes in TC precipitation forecasts owing to the scale-awareness and

Writing – review & editing: Xuan Wang, Zhe-Min Tan

parameterized cloud–radiation feedback. In the evaluation of microphysics schemes, the Thompson aerosol-aware scheme was denoted of superiority in hurricane precipitation forecasts and was further improved based on the polarimetric radar observations (Brown et al., 2016, 2017). Several boundary layer parameterization schemes have been improved for diffusion and mixing within TC boundary layer based either on the aircraft observations (Gopalakrishnan et al., 2021; Zhang et al., 2015) or the large eddy simulations (Chen, 2022; Chen et al., 2022; Li & Pu, 2021). Besides the deficiency within single scheme, the interaction of different physical parameterization schemes also contributes significantly to the forecast error. With a number of schemes for describing each physical processes in the numerical model, how to select a suitable combination of physical parameterization schemes for TC forecasts becomes an important question.

There are many works on the selection of scheme combinations for TC track and intensity forecasts, mostly focusing on limited number of physical parameterization schemes (e.g., Islam et al., 2015; Osuri et al., 2012; Raju et al., 2011; Srinivas et al., 2013). Di et al. (2019) employed a systematic combinatorial optimization method to efficiently cover all available schemes in the model, and determined the well-performed combinations for TC track, central sea level pressure, and 10 m maximum surface wind, respectively. However, the evaluations in the above studies were all carried out in the deterministic forecast, that is, the single-run forecast with a specific initial field and a specific model set, while changing the physical parameterization schemes to be tested. And the forecast performance was verified by the deterministic forecast error, that is, the difference between a single forecast value and the corresponding observation data.

It is well established that even very small perturbations in the initial conditions or the prediction model can develop into a large range of possibilities after a few days of integration (Lorenz, 1963). So the nonlinear interaction of all kinds of errors in the forecast will eventually lead to the forecast variables fluctuating in a distribution rather than being a single value. Also the forecasts with different physical parameterization schemes may produce distinct forecast distributions. The deterministic forecast error is only a sample of all possibilities, thus has inherent limitations to represent the overall performance of the combination of physical parameterization schemes. In order to verify the overall performance of various scheme combinations, the evaluation should be carried out in a realistic uncertainty context, which accounts for the lost forecast possibilities.

A practical way to reflect the forecast uncertainty is to employ ensemble perturbations, which will develop into a forecast probability distribution over time. Considering it is impossible to account for all kinds of uncertainty in TC forecasts, this study mainly focuses on the model uncertainty, that is, the nonlinear interactions of model errors, thus utilizes a stochastic kinetic-energy backscatter (SKEB) scheme to create the ensembles related to the model errors. Originating from the concept that the model error can be manifested in the loss of subgrid-scale energy, SKEB accounts for the uncertainty of the energy cascade from the subgrid-scale processes to the resolved flow (Shutts, 2005). The SKEB generates stochastic, spatial and temporal correlated perturbations, and adds the perturbations to the tendency equations of the horizontal wind and the potential temperature at every time step. Through this process, a sprinkle of energy is injected back to the resolved flow as the wind and temperature forcing. Thus the SKEB perturbations address the model errors due to the unrepresented subgrid-scale processes in a very natural way. The SKEB was successfully implemented in the advanced Weather Research and Forecast (WRF) model by Berner et al. (2011), and performed useful in improving the probabilistic weather forecasts in the mid-latitudes. Besides, Berner et al. (2011) showed that the ensemble forecast combining multi-physics schemes with SKEB perturbations performed the best, indicating that the model errors can be more sufficiently represented through the addition of SKEB perturbations to the multi-physics ensembles. SKEB has been widely used for TC ensemble forecast studies, and shows improvements in the ensemble dispersion no matter for TC genesis forecasts (Thatcher & Pu, 2014), or for TC track and intensity forecasts (Judt et al., 2016; Li et al., 2019; Melhauser et al., 2017). Considering the physics as well as the good application of SKEB, it is suitable to add SKEB perturbations to various combinations of physical parameterization schemes for evaluating their forecast performances in the model uncertainty context.

The average of the forecast distribution, or the ensemble mean, generally provides a smaller error than most individual members comprising the ensemble (Murphy, 1988; Tracton & Kalnay, 1993). In addition, the width of the forecast distribution, or the ensemble spread is also an important metric of the ensemble that can reflect the uncertainty in the forecast (Whitaker & Louche, 1998). If the ensemble spread is smaller than the ensemble mean error, the ensemble is seen as under-dispersive and the model is over-confident, and vice versa. For TC ensemble forecast verifications, the forecast error is mostly measured by the ensemble mean error, while the reliability of

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1
2018 TC Cases for Evaluation

Number	Name	Category	Simulation period	Days
1	Maria	5	2018-07-05_00~2018-07-11_00	6
2	Cimaron	4	2018-08-19_00~2018-08-24_00	5
3	Jebi	5	2018-08-29_00~2018-09-04_00	6
4	Mangkhut	5	2018-09-09_00~2018-09-17_00	8
5	Kong-rey	5	2018-09-30_00~2018-10-06_00	6
6	Yutu	5	2018-10-23_12~2018-10-30_00	6.5

Note. The category refers to Saffir-Simpson Intensity Scale.

the probability distribution is separately evaluated by the spread–error relationship (e.g., Aemisegger, 2009). However, substituting the ensemble mean error for the deterministic forecast error will still lose the forecast possibilities, because the ensemble mean is also a deterministic forecast. And an objective selection for physical parameterization schemes is hard to achieve since the ensemble mean error and the spread–error relationship may exhibit different performances among the ensemble forecasts with various physical parameterization schemes. Therefore, the performance of the ensemble forecast will be better measured by a metric including the characteristics of the ensemble mean error and the ensemble spread. In other words, the evaluations of scheme combinations in the uncertainty context should be based on the overall quality of the forecast distributions.

In previous studies, the well-performed scheme combination was either selected for the track forecast and the intensity forecast separately (Di et al., 2019), or selected for all variables depending on the subjective judgment (Osuri et al., 2012; Raju et al., 2011; Srinivas et al., 2013). With accurate intensity but large-biased track, the TC forecast is hard to trust. With accurate track but large-biased intensity, the predicted vortex structure may be false. Thus it is expected to obtain a scheme combination performing good in both track and intensity forecasts.

This study proposes an uncertainty-informed framework of evaluating and selecting the combinations of physical parameterization schemes for TC track and intensity forecasts. More specifically, this study attempts to explore the following questions: (a) What is the difference between the performance of the scheme combination in the uncertainty context and that in the deterministic context? (b) How to evaluate the forecast performance of the scheme combination in the context of model uncertainty? (c) How to further quantify the multivariate forecast performance of the scheme combination for TC forecasts? To address these questions, a few combinations of physical parameterization schemes are first selected from a number of schemes, then the deterministic forecasts and the SKEB-perturbed ensemble forecasts with these combinations are carried out. Instead of the deterministic forecast error and the ensemble mean error, the ensemble Continuous Ranked Probability Score (eCRPS, Gneiting & Raftery, 2007) which has the ability of measuring the overall quality of the forecast distribution is employed for the evaluations of ensemble forecasts. Furthermore, the forecast distributions of TC track and intensity forecasts are combined and evaluated through a multivariate extension of eCRPS, thus obtaining the well-performed combination from the evaluated schemes in the context of model uncertainty.

The paper proceeds as follows. Section 2 presents an overview of TC cases and model set-up, along with the experimental design. Section 3 shows the results of the deterministic forecasts and the ensemble forecasts with various physical parameterization combinations. Section 4 displays the performances of various combinations quantified by eCRPS in track ensemble forecasts and intensity ensemble forecasts, respectively. Section 5 exhibits the multivariate performance of the combinations. Section 6 conducts a further validation of the evaluation results. The conclusion and discussion are given in Section 7.

2. Methodology

2.1. TC Cases and Model Set-Up

A total of six TC cases during 2018 over the western North Pacific (WNP) are considered in the evaluation of physical schemes, with intensity and simulation period information listed by Table 1. The TCs were mainly over the ocean and all experienced intensification and weakening during the simulation period, with the lifetime maximum intensity of at least category 4 based on the Saffir-Simpson scale (Simpson & Saffir, 1974). The intense TCs cause the major portion of destructions from all TCs (Pielke et al., 2008), drawing much attention on their forecast errors. And the intensities of intense TCs are usually underestimated while the intensities of weak TCs tend to be overestimated (Huang et al., 2021; Lei et al., 2020). So only intense TCs are taken as evaluated cases in the present study to avoid confusing the properties of all kinds of TCs. Though selecting TCs within 1 year may cause lack of time variability, the climate regimes of the cases are guaranteed to be as similar as possible. The 2018 WNP TC season was known to be highly active and the number of severe TCs was abnormally large (Wu et al., 2020), the unusual tracks and the extreme intensities of which brought great challenge to numerical

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

forecasts (Lei et al., 2020). Therefore, using 2018 WNP major TCs for evaluating physical parameterizations has practical meaning.

The TC forecasts in this study are carried out by the advanced Weather Research and Forecast (WRF) model version 4.0 (Skamarock et al., 2019). The model is configured with two-way interactive three nested domains, of which the grid points are 264 * 190, 679 * 454, 307 * 307 with grid spacing of 27, 9, and 3-km, respectively. There are 45 vertical levels and the model top is at 50 hPa. All TCs have the same domain except for the center location of the innermost vortex-following domain. The initial and boundary conditions are obtained from NCEP GDAS/FNL 0.25° operational global analysis of 6 hr interval. The observed TC positions and intensities come from the International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al., 2010). And the forecast output is compared with the observed data every 3 hr interval for the evaluation.

The WRF model offers a number of physical parameterization schemes and needs to be customized by users. TC track and intensity are considered most sensitive to cumulus convection, microphysics and planetary boundary layer (PBL) physical processes (Di et al., 2019; Raju et al., 2011; Srinivas et al., 2013). Thus in the present study only schemes of the three physical processes are tested for identifying the proper scheme combinations for TC forecasts. The cumulus parameterization schemes considered in this study are: Kain-Fritsch scheme (KF, Kain, 2004), Betts-Miller-Janjic scheme (BMJ, Janjic, 1994), Grell-Freitas ensemble scheme (GF, Grell & Freitas, 2014), Grell-3D scheme (G3, Grell & Dévényi, 2002), Tiedtke scheme (Tiedtke, 1989; Zhang et al., 2011), New Simplified Arakawa-Schubert (NSAS, Kwon & Hong, 2017). It is noticeable that the 3-km resolution is convection-permitting so cumulus schemes are only employed on the 27- and 9-km domains. The microphysics parameterization schemes considered in this study are: Purdue Lin scheme (Lin et al., 1983), WRF Single-Moment 5-class scheme (WSM5, Hong et al., 2004), Ferrier (Eta) scheme (Rogers et al., 2001), WRF Single-Moment 6-class scheme (WSM6, Hong & Lim, 2006), New Thompson scheme (Thompson et al., 2008), WRF Double-Moment 6-class scheme (WDM6, Lim & Hong, 2010). The PBL parameterization schemes considered in this study are: Yonsei University scheme (YSU, Hong et al., 2006), Mellor-Yamada-Janjic scheme (MYJ, Janjic, 1996), Mellor-Yamada Nakanishi and Niino Level 3 scheme (MYNN3, Nakanishi & Niino, 2006), ACM2 scheme (Pleim, 2007), BouLac scheme (Bougeault & Lacarrere, 1989), UW scheme (Bretherton & Park, 2009). Most PBL schemes are matched with the revised MM5 surface layer scheme (Jimenez et al., 2012) except for MYJ, which has to be matched with Janjic Eta surface layer scheme (Janjic, 1996). In addition to the above physical processes, other parameterization configurations of all experiments are kept the same, including the unified Noah land surface model (Tewari et al., 2004), the Rapid Radiative Transfer Model for GCM (RRTMG) longwave and shortwave radiation (Iacono et al., 2008) with the GHG concentration modified to 2018 level, the Donelan and Garratt formulation for air-sea flux parameterization (Donelan et al., 2004; Garratt, 1994), and a 1D ocean model based on Pollard et al. (1973) to turn on the ocean temperature feedback.

2.2. Experimental Design

As mentioned in the introduction, the ensemble forecasts with the testing scheme combinations should be performed in order to evaluate the combinations in the uncertainty context. Considering it is expensive to construct ensembles for every possible combination sampling from all schemes of the three physical processes, a few comparable combinations will be better selected first. So the selection approach similar to the hiring process consisting of the "pre-employment test" and the "interview" is performed. Due to the large number of job applicants, a simple test in advance is useful to shortlist good candidates so that everyone moving to the formal interview is ensured to meet the basic standards for the job (e.g., Hoffman et al., 2018).

In this study, the single scheme sensitivity experiments are conducted as the "pre-employment test" to pick the schemes with relatively low deterministic forecast errors. The combination of KF (cumulus), Lin (microphysics) and YSU (PBL) schemes is set to be the control experiment (CTL), considering that these classic schemes are widely used in TC high resolution studies (e.g., Choudhury & Das, 2017; Mohan et al., 2019; Nekkali et al., 2022; Rogers, 2010), and the Lin scheme has been reported to produce the strongest TCs among various microphysics schemes (e.g., Maw & Min, 2017; Tao et al., 2011) which may improve the intensity forecasts for category 4–5 TCs in this study. For each sensitivity experiment, only one physical scheme is changed from the CTL configuration, while two other schemes remain unchanged. There are a total of 15 sensitivity experiments and 1 CTL experiment conducted for six evaluated TCs, as listed by Table 2. The sensitivity experiments with relatively lower deterministic forecast errors indicate better performance of the schemes different from the CTL configuration. The combinations of these schemes are likely to produce lower errors in deterministic forecasts as well as ensemble forecasts. Note that the setup of CTL and sensitivity experiments has no meaning other than narrowing

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Sensitivity Experimental Design

	Cumulus	Microphysics	PBL
CTL	KF	LIN	YSU
1	BMJ	LIN	YSU
2	GF	LIN	YSU
3	G3	LIN	YSU
4	Tiedtke	LIN	YSU
5	NSAS	LIN	YSU
6	KF	WSM5	YSU
7	KF	Ferrier	YSU
8	KF	WSM6	YSU
9	KF	Thompson	YSU
10	KF	WDM6	YSU
11	KF	LIN	MYJ
12	KF	LIN	MYNN3
13	KF	LIN	ACM2
14	KF	LIN	BouLac
15	KF	LIN	UW

the range of combination candidates, which can also be achieved by testing all combinations of the 6 cumulus, 6 microphysics and 6 PBL schemes.

After obtaining the combinations of schemes with relatively lower errors, it is time to "interview" the combinations comprehensively. The deterministic forecasts and the ensembles forecasts with the configurations of these physical parameterization combinations are carried out and compared. As mentioned in the introduction, this study mainly focuses on the model uncertainty and its interaction with the physical parameterization scheme errors, thus employs SKEB perturbations to construct ensemble forecasts for the pre-selected combinations. SKEB adds random perturbations, with prescribed temporal and spatial correlations, to the physical parameterization tendency of horizontal wind and potential temperature at every time step. The temporal correlation is determined by the autoregressive parameter which is the quotient of the time step and a given decorrelation time. The spatial correlation is determined by the perturbed wavenumber spectrum which is correlated with the dimensions of the WRF domain. The default configurations of SKEB in WRF v4.0 are used in this study, including the default total backscattered dissipation rate which controls the perturbation amplitude, the default decorrelation time which determines the temporal correlation, the default addition only to the outermost domain with all wavenumbers involved. The perturbations are set to be vertically incoherent with a westward tilt, just like those in Judt et al. (2016). The SKEB ensemble forecasts with various physical parameterization combinations are conducted for six TCs, each consisting of 10 members. Since the domains and resolutions are identical, the SKEB perturbations adding to the forecasts with various physical parameterization combinations are the same. Thus the resulting differences between various forecast distributions can reflect the performance differences of various schemes which interact with the identical model perturbations.

In the present study, the deterministic forecast error for TC track, minimum sea level pressure (Min SLP) and 10-m maximum surface wind (Max Wind) is computed as the absolute error at every lead time. When pooling all lead times and all cases together, it is expressed as the RMSE averaged over all cases:

$$\mathrm{RMSE} = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[y_k(t) - o_k(t)\right]^2}. \tag{1}$$

where $y_k(t)$ and $o_k(t)$ represent the forecast and observed variables (location, Min SLP or Max Wind) at lead time t for the k th case, T is the total number of lead times. As six TCs have different simulation periods from 5 to 8 days while the frequencies of the forecast output and the observation are equally 3 hr, T varies from 40 to 64 for various cases. N is the total number of evaluated cases, here refers to six.

For ensemble forecasts, the ensemble mean error is calculated the same as the deterministic forecast error, except for substituting the ensemble mean $\overline{y}$ for the single forecast value y. The ensemble spread is computed as the standard deviation σ at every lead time, and is expressed analogously to the average RMSE when pooling all lead times and all cases together, as follows,

$$\mathrm{Spread} = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{(n-1)}\sum_{i=1}^{n}\left[y_{\mathrm{ki}}(t) - \overline{y}_k(t)\right]^2}. \tag{2}$$

where $y_{\mathrm{ki}}(t)$ and $\overline{y}_k(t)$ represent the i th forecast member and the ensemble mean at lead time t for the k th case. n is the total number of ensemble members, here referring to ten.

3. Results

3.1. Single Scheme Sensitivity

A few schemes with relatively low deterministic forecast errors need to be selected to form comparable scheme combinations primarily. Figure 1 presents the deterministic forecast errors of the CTL and the other 15 single

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

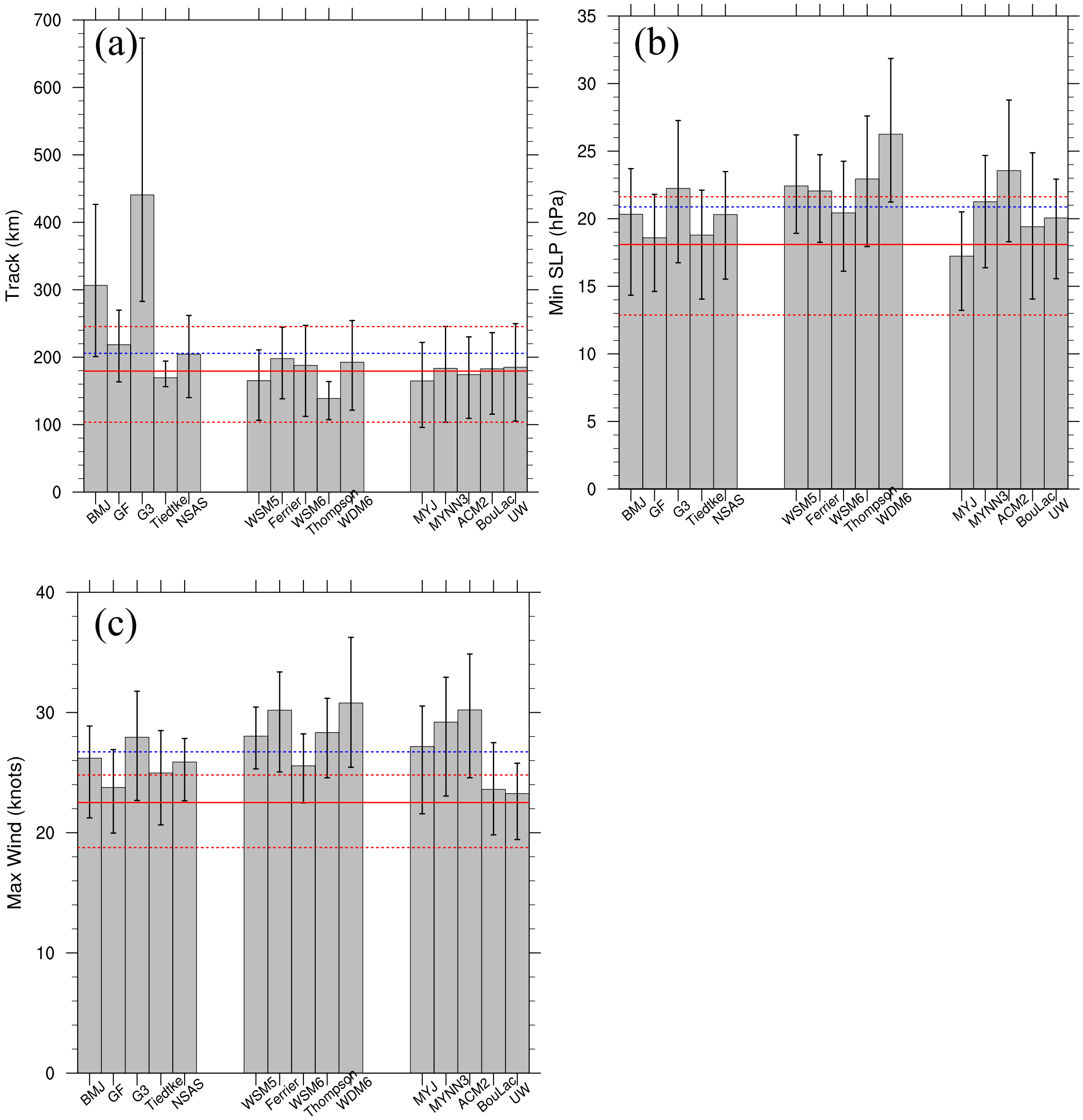

Figure 1. The deterministic forecast errors of the CTL (red solid line) and 15 single scheme sensitive experiments (gray bars) for (a) track, (b) Min SLP, (c) Max Wind, pooling all lead times and evaluated cases together. The vertical lines attached to each error bar indicate the 5%–95% bootstrap confidence intervals, the red dash lines indicate the confidence intervals of the CTL experiment, and the blue dash line represents the average of 16 experiments.

scheme sensitivity experiments listed in Table 2, computed by pooling all lead times and evaluate cases together. Since the error of each scheme is averaged over limited number of TC cases, the 5%–95% confidence interval of the average is estimated by a bootstrap method ($N = 1,000$).

There are diverse sensitivity characteristics of the TC track and intensity forecasts to three types of physical parameterization schemes. The track errors produced by the cumulus schemes vary much more greatly compared to those produced by the microphysics and PBL schemes, with the lowest error from the Tiedtke being 170 km while the highest error from the G3 being up to 440 km (Figure 1a). So the standard deviation of the cumulus scheme errors is 102 km, much larger than that of other two types of schemes (24 and 9 km respectively). But the intensity errors are a bit more uniform among cumulus schemes than among other two types of physical schemes (Figures 1b and 1c), with the standard deviations of Min SLP errors being 1.4, 2.1, and 2.3 hPa and the standard deviations of Max Wind errors being 1.5 knots, 2.0 knots and 3.0 knots, respectively. It is evident that the TC track forecasts are much more sensitive to cumulus schemes than other two types of physical schemes,

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Six Physical Parameterizations Combinations of Cumulus, Microphysics and PBL Schemes

	C 1	C 2	C 3	C 4	C 5	C 6
Cumulus	KF	KF	KF	Tiedtke	Tiedtke	Tiedtke
Microphysics	Lin	Lin	Lin	Lin	Lin	Lin
PBL	YSU	BouLac	UW	YSU	BouLac	UW

whereas the TC intensity forecasts are sensitive to all three types of physical parameterization schemes. The dependence variation of track and intensity forecast errors on the physical processes is similar to that in the previous studies (Di et al., 2019; Li & Pu, 2008; Mohan et al., 2019; Srinivas et al., 2013; Tao et al., 2011).

The deterministic forecast errors of all 16 experiments are ranked for the selection. The schemes sorted from low to high *track errors* are: Thompson, MYJ, WSM5, Tiedtke, ACM2, CTL, BouLac, MYNN3, UW, WSM6, WDM6, Ferrier, NSAS, GF, BMJ and G3. The schemes sorted from low to high *Min SLP errors* are: MYJ, CTL, GF, Tiedtke, BouLac, NSAS, WSM6, UW, BMJ, MYNN3, Ferrier, WSM5, G3, Thompson, ACM2 and WDM6. The schemes sorted from low to high *Max Wind errors* are: CTL, BouLac, UW, GF, WSM6, Tiedtke, NSAS, BMJ, MYJ, G3, WSM5, Thompson, MYNN3, Ferrier, ACM2 and WDM6. The error differences between adjacent schemes are not statistically significant among limited number of TC cases as shown by bootstrap confidence intervals.

The deterministic forecast error of the CTL can be seen as the criterion in comparing the performance of 15 single scheme experiments since there is only one single physical scheme changed from the CTL configuration of each sensitivity experiment. For TC track forecasts, the CTL configuration has an above-average performance among all experiments (Figure 1a). The error differences between most schemes and the CTL are not very significant, except for the G3 and BMJ, which produce significantly larger errors than the CTL. For the TC intensity forecasts, the CTL almost performs the best besides the MYJ in Min SLP forecasts (Figure 1b). Therefore, the CTL configuration is indeed of superiority for TC forecasts and the three schemes combining it are chosen as candidates.

As shown in the sensitivity experiments for three types of physical parametrization schemes, the cumulus schemes with low track errors, as well as the microphysics schemes and PBL schemes with low intensity errors need to be especially considered during the selection. As Tiedtke is the only cumulus scheme among the schemes of lowest track errors and also performs well in intensity forecasts, it is chosen as a candidate. The WSM6 produces lower intensity errors than other four microphysics schemes, but is still not comparable to the default Lin microphysics scheme in CTL (Figures 1b and 1c), and has error rankings just general in the track forecasts among 16 experiments (Figure 1a), so no microphysics scheme is chosen. The BouLac and the UW of PBL schemes perform above-average in the track and Min SLP forecasts and rank almost the best in Max Wind forecasts, thus they are also chosen as candidates. Though the MYJ scheme ranks very well in Min SLP forecasts and track forecasts, it has Max Wind performance lagging much more behind, revealing the poor ability of catching the accurate wind-pressure relationship, which may be caused by the surface layer scheme matched with MYJ not supporting an appropriate air-sea flux parameterization. Finally, the single schemes with relative lower deterministic forecast errors are selected: the KF and Tiedtke for cumulus schemes, the Lin for microphysics schemes, the YSU, BouLac, and UW for PBL schemes. Based on these schemes, six physical parameterization combinations (C1–C6) are constructed as shown in Table 3.

It is interesting that the confidence intervals of the scheme errors exhibit distinct widths. For example, the G3 produces the highest track errors with the widest confidence interval of 390 km, while the Tiedtke has relative lower track errors with the narrowest confidence interval of less than 40 km, suggesting the big differences among the interactions of the TC flow uncertainty and various scheme errors. The performances of some schemes are significantly case dependent whereas the performances of others are stable among TC cases. The uncertainty of TC flows is not as easily quantified as the model uncertainty or the initial condition uncertainty due to the large variance of TC events. But this phenomenon denotes the importance of considering all kinds of uncertainty sources during the evaluations.

Note that the preselection process based on the CTL configuration has the risk of losing other potential well-performed combinations. The schemes not chosen may produce lower errors combining with other configurations and the combinations may have good performances in the uncertainty context. However, the preselection process is a compromise to the limited computation resources. With the relatively small number of combinations, the performance of physical parametrizations in ensemble forecasts can be evaluated efficiently.

3.2. Deterministic and Ensemble Forecast Errors

After obtaining the combinations of physical parameterization schemes with relatively lower errors in Table 3, the deterministic forecasts and the SKEB perturbed ensembles forecasts with these physical parameterization

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

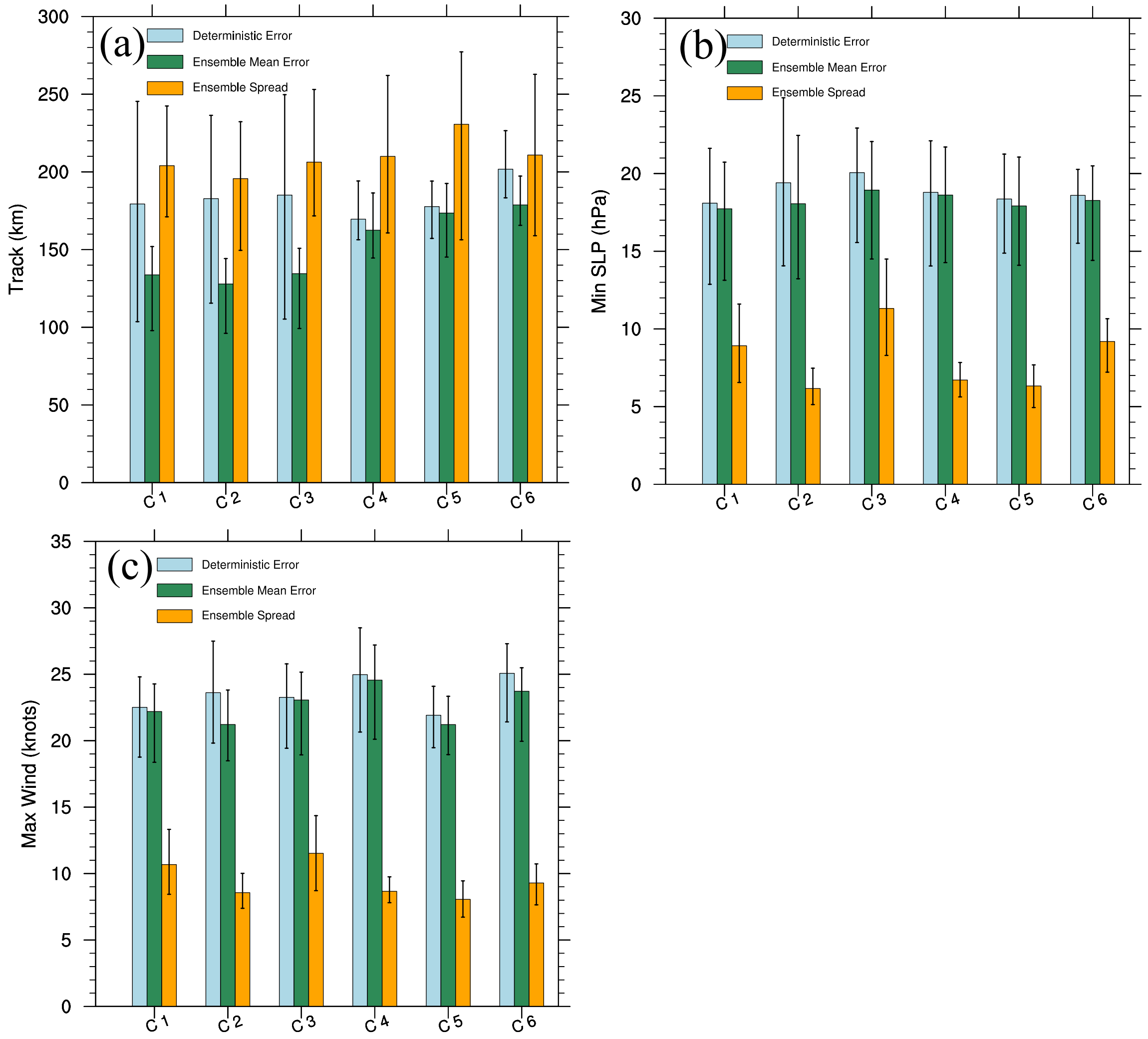

Figure 2. The deterministic error (blue), the ensemble mean error (green) and the ensemble spread (orange) of six combinations for (a) track, (b) Min SLP, (c) Max Wind forecasts, pooling all lead times and evaluated cases together. The vertical lines indicate the 5%–95% bootstrap confidence intervals.

combinations are carried out (Figure 2). The mean of deterministic forecast errors over six combinations are lower than those of 16 experiments as shown in Figure 1, with the former of 183 km, 19 hPa, 24 knots and the latter of 206 km, 21 hPa, 27 knots for track, Min SLP and Max Wind forecasts respectively. The standard deviations of the deterministic forecast errors of six combinations are reduced as well compared to those of 16 experiments, with the former of 11 km, 0.9 hPa, 1.4 knots and the latter of 72 km, 2.2 hPa, 2.5 knots. The reductions of the mean deterministic forecast errors and the standard deviations imply that the obtained combinations are indeed comparable with similar good performances.

Introducing the model uncertainty represented by SKEB perturbations has shown pronounced effects on the performance of combinations. Compared to the deterministic forecast errors, the ensemble mean errors of six combinations are all declined while the declines are more significant in track errors than that in intensity errors according to the bootstrap confidence intervals. The track errors of C1, C2 and C3 show the largest declines (though not significant according to the confidence intervals) and the confidence intervals of the track errors exhibit a sharp narrowing of nearly 33% (Figure 2a), indicating the combinations not only reducing errors but also performing more stable among TC cases in the context of model uncertainty. Thus the ensemble perturbations prove reasonable and effective in reducing the stochastic errors in TC forecasts.

Despite all producing lower ensemble mean errors than the deterministic forecast errors, the six combinations reduce errors by different degrees, leading to changes in the relative performances of six combinations. For

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

track forecasts, C1, C2 and C3 originally perform worse than C4 and C5 in deterministic forecasts, but they surpass the latter two combinations in terms of the ensemble mean errors (Figure 2a). For intensity forecasts, C2 produces higher deterministic forecast errors than C1 but has the ensemble mean errors similar to C1 in Min SLP forecasts and lower than C1 in Max Wind forecasts (Figures 2b and 2c). The performance difference between the deterministic forecast error and the ensemble mean error of the scheme combination denotes the importance of considering the forecast uncertainty for the evaluation, since the deterministic forecast value is only a sample of the forecast distribution and not able to represent the overall performance of the parameterization combination.

Carrying out the evaluations in the uncertainty context, or in the ensemble forecasts, should not neglect the ensemble spread, which estimates the forecast uncertainty and can reflect the reliability of ensemble forecasts compared with the ensemble mean error. Here, the TC track and intensity forecasts are characterized by different spread–error relationships. For track forecasts, the ensemble spreads are larger than the ensemble mean errors, no matter significantly (C1, C2, C3) or not significantly (C4, C5, C6) (Figure 2a). But for Min SLP and Max Wind forecasts, the ensemble spreads are significantly smaller than the ensemble mean errors of all combinations (Figures 2b and 2c). The results indicate that when pooling all lead times together, the ensemble spread efficiently represents the errors in TC track forecasts and even becomes over-dispersive, while showing severe under-dispersion in TC intensity forecasts. The much larger error relative to the spread is mainly due to the systematic bias of the TC intensity forecasts, which is a common issue in current operations and research (e.g., Aemisegger, 2009; Torn, 2016; Zhang, 2018). And for intense TCs the intensity underestimation is especially heavy (e.g., Lei et al., 2022). Besides, the development of SKEB perturbations cannot catch up with that of the initial condition errors in short terms. There has been an intensity error from the initialization time which grows fast during the RI period, while the ensemble spread starts from zero and grows at a much slower rate (Figure 3c). The above reasons lead to the inconsistency of spread–error relationship in TC intensity ensembles, emphasizing the importance of TC initialization or data assimilation as well as developing physical parameterizations more suitable for TC intensifying conditions.

The performances of six combinations are different not only in terms of the ensemble mean error, but also in terms of the ensemble spread. For intensity forecasts, C3 has a significantly larger spread than C2, C4, C5 albeit the ensembles of these combinations all being under-dispersive (Figures 2b and 2c). For track forecasts, the combinations also exhibit different values of the ensemble spread, but are less significant than that of intensity forecasts (Figure 2a). The spread variance denotes that even in an identical uncertainty context represented by SKEB perturbations, the ensemble forecasts with six parameterization combinations still disperse to different degrees and generate probabilistic distributions of different widths. This is just because of the distinct interactions of the model uncertainty and various combination errors. Thus, the ensemble mean error and the ensemble spread need to be considered concurrently for the evaluations. However, an objective selection is hard to achieve since the relative performances of two ensemble metrics are not necessarily consistent for each combination. For example, C4 produces the highest ensemble mean error while C3 has the largest ensemble spread in Max Wind forecasts (Figure 2c). Besides, the error and the spread computed by pooling all lead times and all cases together do not coincide with the error and the spread at every lead time for a single TC case. Therefore, it is necessary to combine the ensemble mean error with the ensemble spread in a more natural way, that is, evaluating the combinations based on the overall quality of the forecast distribution.

4. Assessing the Overall Performance of Scheme Combinations

A number of scores have been raised for the verification of forecast probabilistic distributions (e.g., Gneiting & Raftery, 2007; Roulston & Smith, 2002; Wilks, 2019). Continuous Ranked Probability Score (CRPS) is one of those used most commonly (Matheson & Winkler, 1976). It is defined as the integral of the squared difference between the cumulative distribution function of the probabilistic forecast $F(y)$ and the cumulative distribution function of the observation $F_o(y)$,

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 dy, \tag{3}$$

where

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

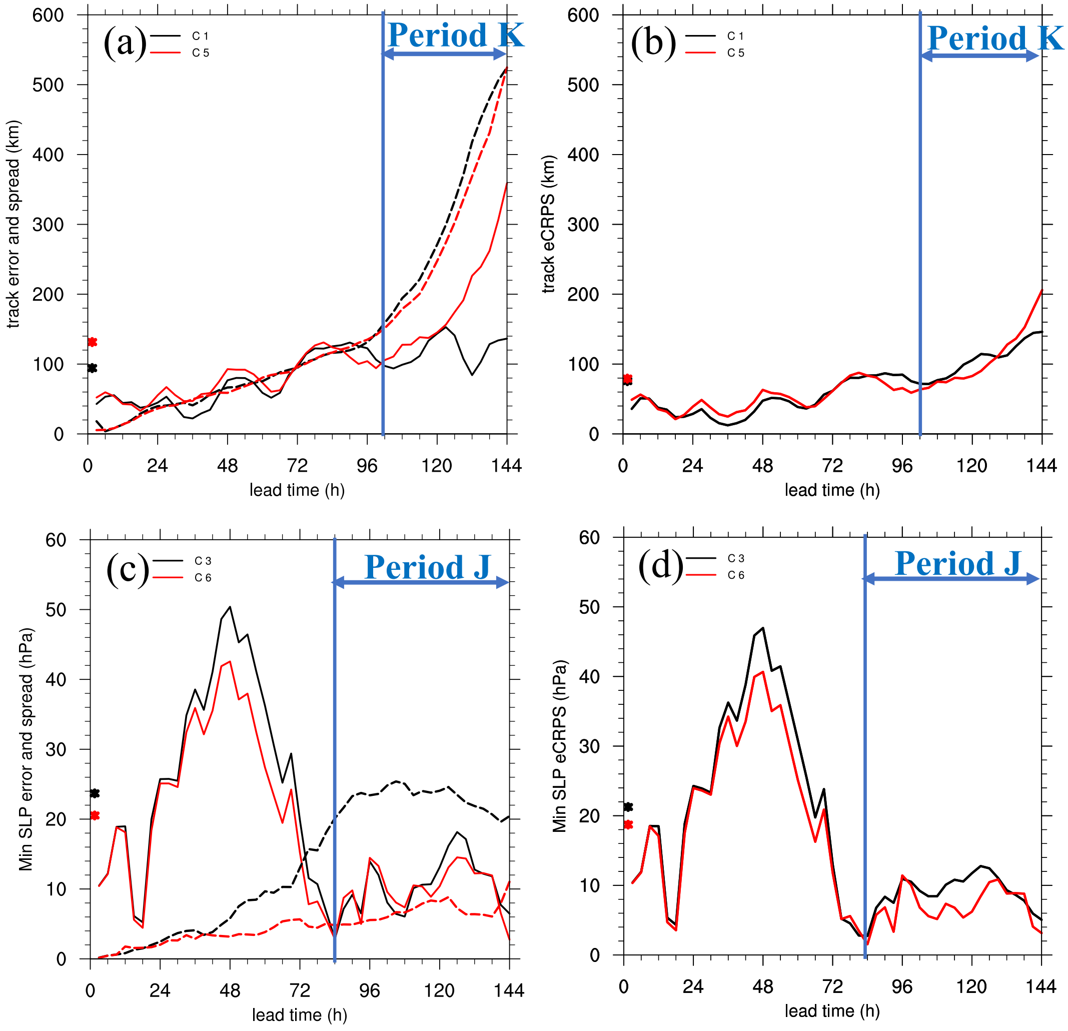

Figure 3. The ensemble mean error (solid line) and the ensemble spread (dashed line) as a function of lead time (a), (c), and the eCRPS as a function of lead time (b), (d) of two combinations for the single TC forecast, where (a), (b) for the track forecasts of Kong-rey (2018), and (c), (d) for the Min SLP forecasts of Jebi (2018). The asterisks next to the left vertical axis indicate the ensemble mean error or the eCRPS pooling all lead times together of two combinations. Period K and Period J represent the time periods of focused analysis for Kong-rey and Jebi, respectively.

$$F_o(y) = \begin{cases} 0, y < o \\ 1, y \geq o \end{cases}, \tag{4}$$

$F_o(y)$ is a cumulative-probability step function that jumps from 0 to 1 at the point where the forecast variable y equals to observation o.

Alternatively, the CRPS can also be formulated as (Gneiting & Raftery, 2007)

$$\text{CRPS} = E_F|Y - o| - \frac{1}{2}E_F|Y - Y'|, \tag{5}$$

where E_F denotes the statistical expectation of the continuous variable in $|\cdot|$, of which the forecast distribution is $F(y)$. Y and Y' are different samples from the continuous forecast value y. The first term reflects the average distance of every forecast sample Y and the observation o. And the second term reflects the average distance of all pairs of two different forecast samples.

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

CRPS has the same units as the forecast variables, and reduces into the absolute error when applied to a single deterministic forecast, thus the score provides a direct way to compare the deterministic forecast and the probabilistic forecast (Gneiting & Raftery, 2007; Hersbach, 2000). Similar to the absolute error, a lower value of CRPS indicates better skill. According to the physics and the hypothetical example in Wilks (2019, pp. 425–426), the forecast distribution tends to be rewarded by CRPS when concentrating around the observed value, which influenced not only by the mean error but also by the spread assuming a unimodal distribution. CRPS is able to quantify the accuracy of the probabilistic forecast, and it is a strictly proper score while the ensemble mean error or the ensemble spread is not (Du, 2021; Smith et al., 2015). "Strictly proper" means that the score achieves the lowest only when the best forecast distribution occurs (Bröcker & Smith, 2007; Gneiting & Raftery, 2007), in which the observation is statistically indistinguishable from any of the ensemble members, thus the forecast bias is zero and the ensemble spread equals to the ensemble mean error. Being strictly proper is important for the verification to be carried out honestly.

Considering that the ensemble forecasts estimate the forecast probabilistic distribution with limited number of ensemble members, a discrete expression of Equation 5 or the "ensemble" CRPS (Gneiting & Raftery, 2007; Wilks, 2019, p. 444), which possesses exactly the same properties as CRPS, is more appropriate to evaluate the quality of ensemble forecasts.

$$\mathrm{eCRPS} = \frac{1}{n}\sum_{i=1}^{n}|y_i - o| - \frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}|y_i - y_j|. \tag{6}$$

where y_i and y_j are different members of the ensembles and n is the total number of ensemble members. Therefore, the performance of various physical parameterization combinations in ensemble forecasts can be quantified and evaluated directly in terms of eCRPS, albeit being both different in the ensemble mean error and in the ensemble spread. Similar to the computation of the ensemble mean error and the ensemble spread, the eCRPS is computed at every lead time for TC track, Min SLP and Max Wind forecasts, and turns to a single value in the same way when pooling all lead times and all evaluated cases together.

In order to examine how eCRPS behaves in reflecting the overall performance of the scheme combinations in ensemble forecasts, the eCRPS as a function of lead time is first analyzed for two combinations in the forecasts of the single TC case, and is compared with the ensemble mean error as well as the ensemble spread as a function of lead time.

Figure 3a shows the times series of the ensemble mean error and the ensemble spread of C1 and C5 in the track forecasts for Kong-rey (2018), and Figure 3b displays the corresponding time series of eCRPS. The track error and spread of two combinations increase monotonically with increasing forecast lead times (Figure 3a), agreeing well with the normal evolution of the track uncertainty (e.g., Judt et al., 2016). Before Period K, the error and the spread of C1 and C5 grow at similar rates. The error of C5 is slightly higher than C1 while the spreads of two combinations are basically the same, thus the differences between two forecast distributions mainly come from the error differences, with C5 exhibiting slightly higher eCRPS than C1. In Period K, the spreads grow much faster and both ensembles become over-dispersive for the two combinations. The error of C5 also grows fast at a similar rate as the spread and reaches to 360 km at the end of the forecast, while the error of C1 keeps fluctuating between 90 and 160 km. Though producing higher errors, C5 exhibits an obviously better spread–error relationship than C1. So during the former part of Period K when the error differences between two combinations are little, C5 has slightly lower eCRPS than C1. Even during the latter part of Period K when the error differences become large, the eCRPS of C5 is not much higher than C1. The increase in errors appears to be offset to some extent by the good consistency between the error and the spread. So C5 exhibits the eCRPS of 200 km at the end of the forecast, much lower than the ensemble mean error. On the contrary, despite no increase in errors during Period K, the over-dispersion of the C1 ensemble forecast is strengthening. This leads to a worsening forecast distribution for C1, which can be reflected by the continuously increasing eCRPS during Period K. When pooling all lead times together, the relative performances of two combinations in terms of the ensemble mean errors are different from that in terms of the quality of forecast distributions, as the mean error of C5 is 40 km higher than C1 while the mean eCRPS of two combinations are of similar values.

Different from the monotonic increase of track errors, the maximum of intensity errors is strongly associated with the rapid intensification (RI) process. Figure 3c displays the times series of the ensemble mean error and the

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ensemble spread of C3 and C6 in the Min SLP forecasts for Jebi (2018), and Figure 3d displays the corresponding time series of eCRPS. Before Period J, the Min SLP errors of two combinations all exhibit the trends of first increasing and then declining, and reach the maximum at the peak intensity of Jebi (2018). The spreads of two combinations grow slowly and are much smaller than the errors, leading to the ensembles for two combinations very under-dispersive at similar degrees. During this period, the quality of the forecast distribution is dominated by the ensemble mean errors, with C3 exhibiting higher eCRPS than C1. In Period J, the errors of two combinations grow again at a slow rate while the spreads of two combinations show different features, with the spread of C3 being much larger than that of C6. During the former part of Period J, the errors of two combinations are of similar values as the spread of C6 but are much lower than the spread of C3, denoting that C3 is over-dispersive while C6 has a good spread–error relationship. Thus the corresponding eCRPS of C3 is higher than C6, albeit the error of C3 is slightly lower. However, during the latter part of Period J, the errors of two combinations grow to the similar values as the spread of C3 and become much larger than the spread of C6, making the spread–error relationship of C3 turn better while that of C6 becomes under-dispersive. The corresponding eCRPS explicitly reflects the changes of the quality of forecast distributions, with C6 exhibiting much higher eCRPS than C3 during the latter part of Period J. When pooling all lead times together, C6 shows a better performance both in terms of the eCRPS and the ensemble mean errors. But the mean eCRPS difference of two combinations is only half of the mean error difference, suggesting that the performance differences of the forecast distributions for two combinations are not so large as denoted by the ensemble mean errors.

The forecasts of two TC cases indicate that the evaluations based on the eCRPS include more comprehensive information of the forecast probability distribution compared to that based on the ensemble mean error, thus are effective in reflecting the overall performance of scheme combinations in ensemble forecasts. In addition, it can be shown in Figure 3 that at the initial forecast time, when the ensemble spread is small and various ensemble members can be seen as a single forecast, the eCRPS is basically equal to the ensemble mean error. As the spread grows, the value of eCRPS is overall reduced relative to the ensemble mean errors. Pooling all lead times together, the eCRPS also exhibits lower value than the ensemble mean error. Since eCRPS provides a direct way to compare the deterministic forecast and the probabilistic forecast, the above phenomenon shows the superiority of the forecast distribution over a single ensemble mean.

Figure 4 shows the eCRPS as well as the ensemble mean error and the deterministic forecast error of six combinations pooling all lead times and all evaluated cases together. Given that eCRPS reduces into the absolute error when applied to a single forecast, the eCRPS is the same as the deterministic forecast error for a deterministic forecast, and is the same as the ensemble mean error for an ensemble mean forecast. No matter in track or intensity forecasts, the eCRPS all decline relative to the ensemble mean errors, just like that the ensemble mean errors all decline relative to the deterministic forecast errors, indicating that the forecast performance continuously improves from the deterministic forecast to the ensemble mean forecast, and finally to the probabilistic forecast which includes the largest amount of forecast information.

As shown in Section 3, the relative performances of six combinations in the ensemble mean forecasts are not consistent with that in the deterministic forecasts. The relative performances of six combinations in the probabilistic forecasts are not consistent with that in the deterministic forecasts as well (Figure 4), confirming that the evaluations based on the deterministic forecast errors is limited. Moreover, the relative performances of six combinations in the probabilistic forecasts have changed from that in the ensemble mean forecasts. Take Min SLP forecasts for example, (Figure 4b), C3 has the highest ensemble mean errors, while C4 and C6 have the second highest errors. But the eCRPS of C3 and C6 are both lower than C4, rendering C4 to become the worst combination. The inconsistency between the eCRPS and the ensemble mean error may be due to the significantly larger spreads (i.e., better spread–error relationship) of C3 and C6 than C4 in Min SLP forecasts (Figure 2b), which affects the overall quality of the forecast probability distributions. Therefore, the evaluations only based on the ensemble mean error are as misleading as based on the deterministic forecast error. The quantification in terms of eCRPS is more appropriate to evaluate the overall performance of the physical parameterization combinations in ensemble forecasts.

The well-performed combinations for TC track, Min SLP and Max Wind forecasts emerge from the evaluated combinations respectively referring to Figure 4. For TC track forecasts, the best combination is C2, while C1 and C3 also produce relative lower values of eCRPS than other combinations. For Min SLP forecasts, albeit little differences among the eCRPS of six combinations, C1 and C5 rank the best. For Max Wind forecasts, C2 and C5 both produce the lowest values of eCRPS with C5 showing a narrower confidence interval than C2.

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

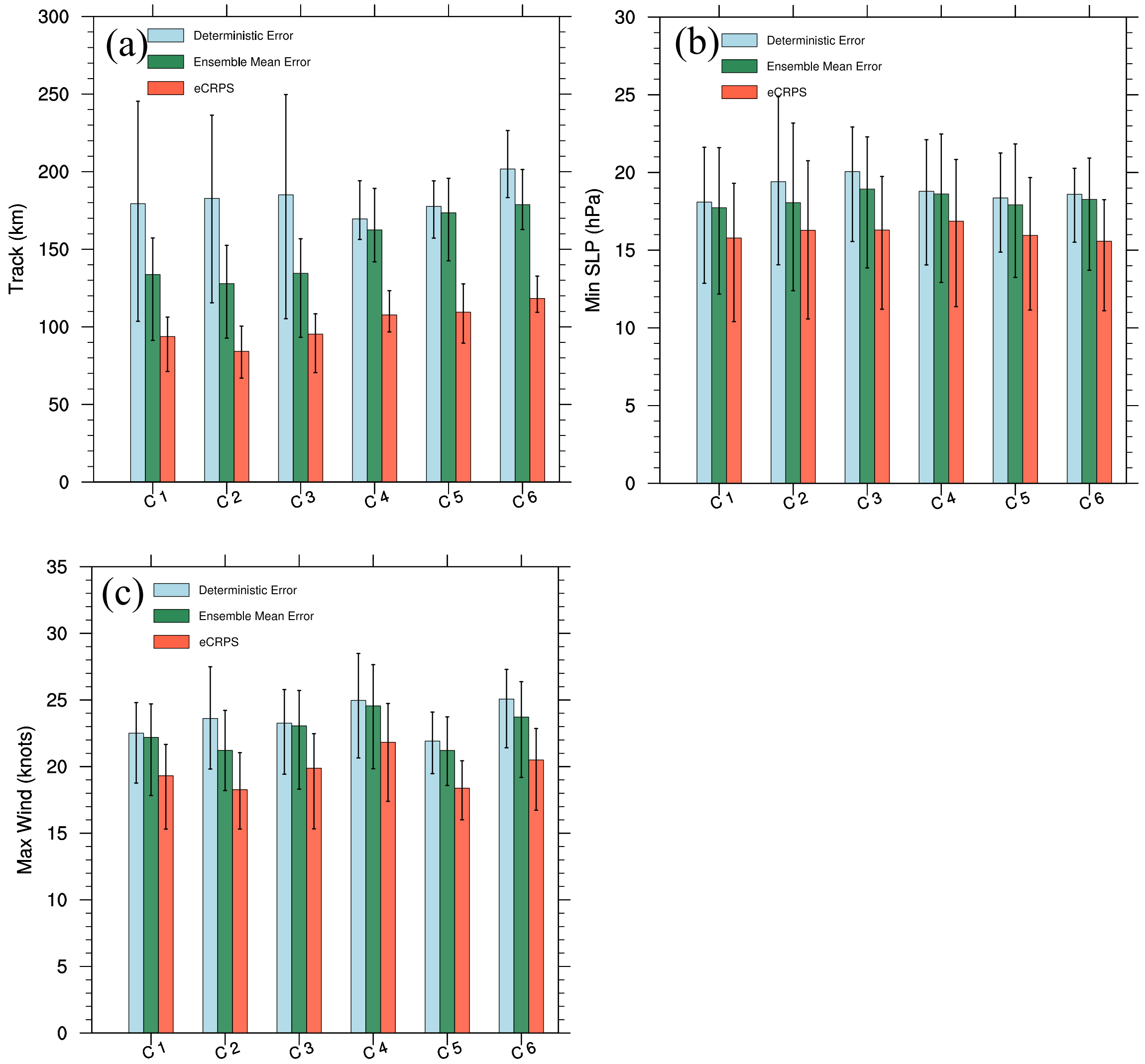

Figure 4. The deterministic error (blue), the ensemble mean error (green) and the eCRPS (red) of six combinations for (a) track, (b) Min SLP, (c) Max Wind forecasts, pooling all lead times and evaluated cases together. The vertical lines indicate the 5%–95% bootstrap confidence intervals.

Note that CRPS is not the only choice for verifying the probabilistic forecast. Another widely-used scoring rule is the Ignorance Score (Roulston & Smith, 2002). Ignorance Score is the only strictly proper score that also exhibits locality, which means the score depends solely on the forecast probability at the observation, not on other features of the forecast distribution (Bröcker & Smith, 2007; Du, 2021). However, Ignorance Score is very sensitive to outliers, and the forecast distribution far away from the observation is penalized heavily by the score (Wilks, 2019, pp. 428–429). Thus when the ensemble forecasts are very under-dispersive, like the RI period in Figure 3c, the Ignorance Score will be abnormally high (not shown). CRPS is not so sensitive to the outliers so the score keeps comparable at all lead times. Moreover, CRPS has the same unit as the observation and turns to the absolute error for a deterministic forecast, thus provides a direct way to compare the deterministic forecast and the probabilistic forecast (Gneiting & Raftery, 2007), which meets the request of this study well. In general, the preference of scoring rules depends on the circumstances of forecast users, if the locality is viewed as a desirable property, then the Ignorance Score should be recommended (Du, 2021).

5. Assessing the Multivariate Performance

Instead of making a subjective selection according to the results of Section 4, this section employs a multivariate extension of eCPRS, the ensemble Energy Score (eES), to quantitatively assess the multivariate performance

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

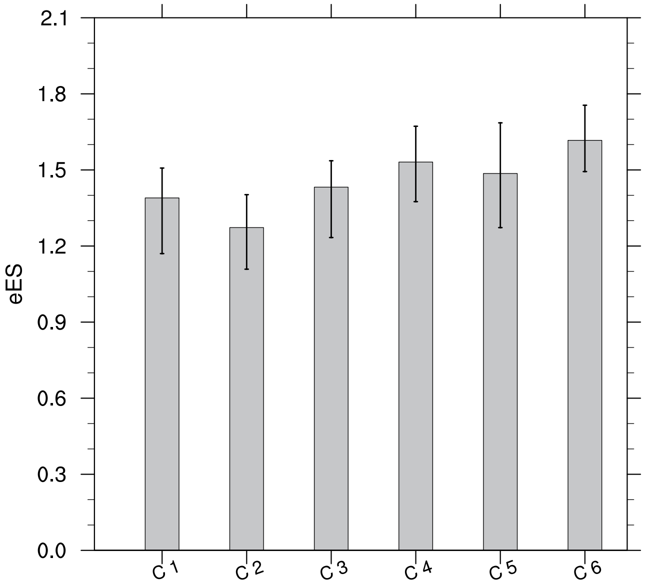

Figure 5. The standardized eES of six combinations, pooling all lead times and evaluated cases together. The vertical lines indicate the 5%–95% bootstrap confidence intervals.

of six combinations in TC forecasts. The eES is a commonly used score for multivariate verifications of ensemble forecasts (Gneiting et al., 2008). It combines the forecast probability distributions of different variables as a vector distribution and assesses the overall quality of the vector distribution.

$$\text{eES} = \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{y}_i - \boldsymbol{o}\| - \frac{1}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|. \tag{7}$$

where the Euclidean distance in the d-dimensional space of the vector ensemble $\|\cdot\|$ substitute for the absolute value in Equation 6. In case of $d = 1$, the eES reduces to the eCPRS. Similar to eCRPS, the eES is strictly proper and a lower value of the score indicates better performance. Since eES does not make any distinctions between the components of the forecast vector, the absolute differences ($|y_i - o|$ and $|y_i - y_j|$) of each variable are standardized by the standard deviation of all absolute errors (pooling all lead times and all TCs together) to be comparable in magnitude. Then the intensity ensemble averaged over the Min SLP and Max Wind ensembles is combined with the track ensemble as a vector ensemble to compute the eES at every lead time.

Figure 5 shows the eES of six scheme combinations, pooling all lead times and evaluated cases together as well. Obviously, C2 has the lowest value of eES, exhibiting the best multivariate performance. It is reasonable as C2 is the best combination in track forecasts and performs relatively good in intensity forecasts. Besides, C1 and C3 rank the second and the third respectively in terms of eES among six combinations, also exhibiting a relative good multivariate performance. The multivariate performance of the combinations seems consistent with the track performance, but in fact there are indeed influences of intensity performance on eES. For example, the eES of C5 is smaller than C4, which is consistent with the relative performance in Min SLP and Max Wind forecasts (Figures 4b and 4c) but is not reflected in track forecasts (Figure 4a).

It is noteworthy that the eES was challenged on not sufficiently sensitive to the correlations among the components of the forecast vector (Pinson & Girard, 2012). There are other multivariate scores such as the variogram score (Scheuerer & Hamill, 2015) and the David-Sebastiani score (Dawid & Sebastiani, 1999) addressing this issue. However, these scores need to compute the differences or the covariance matrix between various components of the vector. They are not applicable for the TC track and intensity forecasts, as the location of TC is in fact a two-dimensional variable while Min SLP and Max Wind are one-dimensional variables. Considering the non-negligible correlations between the location and intensity for TCs, seeking a new variable or index which can reflect the multivariable features of TCs may provide an alternative way, so that the multivariate verifications will turn into the univariate ones based on the standard univariate scores (e.g., eCRPS).

6. Validation

In order to validate the well-performed combination C2 and other two combinations C1 and C3, these combinations are further analyzed with three intense TC cases, Lingling, Lekima and Hagibis in 2019 WNP (Table 4). For the three 2019 TC forecasts, the domains, the model set-up and the forcing data are the same as the 2018 TC forecasts.

Table 4
2019 TC Cases for Validation

Number	Name	Category	Simulation period	Days
1	Lingling	4	2019-09-03_06~2019-09-07_06	4
2	Lekima	4	2019-09-06_00~2019-09-10_00	4
3	Hagibis	5	2019-10-06_12~2019-10-12_12	6

Figure 6 shows the performances of C1, C2 and C3 in the track and intensity forecasts of three 2019 TC cases. The ensemble mean errors of three 2019 TCs all decline relative to the deterministic forecast errors, and the eCRPS all decline relative to the ensemble mean errors, just like the results of pooling all 2018 TCs together (Figure 4), further verifying the superiority of the probabilistic forecast over the ensemble mean or the deterministic forecast. Besides, the ensemble spreads are smaller than the ensemble mean errors both in track and intensity forecasts, which is slightly different from the overall situation of 2018 TCs that are under-dispersive in intensity forecasts but

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

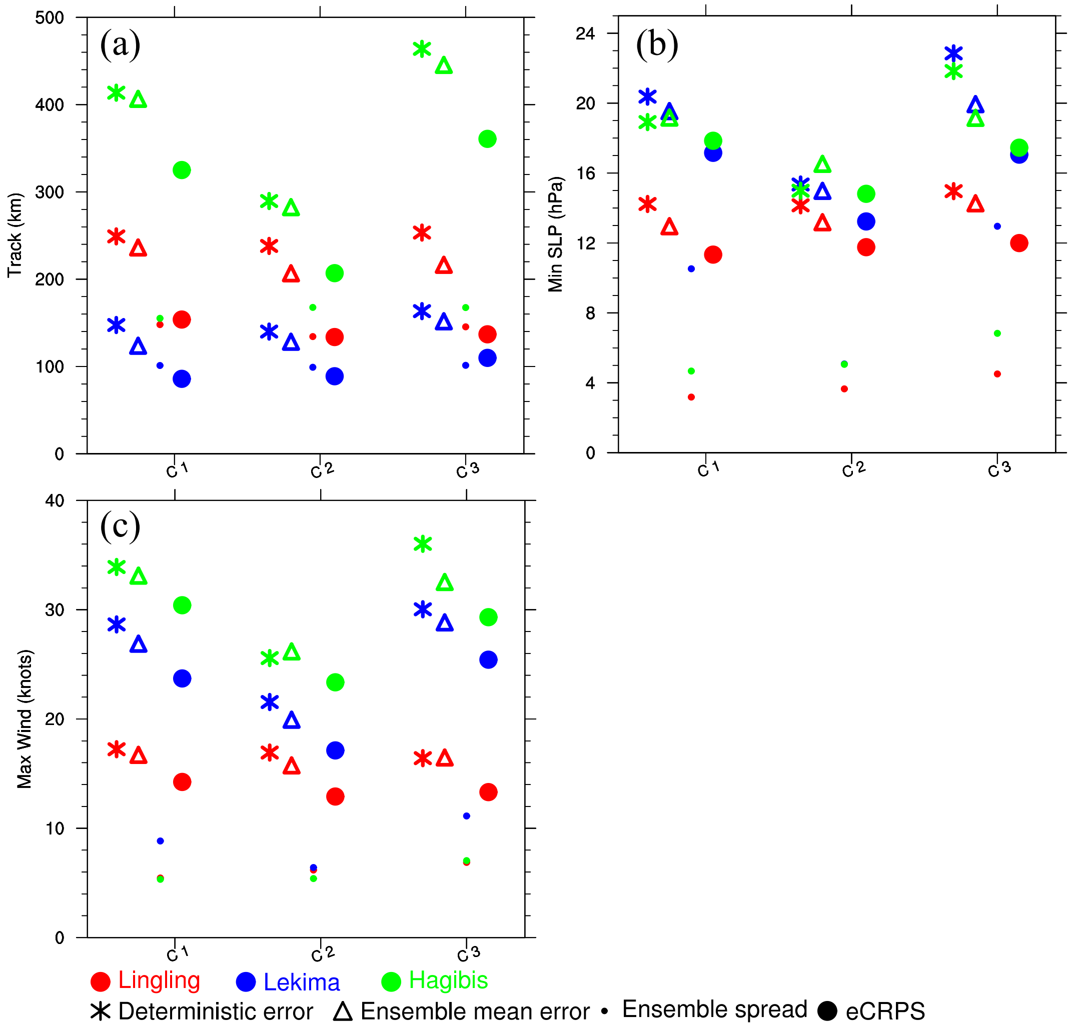

Figure 6. The deterministic error (asterisk), the ensemble mean error (triangle), the ensemble spread (dot) and the eCRPS (circle) of the selected combinations for (a) track, (b) Min SLP, (c) Max Wind forecasts, pooling all lead times together. Red, blue and green represent the three TC cases Lingling (2019), Lekima (2019) and Hagibis (2019), respectively.

over-dispersive in track forecasts (Figure 2). However, the under-dispersion of 2019 TC track forecasts is reasonable considering the much shorter simulation periods of three 2019 TCs than 2018 cases, as the ensemble spread grows fast at longer lead times for TC track forecasts (see Figure 3a).

Despite all showing declines from the deterministic forecast error to the ensemble mean error and finally to the eCRPS, the improvements of C1, C2 and C3 are of different degrees, similar to pooling all 2018 evaluated cases together (Figure 2). For example, in the Min SLP forecast for Hagibis (2019) (Figure 6b), the deterministic forecast error of C3 is higher than C1 while C3 performs better than C1 in terms of the ensemble mean errors and the eCRPS. Moreover, there are also spread variances among the combinations for individual 2019 TC forecasts, consistent with 2018 TC forecasts (Figures 2b and 2c). In the intensity forecasts of three cases, C3 still exhibits the largest spread among three combinations (Figures 6b and 6c), especially for Lekima (2019).

C2 basically produces the lowest eCRPS no matter in track forecasts (Figure 6a) or intensity forecasts (Figures 6b and 6c), with the superiority of this combination most obvious for Haigibis (2019). When it turns to the multivariate performance (Figure 7), C2 also produces the lowest eES among the selected combinations for all three 2019 TC cases. The performance is consistent with that obtained from evaluating the 2018 TC forecasts (Figure 5), thus effectively validates the selected combination of C2, which refers to the combination of the KF cumulus scheme, the Lin microphysics scheme and the BouLac PBL scheme.

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

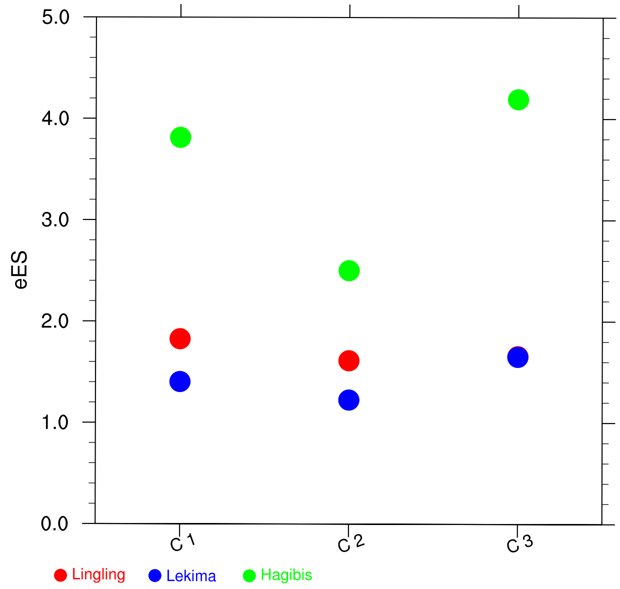

Figure 7. The standardized eES of the selected combinations, pooling all lead times together. Red, blue and green represent the three TC cases Lingling (2019), Lekima (2019) and Hagibis (2019), respectively.

7. Conclusion and Discussion

Much efforts have been made to select the proper physical parameterization schemes for TC track and intensity forecasts. In general, the evaluations of physical parameterization schemes and their combined performance were all based on the deterministic forecast error, which had inherent limitations to represent the overall performance due to the existence of the model uncertainty. This study introduces a realistic context of model uncertainty represented by the SKEB ensemble perturbations, and evaluates the combinations of cumulus, microphysics and PBL schemes in the model uncertainty context. Six 2018 WNP TCs with the intensity of at least category 4 are taken as evaluated cases. The deterministic forecasts and the SKEB-perturbed ensemble forecasts with some pre-selected parameterization combinations are performed and compared.

It is found that introducing the model uncertainty has pronounced effects on the performance of scheme combinations. The ensemble mean errors of all combinations are declined compared to the deterministic forecast errors, both in track and intensity forecasts. But the scheme combinations reduce errors by different degrees, leading to changes in the relative performances. There are differences not only between the ensemble mean errors but also between the ensemble spreads of various combinations, indicating that the errors of various physical parameterization combinations interact with the identical perturbations in different ways, and contribute to different forecast distributions.

The overall quality of the forecast distribution is quantified by the eCRPS, which shows capable of reflecting more features of the forecast than only by the ensemble mean error or the ensemble spread. The performances of scheme combinations in the model uncertainty context, evaluated by eCRPS, are improved from those in the deterministic context. Moreover, the relative performances of scheme combinations in terms of eCRPS are changed from those in terms of the ensemble mean errors, as the probabilistic forecasts exhibit more comprehensive information than the ensemble mean forecasts.

In order to obtain a well-performed scheme combination in both track and intensity forecasts, a multivariate extension of eCRPS, the eES, is employed to evaluate the overall quality of the combined forecast distributions of track, Min SLP and Max Wind. Finally, the KF cumulus scheme, the Lin microphysics scheme and the BouLac PBL scheme are identified to be the well-performed combination among the evaluated schemes for the forecasts of 2018 WNP intense TCs. The evaluation results are further validated in the forecasts of three intense TCs in 2019 WNP.

This study accounts for the nonlinear interactions of model errors through SKEB ensembles to evaluate the performance of combinations of physical parameterization schemes. But there are other uncertainty sources not included, especially the initial condition errors. The nonlinear interactions between initial condition errors and the physical parameterization scheme errors should also have influences on the performance of scheme combinations. However, it is impossible to reflect all kinds of uncertainty sources by limited members. Though it will be beneficial to reflect the combination performance more comprehensively by including the initial uncertainty, this study may be treated as the first step toward showing the influence of model uncertainty on the performance of scheme combinations.

Caution should be exercised as that although consistent results are found among the evaluation and validation processes, they may not be universally applicable due to the limited number of TC cases and experiments. More cases and experiments are needed for improving the statistical significance of the combination performance. The selected combination can only be viewed as the best-performed one among the evaluated schemes for the forecasts of six intense TCs. Some newly developed parameterization schemes which were denoted of physical superiority should also be included in the future. Moreover, the errors, the spreads as well as the scores are computed pooling all lead times of the TC cases together to explicitly compare the performance of various combinations, thus the error evolutions of different time lengths are reduced to single values. Strictly speaking, various lead

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

times are not equal. So the relative performances of combinations may change if the time lengths for computing are altered.

Note that the superiority of the selected combination is only for TC track and intensity forecasts. The forecast performance of TC size has not been considered, because it is difficult to quantify the multivariate performance of different scheme combinations considering the complex relationship between TC size and TC intensity as well as that between different size metrics (Guo & Tan, 2017). Including the size evaluation should be carried out with more cautions in the future work.

Data Availability Statement

This manuscript uses Weather Research and Forecast (WRF) model version 4.0 (National Center for Atmospheric Research, 2018) for TC forecasts, which is available at http://dx.doi.org/10.5065/D6MK6B4K. The input data for the forecasts comes from the GDAS analysis (National Centers for Environmental Prediction, 2015), which is available at https://doi.org/10.5065/D65Q4T4Z. The observed TC positions and intensities come from the International Best Track Archive for Climate Stewardship (IBTrACS, Knapp et al., 2018), which is available at https://doi.org/10.25921/82ty-9e16. The figures are plotted by NCAR Command Language 6.6.2 (National Center for Atmospheric Research, 2019), which is accessible at http://dx.doi.org/10.5065/D6WD3XH5.

Acknowledgments
This study is jointly supported by the National Natural Science Foundation of China under Grants 42192555 and 61827901. We would like to thank the reviewers and the editors for their thorough comments, which help make great improvements on this paper.

References

Aemisegger, F. (2009). *Tropical cyclone forecast verification: Three approaches to the assessment of the ECMWF model (Master's thesis)*. Eidgenssische Technische Hochschule Zürich. Retrieved from https://iacweb.ethz.ch/doc/publications/TC_MasterThesis.pdf

Berner, J., Ha, S. Y., Hacker, J. P., Fournier, A., & Snyder, C. (2011). Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Monthly Weather Review*, *139*(6), 1972–1995. https://doi.org/10.1175/2010MWR3595.1

Bougeault, P., & Lacarrere, P. (1989). Parameterization of orography-induced turbulence in a mesobeta-scale model. *Monthly Weather Review*, *117*(8), 1872–1890. https://doi.org/10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2

Bretherton, C. S., & Park, S. (2009). A new moist turbulence parameterization in the community atmosphere model. *Journal of Climate*, *22*(12), 3422–3448. https://doi.org/10.1175/2008JCLI2556.1

Bröcker, J., & Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, *22*(2), 382–388. https://doi.org/10.1175/WAF966.1

Brown, B. R., Bell, M. M., & Frambach, A. J. (2016). Validation of simulated hurricane drop size distributions using polarimetric radar. *Geophysical Research Letters*, *43*(2), 910–917. https://doi.org/10.1002/2015GL067278

Brown, B. R., Bell, M. M., & Thompson, G. (2017). Improvements to the snow melting process in a partially double moment microphysics parameterization. *Journal of Advances in Modeling Earth Systems*, *9*(2), 1150–1166. https://doi.org/10.1002/2016MS000892

Chen, X. (2022). How do planetary boundary layer schemes perform in hurricane conditions: A comparison with large-eddy simulations. *Journal of Advances in Modeling Earth Systems*, *14*(10), 1–18. https://doi.org/10.1029/2022MS003088

Chen, X., Bryan, G. H., Hazelton, A., Marks, F. D., & Fitzpatrick, P. (2022). Evaluation and improvement of a TKE-based Eddy-Diffusivity Mass-Flux (EDMF) planetary boundary layer scheme in hurricane conditions. *Weather and Forecasting*, *37*(6), 935–951. https://doi.org/10.1175/waf-d-21-0168.1

Choudhury, D., & Das, S. (2017). The sensitivity to the microphysical schemes on the skill of forecasting the track and intensity of tropical cyclones using WRF-ARW model. *Journal of Earth System Science*, *126*(4), 1–10. https://doi.org/10.1007/s12040-017-0830-2

Dawid, A. P., & Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, *27*(1), 65–81. https://doi.org/10.1214/aos/1018031101

Di, Z., Gong, W., Gan, Y., Shen, C., & Duan, Q. (2019). Combinatorial optimization for WRF physical parameterization schemes: A case study of three-day typhoon simulations over the Northwest Pacific Ocean. *Atmosphere*, *10*(5), 233. https://doi.org/10.3390/atmos10050233

Donelan, M. A., Haus, B. K., Reul, N., Plant, W. J., Stiassnie, M., & GraberSaltzman, H. C. E. S. (2004). On the limiting aerodynamic roughness of the ocean in very strong winds. *Geophysical Research Letters*, *31*(18), L18306. https://doi.org/10.1029/2004GL019460

Du, H. (2021). Beyond strictly proper scoring rules: The importance of being local. *Weather and Forecasting*, *36*(2), 457–468. https://doi.org/10.1175/WAF-D-19-0205.1

Garratt, J. R. (1994). Review: The atmospheric boundary layer. *Earth-Science Reviews*, *37*(1–2), 89–134. https://doi.org/10.1016/0012-8252(94)90026-4

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, *17*(2), 211–235. https://doi.org/10.1007/s11749-008-0114-x

Gopalakrishnan, S., Hazelton, A., & Zhang, J. A. (2021). Improving hurricane boundary layer parameterization scheme based on observations. *Earth and Space Science*, *8*(3), e2020EA001422. https://doi.org/10.1029/2020EA001422

Grell, G. A., & Dévényi, D. (2002). A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophysical Research Letters*, *29*(14), 38-1–38–4. https://doi.org/10.1029/2002GL015311

Grell, G. A., & Freitas, S. R. (2014). A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmospheric Chemistry and Physics*, *14*(10), 5233–5250. https://doi.org/10.5194/acp-14-5233-2014

Guo, X., & Tan, Z.-M. (2017). Tropical cyclone fullness: A new concept for interpreting storm intensity. *Geophysical Research Letters*, *44*(9), 4324–4331. https://doi.org/10.1002/2017GL073680

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. *Quarterly Journal of Economics*, *133*(2), 765–800. https://doi.org/10.1093/qje/qjx042

Hong, S. Y., Dudhia, J., & Chen, S. H. (2004). A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Monthly Weather Review*, *132*(1), 103–120. https://doi.org/10.1175/1520-0493(2004)132<0103:ARATIM>2.0.CO;2

Hong, S. Y., & Lim, J. O. J. (2006). The WRF single-moment 6-class microphysics scheme (WSM6). *Journal of the Korean Meteorological Society*, *42*(2), 129–151.

Hong, S. Y., Noh, Y., & Dudhia, J. (2006). A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, *134*(9), 2318–2341. https://doi.org/10.1175/MWR3199.1

Huang, X., Peng, X., Fei, J., Cheng, X., Ding, J., & Yu, D. (2021). Evaluation and error analysis of official tropical cyclone intensity forecasts during 2005–2018 for the Western North Pacific. *Journal of the Meteorological Society of Japan. Series II*, *99*(1), 139–163. https://doi.org/10.2151/jmsj.2021-008

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., & Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *Journal of Geophysical Research*, *113*(D13), D13103. https://doi.org/10.1029/2008JD009944

Islam, T., Srivastava, P. K., Rico-Ramirez, M. A., Dai, Q., Gupta, M., & Singh, S. K. (2015). Tracking a tropical cyclone through WRF–ARW simulation and sensitivity of model physics. *Natural Hazards*, *76*(3), 1473–1495. https://doi.org/10.1007/s11069-014-1494-8

Janjic, Z. I. (1994). The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Monthly Weather Review*, *122*(5), 927–945. https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2

Janjic, Z. I. (1996). The surface layer in the NCEP Eta Model. In *Eleventh conference on numerical weather prediction*. (pp. 354–355). American Meteorological Society. 19-23 August.

Jimenez, P. A., Dudhia, J., Gonzalez-Rouco, J. F., Navarro, J., Montavez, J. P., & García-Bustamante, E. (2012). A revised scheme for the WRF surface layer formulation. *Monthly Weather Review*, *140*(3), 898–918. https://doi.org/10.1175/MWR-D-11-00056.1

Judt, F., Chen, S. S., & Berner, J. (2016). Predictability of tropical cyclone intensity: Scale-dependent forecast error growth in high-resolution stochastic kinetic-energy backscatter ensembles. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 43–57. https://doi.org/10.1002/qj.2626

Kain, J. S. (2004). The Kain-Fritsch convective parameterization: An update. *Journal of Applied Meteorology*, *43*(1), 170–181. https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2

Knapp, K. R., Diamond, H. J., Kossin, J. P., Kruk, M. C., & Schreck, C. J. (2018). International Best Track Archive for Climate Stewardship (IBTrACS) Project, version 4. [Dataset]. NOAA National Centers for Environmental Information. https://doi.org/10.25921/82ty-9e16

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., & Neumann, C. J. (2010). The International Best Track Archive for Climate Stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, *91*(3), 363–376. https://doi.org/10.1175/2009BAMS2755.1

Kwon, Y. C., & Hong, S. Y. (2017). A mass-flux cumulus parameterization scheme across gray-zone resolutions. *Monthly Weather Review*, *145*(2), 583–598. https://doi.org/10.1175/MWR-D-16-0034.1

Lei, L., Ge, Y., Tan, Z., & Bao, X. (2020). An evaluation and improvement of tropical cyclone prediction in the Western North Pacific basin from global ensemble forecasts. *Science China Earth Sciences*, *63*(1), 12–26. https://doi.org/10.1007/s11430-019-9480-8

Lei, L., Ge, Y., Tan, Z. M., Zhang, Y., Chu, K., Qiu, X., & Qian, Q. (2022). Evaluation of a regional ensemble data assimilation system for typhoon prediction. *Advances in Atmospheric Sciences*, *39*(11), 1816–1832. https://doi.org/10.1007/s00376-022-1444-4

Li, G., Curcic, M., Iskandarani, M., Chen, S. S., & Knio, O. M. (2019). Uncertainty propagation in coupled atmosphere-wave-ocean prediction system: A study of Hurricane Earl (2010). *Monthly Weather Review*, *147*(1), 221–245. https://doi.org/10.1175/MWR-D-17-0371.1

Li, X., & Pu, Z. (2008). Sensitivity of numerical simulation of early rapid intensification of Hurricane Emily (2005) to cloud microphysical and planetary boundary layer parameterizations. *Monthly Weather Review*, *136*(12), 4819–4838. https://doi.org/10.1175/2008MWR2366.1

Li, X., & Pu, Z. (2021). Vertical eddy diffusivity parameterization based on a large-eddy simulation and its impact on prediction of hurricane landfall. *Geophysical Research Letters*, *48*(2), 1–9. https://doi.org/10.1029/2020GL090703

Lim, K. S. S., & Hong, S. Y. (2010). Development of an effective double-moment cloud microphysics scheme with prognostic Cloud Condensation Nuclei (CCN) for weather and climate models. *Monthly Weather Review*, *138*(5), 1587–1612. https://doi.org/10.1175/2009MWR2968.1

Lin, Y. L., Farley, R. D., & Orville, H. D. (1983). Bulk parameterization of the snow field in a cloud model. *Journal of Climate and Applied Meteorology*, *22*(6), 1065–1092. https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

Ma, L. M., & Tan, Z. M. (2009). Improving the behavior of the cumulus parameterization for tropical cyclone prediction: Convection trigger. *Atmospheric Research*, *92*(2), 190–211. https://doi.org/10.1016/j.atmosres.2008.09.022

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Maw, K. W., & Min, J. (2017). Impacts of microphysics schemes and topography on the prediction of the heavy rainfall in Western Myanmar associated with tropical cyclone ROANU (2016). *Advances in Meteorology*, *2017*, 1–22. https://doi.org/10.1155/2017/3252503

Melhauser, C., Zhang, F., Weng, Y., Jin, Y., Jin, H., & Zhao, Q. (2017). A multiple-model convection-permitting ensemble examination of the probabilistic prediction of tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Weather and Forecasting*, *32*(2), 665–688. https://doi.org/10.1175/WAF-D-16-0082.1

Mohan, P. R., Srinivas, C. V., Yesubabu, V., Baskaran, R., & Venkatraman, B. (2019). Tropical cyclone simulations over Bay of Bengal with ARW model: Sensitivity to cloud microphysics schemes. *Atmospheric Research*, *230*(March), 104651. https://doi.org/10.1016/j.atmosres.2019.104651

Murphy, J. M. (1988). The impact of ensemble forecasts on predictability. *Quarterly Journal of the Royal Meteorological Society*, *114*(480), 463–493. https://doi.org/10.1002/qj.49711448010

Nakanishi, M., & Niino, H. (2006). An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*, *119*(2), 397–407. https://doi.org/10.1007/s10546-005-9030-8

National Center for Atmospheric Research. (2019). The NCAR Command Language (version 6.6.2). [Software]. https://doi.org/10.5065/D6WD3XH5

National Center for Atmospheric Research. (2018). Weather Research & Forecasting Model (version 4.0). [Software]. https://doi.org/10.5065/D6MK6B4K

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2023, 4, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003381 by Nanjing University, Wiley Online Library on [06/04/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

National Centers for Environmental Prediction. (2015). NCEP GDAS/FNL 0.25 degree global tropospheric analyses and forecast grids. [Dataset]. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. https://doi.org/10.5065/D65Q4T4Z

Nekkali, Y. S., Osuri, K. K., & Das, A. K. (2022). Numerical modeling of tropical cyclone size over the Bay of Bengal: Influence of microphysical processes and horizontal resolution. *Meteorology and Atmospheric Physics*, *134*(4), 1–13. https://doi.org/10.1007/s00703-022-00915-4

Osuri, K. K., Mohanty, U. C., Routray, A., Kulkarni, M. A., & Mohapatra, M. (2012). Customization of WRF-ARW model with physical parameterization schemes for the simulation of tropical cyclones over North Indian Ocean. *Natural Hazards*, *63*(3), 1337–1359. https://doi.org/10.1007/s11069-011-9862-0

Pielke, R. A., Gratz, J., Landsea, C. W., Collins, D., Saunders, M. A., & Musulin, R. (2008). Normalized hurricane damage in the United States: 1900–2005. *Natural Hazards Review*, *9*(1), 29–42. https://doi.org/10.1061/(asce)1527-6988

Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12–20. https://doi.org/10.1016/j.apenergy.2011.11.004

Pleim, J. E. (2007). A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *Journal of Applied Meteorology and Climatology*, *46*(9), 1383–1395. https://doi.org/10.1175/JAM2539.1

Pollard, R. T., Rhines, P. B., & Thompson, R. O. (1973). The deepening of the wind-mixed layer. *Geophysical Fluid Dynamics*, *4*(4), 381–404. https://doi.org/10.1080/03091927208236105

Raju, P. V. S., Potty, J., & Mohanty, U. C. (2011). Sensitivity of physical parameterizations on prediction of tropical cyclone Nargis over the Bay of Bengal using WRF model. *Meteorology and Atmospheric Physics*, *113*(3), 125–137. https://doi.org/10.1007/s00703-011-0151-y

Rogers, E., Black, T., Ferrier, B., Lin, Y., Parrish, D., & Di Mego, G. (2001). Changes to the NCEP Meso Eta Analysis and Forecast System: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3DVAR analysis. *NWS Technical Procedures Bulletin*, *488*, 15.

Rogers, R. (2010). Convective-scale structure and evolution during a high-resolution simulation of tropical cyclone rapid intensification. *Journal of the Atmospheric Sciences*, *67*(1), 44–70. https://doi.org/10.1175/2009JAS3122.1

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*(6), 1653–1660. https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2

Scheuerer, M., & Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, *143*(4), 1321–1334. https://doi.org/10.1175/MWR-D-14-00269.1

Shi, X., & Wang, Y. (2022). Impacts of cumulus convection and turbulence parameterizations on the convection-permitting simulation of typhoon precipitation. *Monthly Weather Review*, *150*(11), 2977–2997. https://doi.org/10.1175/MWR-D-22-0057.1

Shutts, G. (2005). A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, *131*(612), 3079–3102. https://doi.org/10.1256/qj.04.106

Simpson, R. H., & Saffir, H. (1974). The hurricane disaster potential scale. *Weatherwise*, *27*(8), 169. https://doi.org/10.1080/00431672.1974.9931702

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., et al. (2019). A description of the advanced research WRF model version 4 (No. NCAR/TN-556+STR). https://doi.org/10.5065/1dfh-6p97

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., & Du, H. (2015). Towards improving the framework for probabi-listic forecast evaluation. *Climatic Change*, *132*(1), 31–45. https://doi.org/10.1007/s10584-015-1430-2

Srinivas, C. V., Bhaskar Rao, D. V., Yesubabu, V., Baskaran, R., & Venkatraman, B. (2013). Tropical cyclone predictions over the Bay of Bengal using the high-resolution advanced research weather research and forecasting (ARW) model. *Quarterly Journal of the Royal Meteorological Society*, *139*(676), 1810–1825. https://doi.org/10.1002/qj.2064

Tao, W. K., Shi, J. J., Chen, S. S., Lang, S., Lin, P. L., Hong, S. Y., et al. (2011). The impact of microphysical schemes on hurricane intensity and track. *Asia-Pacific Journal of Atmospheric Sciences*, *47*(1), 1–16. https://doi.org/10.1007/s13143-011-1001-z

Tewari, M., Chen, F., Wang, W., Dudhia, J., LeMone, M., Mitchell, K., et al. (2004). Implementation and verification of the unified NOAH land surface model in the WRF model. In *Proceedings of the 20th conference on weather analysis and forecasting/16th conference on numerical weather prediction* (Vol. 14).

Thatcher, L., & Pu, Z. (2014). Characteristics of tropical cyclone genesis forecasts and underdispersion in high-resolution ensemble forecasting with a stochastic kinetic energy backscatter scheme. *Tropical Cyclone Research and Review*, *3*(4), 203–217. https://doi.org/10.6057/2014TCRR04.01

Thompson, G., Field, P. R., Rasmussen, R. M., & Hall, W. D. (2008). Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Monthly Weather Review*, *136*(12), 5095–5115. https://doi.org/10.1175/2008MWR2387.1

Tiedtke, M. (1989). A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, *117*(8), 1779–1800. https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2

Torn, R. D. (2016). Evaluation of atmosphere and ocean initial condition uncertainty and stochastic exchange coefficients on ensemble tropical cyclone intensity forecasts. *Monthly Weather Review*, *144*(9), 3487–3506. https://doi.org/10.1175/MWR-D-16-0108.1

Tracton, M. S., & Kalnay, E. (1993). Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather and Forecasting*, *8*(3), 379–398. https://doi.org/10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2

Whitaker, J. S., & Louche, A. F. (1998). The relationship between ensemble spread and ensemble mean skill. *Monthly Weather Review*, *126*(12), 3292–3302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Elsevier. https://doi.org/10.1016/C2017-0-03921-6

Wu, L., Zhang, H., Feng, T., & Tang, Y. (2020). Tropical cyclones and multiscale climate variability: The active Western North Pacific typhoon season of 2018. *Science China Earth Sciences*, *63*(1), 1–11. https://doi.org/10.1007/s11430-019-9474-4

Zhang, C., Wang, Y., & Hamilton, K. (2011). Improved representation of boundary layer clouds over the southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme. *Monthly Weather Review*, *139*(11), 3489–3513. https://doi.org/10.1175/MWR-D-10-05091.1

Zhang, J. A., Nolan, D. S., Rogers, R. F., & Tallapragada, V. (2015). Evaluating the impact of improvements in the boundary layer parameterization on hurricane intensity and structure forecasts in HWRF. *Monthly Weather Review*, *143*(8), 3136–3155. https://doi.org/10.1175/MWR-D-14-00339.1

Zhang, X. (2018). A GRAPES-based mesoscale ensemble prediction system for tropical cyclone forecasting: Configuration and performance. *Quarterly Journal of the Royal Meteorological Society*, *144*(711), 478–498. https://doi.org/10.1002/qj.3220