Precipitation Forecast Experiments Using the Weather Research and Forecasting (WRF) Model at Gray-Zone Resolutions

JI-YOUNG HAN AND SONG-YOU HONG

Korea Institute of Atmospheric Prediction Systems, Seoul, South Korea

(Manuscript received 12 February 2018, in final form 29 September 2018)

ABSTRACT

In the Weather Research and Forecasting (WRF) community, a standard model setup at a grid size smaller than 5 km excludes cumulus parameterization (CP), although it is unclear how to determine a cutoff grid size where convection permitting can be assumed adequate. Also, efforts to improve high-resolution precipitation forecasts in the range of 1–10 km (the so-called gray zone for parameterized precipitation physics) have recently been made. In this study, we attempt to statistically evaluate the skill of a gray-zone CP with a focus on the quantitative precipitation forecast (QPF) in the summertime. A WRF Model simulation with the gray-zone simplified Arakawa–Schubert (GSAS) CP at 3-km spatial resolution over East Asia is evaluated for the summer of 2013 and compared with the results from a conventional setup without CP. A statistical evaluation of the 3-month simulations shows that the GSAS demonstrates a typical distribution of the QPF skill, with high (low) scores and bias in the light (heavy) precipitation category. The WRF without CP seriously suppresses light precipitation events, but its skill for heavier categories is better. Meanwhile, a new set of precipitation data, which is simply averaged precipitation from the two simulations, demonstrates the best skill in all precipitation categories. Bearing in mind that high-resolution QPF requires essential challenges in model components, along with complexity in precipitating convection mechanisms over geographically different regions, this proposed method can serve as an alternative for improving the QPF for practical usage.

1. Introduction

For grid spacing smaller than 5 km, most regional atmospheric models including the Weather Research and Forecasting (WRF) Model have typically been setup without using cumulus parameterizations (CPs). The underlying assumption is that this grid size is nominally sufficient to resolve mesoscale convective systems explicitly and represent the development and evolution of deep moist convection (e.g., Weisman et al. 1997, 2008; Kain et al. 2008), in which the precipitation process can be represented by only the grid-resolvable precipitation physics through microphysics schemes (MPSs). With this hypothesis, in order to improve the predictability of precipitating convection using the WRF Model, there have been continuous efforts to increase the complexity in MPSs by adding the number concentration of hydrometeors (e.g., Milbrandt and Yau 2005; Morrison et al. 2005; Thompson et al. 2008; Lim and Hong 2010) or by introducing spectral bin microphysics (e.g., Lynn et al. 2005).

At a resolution whose grid size is smaller than 5 km, it has been reported that severe meteorological phenomena are realistically reproduced in terms of mesoscale characteristics associated with accompanying precipitation (e.g., Bryan and Morrison 2012; McMillen and Steenburgh 2015; Jee and Kim 2017). But, it is concluded by Bryan et al. (2003) that horizontal grid spacing of order 100 m is required to adequately resolve deep moist convection. Some studies identified the sensitivity of simulation results to MPSs, but there is no specific MPS that outperforms the others (Clark et al. 2012), though a complex MPS tends to outperform a simpler one (e.g., McCumber et al. 1991; Reisner et al. 1998; Putnam et al. 2017). The reason lies in the fact that precipitating convection is not fully resolved by grid-scale forcing with grid spacing of 5 km or less. This issue, the so-called gray zone, has been raised in the modeling community (Hong and Dudhia 2012). Problems appear in physics as numerical weather prediction (NWP) models go to finer
scales where there are “gray zones” in which the explicit model dynamics are partly capable of resolving features that are parameterized at coarser scales. At gray-zone resolutions, there tends to be only a partial representation of convective processes without clear separation between grid-scale and subgrid-scale moist processes (Arakawa 2004).

To address scale awareness for improving high-resolution precipitation forecasts in the range of 1–10 km (the gray zone for parameterized precipitation physics), several modifications have been introduced to conventional CPs (e.g., Arakawa and Wu 2013; Grell and Freitas 2014; Zheng et al. 2016; Kwon and Hong 2017). A major point of gray-zone CPs is to suppress the strength of convective updrafts at higher resolutions by adding a resolution-dependent parameter to conventional CPs. All these studies are efforts to represent a smooth transition from CP-induced to MPS-induced precipitation at gray-zone scales. Although the results with gray-zone CPs are promising to quantitative precipitation forecasts (QPFs) by suppressing excessive precipitation in extremely high-precipitation categories, evaluations in the previous literature are limited to a few cases of precipitation events.

The purpose of this study is to statistically evaluate a gray-zone CP with a focus on QPFs during the summertime. A WRF Model simulation with the gray-zone simplified Arakawa–Schubert (GSAS) CP (Kwon and Hong 2017) at a 3-km spatial resolution over East Asia is evaluated for the summer of 2013. The statistical skill scores from 92 cases of 48-h forecast experiments are compared with the values from a conventional setup without CP. Recognizing the advantages of each model setup, a simple averaged output of the two experiments with and without CP is also evaluated. The following section describes the experimental setup and the results of the study are reported in section 3, after which we end this paper with concluding remarks.

2. Model and experimental setup

a. Model description

The Advanced Research version of the WRF Model (version 3.7.1; Skamarock et al. 2008) is employed in this study. The physics parameterization schemes used are the WRF single-moment 5-class (WSM5) (Hong et al. 2004), the Rapid Radiative Transfer Model for general circulation models (RRTMG; Iacono et al. 2008), the unified Noah land surface model (Chen and Dudhia 2001), the Yonsei University (YSU) planetary boundary layer (PBL) scheme (Hong et al. 2006), and the Global/Regional Integrated Model System (GRIMs) shallow convection scheme (Hong et al. 2012).

An updated version of the simplified Arakawa–Schubert (SAS) deep convection scheme is used as a CP. The CP used in this study is a mass-flux scheme based on the SAS convection scheme (Pan and Wu 1995; Han and Pan 2011), but with several modifications to the subgrid-scale cloud properties that has been revised by the Korea Institute of Atmospheric Prediction Systems (KIAPS). The modified autoconversion from cloud condensate to convective precipitation is one of them (Han et al. 2016). The CP version also has a function of scale awareness that enables a mass-flux CP to work seamlessly at various grid sizes across gray-zone resolutions proposed by Kwon and Hong (2017). The dependence on the convective updraft fraction is taken into account in the calculations of the threshold for the trigger condition, the cloud-base mass flux, and the amount of condensed water that is detrained into grid-scale condensate. These modifications result in a suppressed trigger of convection, reduced strength of convection, and increased detrainment of convective cloud water at higher resolutions. Kwon and Hong (2017) evaluated the scale-aware CP for a heavy rainfall event over Korea at gray-zone resolutions and showed that it reasonably simulates the precipitation pattern, reproducing a well-organized single rainfall core as in the observations.

b. Experimental design

Three computational domains with one-way nesting are used (Fig. 1). A 3-km horizontal grid-spacing domain covering the Korean Peninsula (domain 3, with $355 \times 352$ grid points) is nested within a 9-km grid-spacing
domain (domain 2, with 259 × 223 grid points), which in turn is nested within a 27-km grid-spacing domain covering the East Asian region (domain 1, with 178 × 150 grid points). All domains have 51 vertical layers and a model top at 50 hPa.

For statistical evaluation of the precipitation forecast skill, 48-h forecasts starting at every 0000 UTC during the East Asian summer monsoon period [July–September (JAS)] of 2013 are performed. The National Centers for Environmental Prediction (NCEP) Final Analysis (FNL) dataset with a horizontal resolution of 1° at 6-h intervals is used as the initial and boundary conditions. Sea surface temperature (SST) forcing is provided by the NCEP daily real-time global SST (RTG_SST) analysis (Thiébaux et al. 2003) with a horizontal resolution of 0.5°. To compare the features of high-resolution simulations of precipitation without and with CP adopting scale awareness and assess their skill for the forecasts of precipitation, two experiments are performed: one without CP only for a 3-km domain (CP-off experiment), which is a typical experimental setup at this high resolution where it is assumed that cumulus convection is explicitly resolved, and the other with CP for all domains (CP-on experiment). Then, a new precipitation forecast dataset, which is an average of the simulation results with and without CP, is evaluated to determine its skill for the QPF.

c. Data and methodology

Two observational precipitation datasets are used to statistically evaluate the precipitation forecast skill. One is the Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis (TMPA) 3B42 (version 7) 3-hourly precipitation product (Huffman et al. 2007), which covers from 50°S to 50°N with a horizontal resolution of 0.25°. The other is automatic weather station (AWS) rain gauge observations over South Korea operated by the Korea Meteorological Administration (KMA). Hourly surface gauge data from 617 stations are used in this study, and the average separation between stations is approximately 13 km. The precipitation forecast data are interpolated in space to 0.25° grid resolution and to the AWS observation points for verification against the TMPA and rain gauge observations, respectively, using an inverse distance squared algorithm within the NCAR Command Language (NCL: https://www.ncl.ucar.edu). Although the TMPA data have a relatively lower spatial resolution compared to the AWS data, they are used to validate rainfall forecasts over the innermost domain including the Korean Peninsula and adjacent oceans (see Fig. 1) because the AWS rain gauge observations are only available over the land area of South Korea.

For quantitative evaluation of the precipitation forecast skill, standard verification measures of the equitable threat score (ETS) and bias score that are most widely used to verify the precipitation forecasts are calculated at different precipitation thresholds. The ETS measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for correct predictions due to random chance (Ebert et al. 2003), and is defined by

\[ ETS = \frac{H - H_{\text{random}}}{H + M + F - H_{\text{random}}}, \]

where

\[ H_{\text{random}} = \frac{(H + F)(H + M)}{N}. \]

Here, \( H \) is the number of correct forecasts of events occurring (hits), \( M \) is the number of events observed but not forecasted (misses), \( F \) is the number of events forecasted but not observed (false alarms), \( Z \) is the number of correct forecasts of events not occurring (correct negatives), \( N \) (= \( H + M + F + Z \)) is the sample size, and \( H_{\text{random}} \) is the number of hits that could be expected due purely to random chance. The bias score is the ratio of the number of forecasted to observed events, which is given by

\[ \text{bias score} = \frac{H + F}{H + M}, \]

and indicates whether the model forecasts have a tendency to underestimate (bias score less than 1) or overestimate (bias score greater than 1) events. Note that a higher ETS and a bias score close to 1 indicate better forecast performance.

Another score used in this study is the symmetric extreme dependency score (SEDS; Hogan et al. 2009) defined by

\[ SEDS = \frac{\ln[(H + M)/N] + \ln[(H + F)/N]}{\ln(H/N)} - 1. \]

Unlike the ETS, which has an intrinsic dependence on the observed frequency of occurrence and tends to the trivial noninformative limit for increasing precipitation thresholds (Stephenson et al. 2008), the SEDS provides robust verification for rare extreme events.

3. Results and discussion

To statistically assess the short-range precipitation forecast skill of high-resolution simulations with GSAS, the ETS and bias score for precipitation at 3-km resolution for forecast day 2 during the period JAS 2013 are...
calculated for the CP-on experiment (red) and compared to those from the CP-off experiment (blue), as shown in Fig. 2. For light precipitation with thresholds below 5 mm day$^{-1}$, the ETS against AWS rain gauge observations over South Korea is considerably higher in the CP-on experiment than in the CP-off experiment (Fig. 2a). On the contrary, for heavy precipitation categories, a lower ETS is observed in the CP-on experiment. A consistent result, that is, a better score for light (heavy) precipitation in the CP-on (CP-off) experiment, is also obtained from the statistical skill of the ETS calculated against the TMPA observation over the 3-km domain (Fig. 2c). To see if there is a statistically significant difference in the skill scores, we conducted a hypothesis test proposed by Hamill (1999), which is based on the remapping technique. The results show that the differences in the bias score between the CP-off and CP-on experiments are statistically significant at the 95% confidence level for all precipitation thresholds, and the differences in the ETS are statistically significant for light precipitation below 5 mm day$^{-1}$ and heavy precipitation above 20 mm day$^{-1}$ (not shown). Figures 2b and 2d indicate that a systematic bias of underestimation of rainfall for light-precipitation categories below 5 mm day$^{-1}$ is pronounced in the CP-off experiment, especially against the TMPA observations, whereas the CP-on experiment overestimates light precipitation with thresholds of 0.5 and 1 mm day$^{-1}$ against AWS rain gauge observations and a threshold of 0.5 mm day$^{-1}$ against the TMPA observation. Meanwhile, both experiments systematically
underestimate rainfall for heavier precipitation categories, and it is significantly pronounced in the CP-on experiment. A comparison of the skill scores for forecast day 1 shows a similar behavior with that for forecast day 2; that is, a more pronounced bias of a lack of light (heavy) rain in the CP-off (CP-on) experiment and a lower ETS for light (heavy) precipitation categories (not shown).

Recognizing the systematic biases in high-resolution precipitation forecasts observed in the CP-off and CP-on experiments, we also evaluate the statistical performance of a new set of precipitation forecasts that is the simple averaging of precipitation data from the two experiments (hereafter called the AVG experiment; green in Fig. 2). The AVG experiment reveals an overall improvement of the precipitation forecast skill compared to both the CP-off and CP-on experiments. The ETS is higher than both experiments for all precipitation categories against AWS rain gauge observations over South Korea (Fig. 2a) and for all categories except a threshold of 5 mm day$^{-1}$ against the TMPA observation over the 3-km domain (Fig. 2c). Light precipitation tends to be overestimated in the AVG experiment, with a smaller bias than the CP-on experiment (Figs. 2b,d). A systematic bias of a lack of heavier precipitation events is also observed in the AVG experiment, but the bias is significantly reduced compared to the CP-on experiment.

To compare the bias patterns of the precipitation field from each experiment, a representative forecast of 24-h accumulated precipitation amount at 0000 UTC 4 July 2013 for the 3-km domain is provided in Fig. 3, in which...
the model results are regridded to the resolution of the TMPA dataset. The overall pattern of observed precipitation, an east–west-oriented rainband across the southern coast of the Korean Peninsula (Fig. 3a), is reproduced in both the CP-off and CP-on experiments. In the CP-off experiment, however, light to heavy precipitation over the West Sea and south coast of Korea is significantly underestimated, while heavy precipitation over the Korean Strait between Korea and Japan is remarkably enhanced compared to the observations (Fig. 3b). In contrast, the CP-on experiment shows excessive areal coverage for light precipitation, which is attributed to convective precipitation, and weakened precipitation activity (Fig. 3c). Figure 3d indicates that the AVG experiment exhibits a similar spatial coverage of light precipitation as observed in the CP-on experiment, but the amount is reduced. On the other hand, the pattern for heavy precipitation is similar to that in the CP-off experiment, but with reduced intensity.

Figure 4 shows the probability density function (PDF) of precipitation on a logarithmic scale during the period of July 2013 averaged over the 3-km domain. The CP-off experiment overestimates the fraction of grid points with almost no precipitation and extremely heavy rainfall exceeding 70 mm day$^{-1}$, but underestimates light-to-heavy precipitation with intensity between them in comparison with the observations. In the CP-on experiment, on the other hand, the area fractions with almost no precipitation and rainfall intensity greater than 10 mm day$^{-1}$ are underestimated, whereas the precipitation between them is overestimated. The PDF of precipitation from the AVG experiment shows that the area fraction with almost no rain becomes similar to that of the TMPA observations. Although the AVG experiment tends to overestimate areas of light precipitation, the bias is smaller than the CP-on experiment. Also, the biases for rainfall in heavier precipitation categories are the smallest in the AVG experiment.

It is found that because a lack of light rain in the CP-off experiment causes the large increase in misses and the large decrease in hits for light precipitation, the skill score for light precipitation is much lower in the CP-off experiment than in the CP-on experiment. On the other hand, in the CP-on experiment, the increase in misses and the decrease in hits for heavy rainfall resulting from a severe deficiency of heavy precipitation lead to a lower score for the heavy precipitation categories compared with the CP-off experiment. Meanwhile, the AVG experiment takes advantage of light (heavy) precipitation from the CP-on (CP-off) experiment. Better forecast skill for light precipitation compared to the CP-on experiment is attributable to the reduction in false alarms by alleviating the bias of excessive light rain observed in the CP-on experiment. Similarly, the decreased number of false alarms for heavy precipitation compared to the CP-off experiment results in an improvement to the forecast skill for heavy rainfall. The forecast performance of the AVG experiment also shows the improvement over the others in terms of the root-mean-square error (RMSE) and pattern correlation (PC) of precipitation (Table 1). The differences in the RMSEs between all experiments are found to be statistically significant at the 95% confidence level, while there is a statistically significant difference in the PC only between the CP-on and AVG experiments.

Because comparisons of precipitation skill scores from competing forecasts can cause misleading conclusions if they have different biases (Mason 1989; Hamill 1999), the scores for percentile thresholds are calculated to remove the effects of the bias in rainfall amounts on the ETS. Figures 5a–c display the bias score, ETS, and SEDS for different percentile thresholds from the
FIG. 5. (a) Bias score, (b) ETS, and (c) SEDS for different percentile thresholds for precipitation forecasts from the CP-off (blue) and CP-on (red) experiments for forecast day 2 during the period JAS 2013 over the 3-km domain against the TMPA observations. (d)–(f) As in (a)–(c), but for the scores after the precipitation thresholds of the CP-on experiment are adjusted to achieve a bias score similar to that of the CP-off experiment. Confidence intervals referenced to the CP-on experiment indicate the 2.5 and 97.5 percentile values of the resampled distribution.
CP-off and CP-on experiments. The confidence intervals referenced to the CP-on experiment indicate the 2.5 and 97.5 percentile values of the resampled distribution. Figure 5b shows that the CP-on experiment has a higher ETS than the CP-off experiment for percentile thresholds up to 95% but a lower ETS for the 98% threshold that can be considered extreme. The difference is statistically significant only for the 98% threshold. Note that the differences in scores for lower percentile thresholds become less statistically significant due to their smaller sample size. Because the ETS has an intrinsic dependence on the observed frequency of occurrence and tends to the trivial noninformative limit for increasing precipitation thresholds (Stephenson et al. 2008), the SEDS, which provides robust verification for rare extreme events, is also calculated. The SEDS indicates that the CP-on experiment consistently outperforms the CP-off experiment for thresholds below 95%, while the CP-off experiment is better for more extreme events (Fig. 5c).

The bias adjustment procedure proposed by Hamill (1999) is also applied in this study to assess the effects of the differences in bias scores between the experiments on the skill scores. The scores in Figs. 5d–f are obtained after adjusting the precipitation thresholds of the CP-on experiment to achieve a bias score similar to that of the CP-off experiment (Fig. 5d). The ETS after performing the bias adjustment is higher in the CP-on experiment than in the CP-off experiment for all precipitation thresholds considered in Fig. 2 (up to 25 mm day$^{-1}$) in a statistically significant way (Fig. 5e). Compared to the score calculated without accounting for the bias difference between the two experiments (Fig. 2c), the difference in the ETS for light precipitation below 5 mm day$^{-1}$ is reduced, and the CP-on experiment performs better even for precipitation thresholds of 10–25 mm day$^{-1}$. This indicates that the ETS for light (heavy) precipitation categories is inflated in the CP-on (CP-off) experiment due to its systematic bias of overestimation of light (heavy) precipitation. Meanwhile, the ETS and SEDS from the CP-off experiment are higher than those from the CP-on experiment for more extreme precipitation events with higher thresholds above 50 mm day$^{-1}$, which is consistent with the result with the percentile threshold. However, unlike that, the differences are not statistically significant at the 5% level.

The ETS and SEDS calculated using the percentile threshold and the fixed threshold, but after the bias adjustment for precipitation forecasts from the AVG experiment, are compared with those from the CP-off experiment in Fig. 6. The AVG experiment shows better performance than the CP-off experiment for all precipitation thresholds. The score differences in the ETS for the percentile threshold are statistically significant only for the 98% threshold (Fig. 6b), but at thresholds greater or equal to 90% for the SEDS (Fig. 6c). Meanwhile, statistically significant differences are found in the ETS and SEDS after the bias adjustment at thresholds up to 40 mm day$^{-1}$ and all thresholds, respectively (Figs. 6e,f).

A comparison between the CP-on and AVG experiments indicates that the AVG experiment performs better for all percentile thresholds, but the differences are statistically significant only for the 98% threshold (Figs. 7b,c). For the scores with the fixed threshold after performing the bias adjustment, on the other hand, no significant difference is found in the ETS (Fig. 7e), but a higher SEDS is observed from the AVG experiment for more extreme precipitation events above 30 mm day$^{-1}$ with statistically significant differences at thresholds greater or equal to 50 mm day$^{-1}$ (Fig. 7f).

4. Summary and conclusions

As one of the efforts to improve precipitation forecasts at gray-zone resolutions, CPs with scale-aware capability have recently been developed. In this study, the high-resolution precipitation forecast skill of a gray-zone CP is statistically evaluated in comparison with the typical model setup without CP, focusing on the QPF in the summertime. We performed 48-h forecast experiments with and without GSAS at 3-km resolution over the East Asia region during the summer monsoon period of 2013 using the WRF Model. From the statistical evaluation of 92 case runs, it is found that the experiment without CP suffers from a systematic underestimation of light precipitation, while a systematic bias of a lack of heavy precipitation is observed in the experiment with CP. As a result, by increasing misses and decreasing hits, the skill score for light precipitation is lower without CP, whereas for heavy precipitation a lower score is obtained with CP. Meanwhile, a new precipitation forecast dataset, which is the simple averaging of the output from the two experiments, exhibits an overall improvement in forecast skill for all precipitation categories by taking advantage of light (heavy) precipitation from the experiment with (without) CP.

It is evident from this study that further development of gray-zone CPs is required to overcome the systematic bias found in high-resolution precipitation forecasts. To improve high-resolution QPFs, continuous efforts should be made to develop the model components. However, there are many difficulties, for example, the diversity of the mechanisms responsible for precipitating convection occurring in geographically different areas. Considering this, our method, which can be considered as a kind of multiphysics ensemble forecast with two ensemble
FIG. 6. As in Fig. 5, but for the CP-off (blue) and AVG (green) experiments. The precipitation thresholds of the AVG experiment are adjusted to achieve a bias score similar to that of the CP-off experiment in (d)–(f).
FIG. 7. As in Fig. 5, but for the CP-on (red) and AVG (green) experiments. The precipitation thresholds of the CP-on experiment are adjusted to achieve a bias score similar to that of the AVG experiment in (d)–(f).
members using different model physics configurations, can be proposed as an alternative for improving high-resolution QPFs for practical use. A further improvement in the precipitation forecast skill can be achieved by statistically postprocessing the output from the two experiments, for example, using artificial neural networks (Lauret et al. 2014), and optimally combining them. This will be investigated in future work.

Acknowledgments. The authors are grateful to the editor and three anonymous reviewers for providing valuable comments on this work. This work has been carried out through an R&D project on the development of the global numerical weather prediction systems of the KIAPS funded by the KMA.

REFERENCES


