The More Rain, the Better the Model Performs—The Dependency of Quantitative Precipitation Forecast Skill on Rainfall Amount for Typhoons in Taiwan

CHUNG-CHIEH WANG

Department of Earth Sciences, National Taiwan Normal University, Taipei, Taiwan

(Manuscript received 27 April 2014, in final form 18 November 2014)

ABSTRACT

A strong dependency of model performance in quantitative precipitation forecasts (QPFs) as measured by scores such as the threat score (TS) on rainfall amount (i.e., the better the model performs when there is more rain), is demonstrated through real-time forecasts by the 2.5-km Cloud-Resolving Storm Simulator (CReSS) for 15 typhoons in Taiwan in 2010–12. Implied simply from the positive correlation between rain-area sizes and scores, this dependency is expected to exist in all regions, models, and rainfall regimes, while for typhoon QPFs in Taiwan it is also attributed to the model's capability to properly handle (within 72 h) the processes leading to more rain, which are largely controlled by the typhoon's track, size, structure, and environment, and the island's topography. Because of this dependency, the performance of model QPFs for extreme events can be assessed accurately only when forecasts targeted for periods of comparable rainfall magnitude are included for averaging. For the most-rainy 24 h of the top-5 typhoons, the 0–24-h QPFs by CReSS have mean TS of 0.67, 0.67, 0.58, 0.51, and 0.32 at thresholds of 25, 50, 130, 200, and 350 mm, and 0.64, 0.57, 0.37, 0.33, and 0.22 from 48–72-h QPFs, respectively, suggesting superior performance even 2–2.5 days in advance. These scores are strikingly high, precisely because Taiwan can receive extreme rainfall from typhoons. For smaller (nonhazardous) events, the mean scores are progressively lower, but also unimportant and less representative statistically. Therefore, it is inappropriate to average scores over multiple forecasts as those for less-rainy periods would contaminate the result for key periods. The implication for forecasters is also discussed.

1. Introduction and motivation of study

a. Background of research

As one of the most challenging tasks in modern weather prediction, quantitative precipitation forecasts (QPFs) are under constant and heavy demand around the world (e.g., Fritsch et al. 1998; Golding 2000; Fritsch and Carbone 2004; Cuo et al. 2011), especially for heavy to extreme rainfalls because of their hazardous nature. The verification and evaluation of QPFs, produced mostly by numerical models today, are thus crucially important as the results reflect the capabilities of the models and serve as the basis by which model improvement is judged. Among the various verification methods for QPFs, categorical measures based on the 2×2 contingency table (Wilks 1995; Jolliffe and Stephenson 2003) are the most widely used to date, although they are less suitable for rainfall caused by mesoscale systems (such as squall lines) and careful attention is needed to avoid the issue of "double penalty" on models that are capable of predicting the event but not at the correct location and/or time compared to those that cannot predict the event at all (e.g., Ebert and McBride 2000; Davis et al. 2006; Clark et al. 2007). In recent years, various new verification methods have been developed for model QPFs associated with mesoscale rainfall systems (e.g., Davis et al. 2006; Marzban and Sandgathe 2006; Wernli et al. 2008).

In Taiwan, heavy rainfalls brought by typhoons (mainly during July–October) and in the mei-yu season (May–June) are responsible for the majority of weather hazards, and categorical measures such as the threat score (TS) and bias score (BS) have been employed as the major tools for model QPF evaluation. Besides generally easy calculation and interpretation (section 2c), these measures are chosen also for physical and practical reasons. First, when the circulation of a tropical cyclone (TC) or the prevailing monsoon flow impinges on the steep topography of Taiwan (Fig. 1), the rainfall from forced uplift is phase locked to the windward slopes (i.e.,

Corresponding author address: Chung-Chieh Wang, Department of Earth Sciences, National Taiwan Normal University. No. 88, Sec. 4, Ting-Chou Rd., Taipei 11677, Taiwan.
E-mail: cwang@ntnu.edu.tw

DOI: 10.1175/MWR-D-14-00137.1

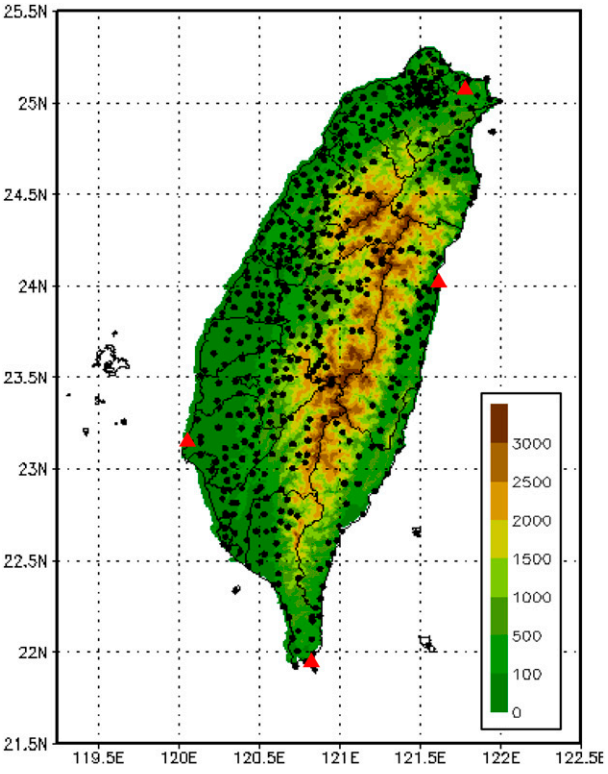

FIG. 1. The topography of Taiwan (m; color) and locations of rain gauges (dots) and Doppler radars (red triangles) operated by the Central Weather Bureau (CWB).

nonmoving) and is often a significant component of total rainfall (e.g., Chang et al. 1993; Cheung et al. 2008; Chang et al. 2013). Second, the observation from a dense network of rain gauges (Hsu 1998) needed for verification is available only over the island, and this limits the applicability of many other alternative methods. From time-averaged scores that value successful forecast of events (i.e., hits) such as the TS, it is long recognized that model QPFs depend on forecast range (the longer the poorer), as forecast errors typically grow with time (e.g., Olson et al. 1995; Ralph et al. 2010). On the other hand, one common characteristic of these measures in individual forecast is that the score decreases with increasing rainfall threshold, as the rain area becomes smaller and it is more difficult for the model to produce "hits" at the same fraction. This very linkage with rain-area size implies that the scores and the inferred model performance also depend on the overall rainfall amount of events, and the QPF is better when there is more rain (since the rain area reaching the same threshold is now larger). This property thus could have important implications on the forecasts for heavy to extreme rainfalls, and on how these forecasts should be perceived and used for hazard prevention and reduction. Although a possible link is previously noted (e.g., Johnson and Olsen 1998; Yuan et al. 2005), such a dependency has not been explored and understood adequately, and this is the objective of the present study.

b. *An example of a good forecast*

Here, an example of a high-resolution forecast at a horizontal grid spacing of 2.5 km, which can resolve the basic structure of convective clouds, for Typhoon (TY) Fanapi (2010) is presented to illustrate the quality of the prediction and the corresponding scores of QPFs. Details about the model, its forecasts, and skill measures will be described shortly in section 2. To verify the evolution of the event, the vertical maximum indicator (VMI) reflectivity composites from land-based radars in Taiwan (cf. Fig. 1) every 4 h from 1600 UTC 18 September to 0000 UTC 20 September 2010 are compared with the real-time forecast made by the author using the Cloud-Resolving Storm Simulator (CReSS; Tsuboki and Sakakibara 2002, 2007) at (starting from) 0000 UTC 17 September for the same period in Fig. 2. With its center labeled every 2 h, TY Fanapi approached Taiwan from almost due east since 18 September, but suddenly deviated southward just prior to landfall (near 0600 UTC 19 September) then moved westward across Taiwan and into the Taiwan Strait (Fig. 2, left column). The clouds associated with Fanapi were rather axisymmetric before landfall but changed to an asymmetric pattern with most deep convection in the southern half (Wang et al. 2013a). The forecast made at 0000 UTC 17 September (Fig. 2, right column), with the TC center determined using 15-min outputs near 2-km height to reduce the influence of topography, shows a storm evolution in excellent agreement with the radar observation. Both the track deflection before landfall and the shift into an asymmetric rainfall structure are well captured by this forecast made two days before 19 September. Careful comparison indicates that the forecast TC is about 3–4 h too early in crossing Taiwan and slightly too far north during departure (Fig. 2), but the track error is within ~100 km and relatively small for the third day.

The total daily (0000–2400 UTC) rainfall distributions from some 400 rain gauges over Taiwan (Fig. 1) during 17–19 September and the corresponding day-1 to day-3 QPFs from the forecast in Fig. 2 (i.e., at 0000 UTC 17 September) are presented in Figs. 3a–c and 3f–h, respectively, while the rest of Fig. 3 will be discussed later. Over Taiwan, the rainfall from Fanapi was received mostly on 19 September (Fig. 3c), while that on 17–18 September was much less, under the influence of the outer circulation (cf. Figs. 3a,b and 2). Because of the high realism of this forecast, the 24-h QPFs for day 1 (0–24 h), day 2 (24–48 h), and day 3 (48–72 h) agree closely with

the observations (Figs. 3a–c and 3f–h), especially for 19 September, although there is some underprediction for 18 September. On 19 September, two rainfall maxima appeared in southern Taiwan (triangles, Fig. 3c) and the one along the southwestern coast (>800 mm) seriously flooded Kaoshiung, the second largest city in Taiwan. Produced by intense rainbands south of the TC during departure, this coastal rainfall on 19 September was also captured by the model (Figs. 2l–p). Thus, overall the model made an excellent forecast for this particular typhoon, including the QPF, and provided valuable information to the forecasters and its users about what was coming at least two days in advance.

Now, let us examine the QPF performance of this decent forecast and plot the widely used TS and BS as functions of 24-h rainfall threshold from 0.05 to 1000 mm (Fig. 4). For any given threshold over an accumulation period (24 h here), the rain areas reaching this value in the observation and forecast can be identified, and TS is defined as their intersection divided by their union area ($0 \le \mathrm{TS} \le 1$, the higher the better), while BS is the ratio of rain-area sizes in the forecast to the observation and thus measures under- or overforecast [$0 \le \mathrm{BS} < \infty$, e.g., Wilks (1995)]. To be consistent and avoid confusion, the low, middle, and high thresholds in this paper refer to their corresponding one-third of the range (i.e., 0.05–50, 75–200, and 250–1000 mm, respectively) while the "extreme" events refer to those with 24-h rainfall ≥750 mm. Of course, not all TYs studied reached the high threshold (≥250 mm) and those that did were rainy only during certain period(s).

In Fig. 4a, the TS values for day 3 (19 September, blue) are very high and decrease only slowly from a perfect 1 at low thresholds ≤2.5 mm, to about 0.6–0.8 across the middle thresholds (75–200 mm), and remain above 0.6 at 250 mm and 0.4 at 350 mm, consistent with visual inspection (cf. Figs. 3c,h). Likewise, the BS for day 3 indicates superior performance with values near unity up to the 350-mm threshold (Fig. 4b). The TS curve for days 2 (red), on the other hand, drops much faster from about 0.8 and is much lower (<0.2) across thresholds of 25–130 mm, while the BS also reflects the underforecast for 18 September noted earlier (cf. Figs. 4a,b and 3b,g). The TS values for day 1 (17 September, black) are very low (<0.2 at 0.05 mm) and decrease quickly to zero and the BS also indicates underprediction (cf. Figs. 3a,f).

c. *Motivation of study*

The TS curves in Fig. 4a for the three days indicate very different skill level in model QPFs, even though they all come from a single forecast starting from 0000 UTC 17 September. Leading to the excellent performance for 19 September (cf. Figs. 2 and 3c,h), the 0–48-h forecast is verified to also have good quality in the characteristics of the TC (not shown) and yet the TSs are so much lower (except at 160 mm for day 2). In this single forecast, the TS shows no correlation with forecast range but depends strongly on the daily rainfall amount (event magnitude) in Taiwan (Figs. 3a–c,f–h, and 4a). A few questions immediately arise regarding this phenomenon of higher QPF scores when there is more rain. How common is this dependency in forecasts, and how strong is it on average? Does it come simply from the variation in rain-area size, or is the model in fact more skillful in predicting larger events and why? What are its importance and implications? What is the true skill of model QPFs, especially for extreme rainfalls that have high potential for hazards? How can we evaluate model QPFs more accurately and effectively? Does this phenomenon also occur in other models, other rainfall regimes, and other regions besides Taiwan? All the above questions motivated the author to carry out the present study, and will be addressed or shed light upon by using real-time CReSS forecasts in three seasons of 2010–12 for a total of 15 typhoons.

The remaining part of the paper is arranged as follows. The model, forecast data, and methodology are described in section 2. More examples using TS and BS and also other verification measures are presented in section 3. In section 4, the dependency of model QPF performance on rainfall amount is further demonstrated using all 15 typhoons, and the origin of this dependency is explored in section 5. While illustrated throughout this paper, the important implications of the phenomenon in model QPF verification/evaluation and to forecasters are further discussed in section 6, and the conclusions are given in section 7.

2. The CReSS model, data, and methodology

a. *The CReSS model and real-time forecasts*

The CReSS model (Tsuboki and Sakakibara 2002, 2007) is a cloud-resolving model suitable for studying convection, rainfall systems, severe weathers, and TCs alike (e.g., Liu et al. 2004; Maesaka et al. 2006; Wang et al. 2011, 2012; Akter and Tsuboki 2012). The model employs a terrain-following vertical coordinate with prognostic equations for momentum, pressure, potential temperature, and mixing ratios of water vapor and other hydrometeors. To properly simulate clouds, an explicit bulk cold rain scheme based on Lin et al. (1983), Cotton et al. (1986), Murakami (1990), Ikawa and Saito (1991), and Murakami et al. (1994) with a total of six species (vapor, cloud water, cloud ice, rain, snow, and graupel)

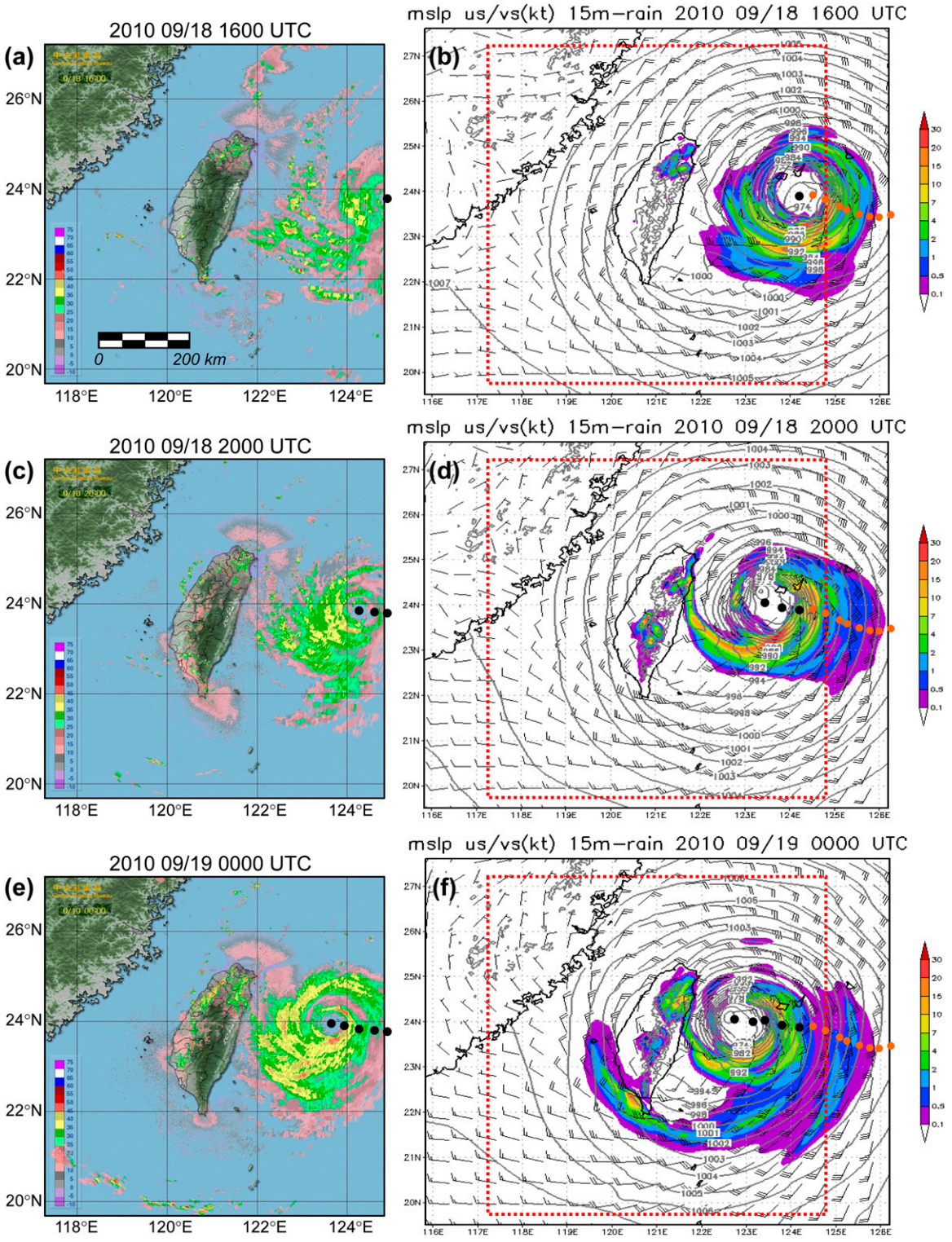

FIG. 2. (a) Radar reflectivity VMI composite (dBZ, scale at lower left, every 5 dBZ from −10 to 75 dBZ; plots provided by the CWB) and (b) CReSS forecast (initial time at 0000 UTC 17 Sep) of sea level pressure [hPa, over ocean only, every 1 (2) hPa above (below) 1000 hPa], surface wind (kt, barbs) at 10-m above-ground level (AGL), terrain-height contours at 1 and 2 km (gray lines, over land only), and 15-min rain (mm, color, scale to the right) valid at 1600 UTC 18 Sep 2010. (c),(d)–(q),(r) As in (a),(b), but every 4 h from 2000 UTC 18 Sep to 0000 UTC 20 Sep 2010. (left) The center of Fanapi is marked every 2 h (black dots), determined using CWB best track and radar composites every 30 min or (right) CReSS forecasts near 2-km height every 15 min (in orange before 1600 UTC 18 Sep 2010). The dotted red box shows the region of radar plots (19.7°–27.2°N, 117.3°–124.8°E).

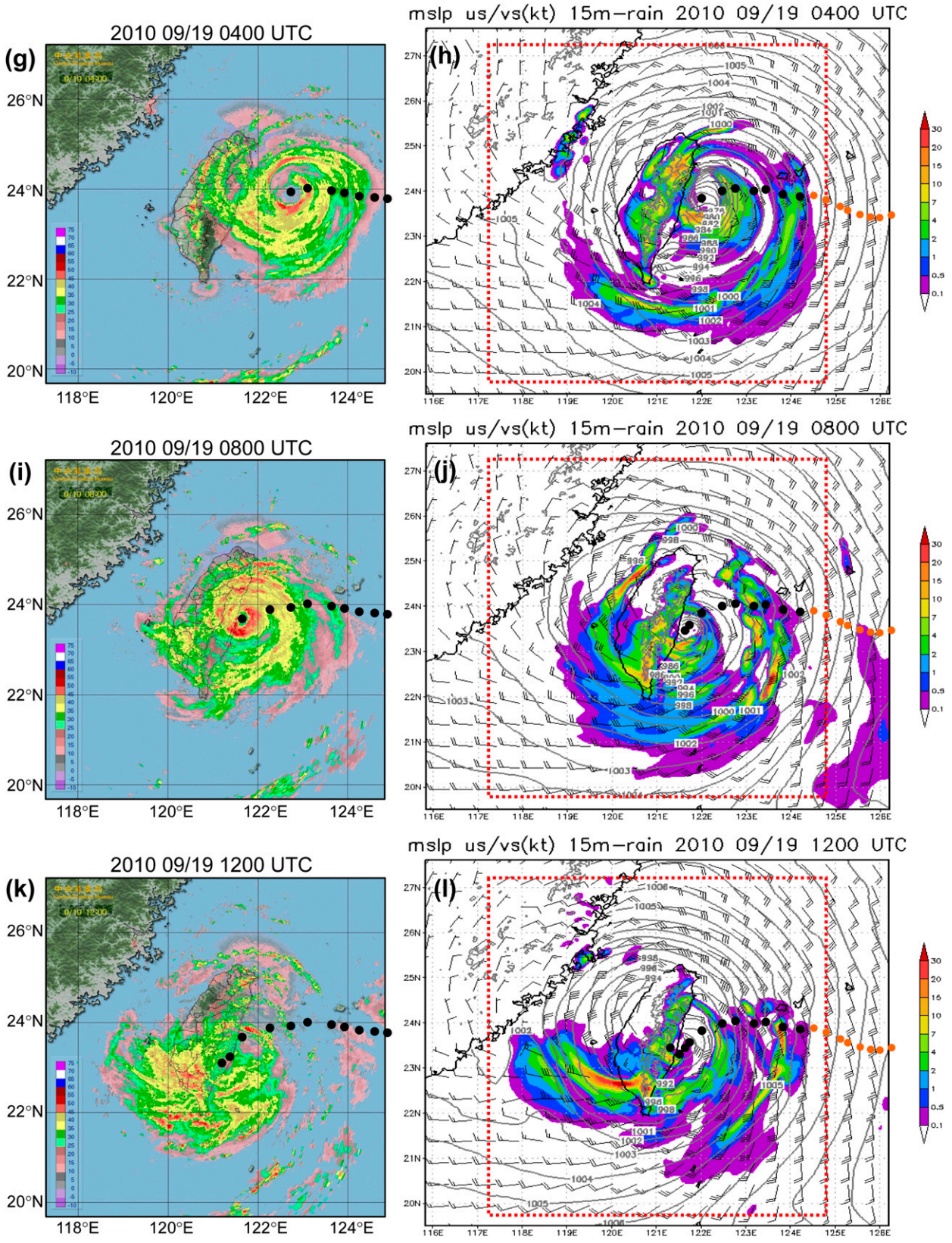

FIG. 2. (*Continued*)

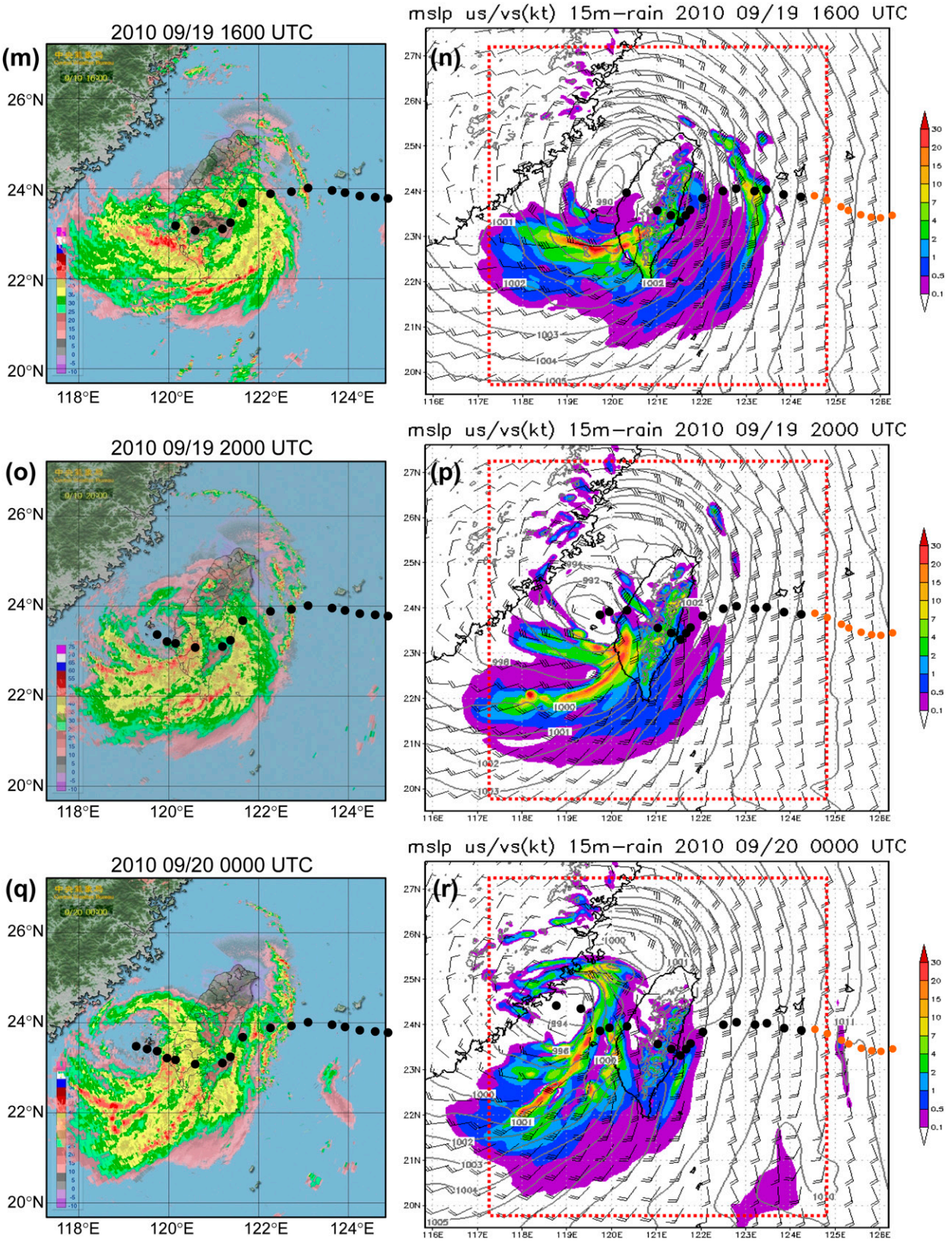

FIG. 2. (*Continued*)

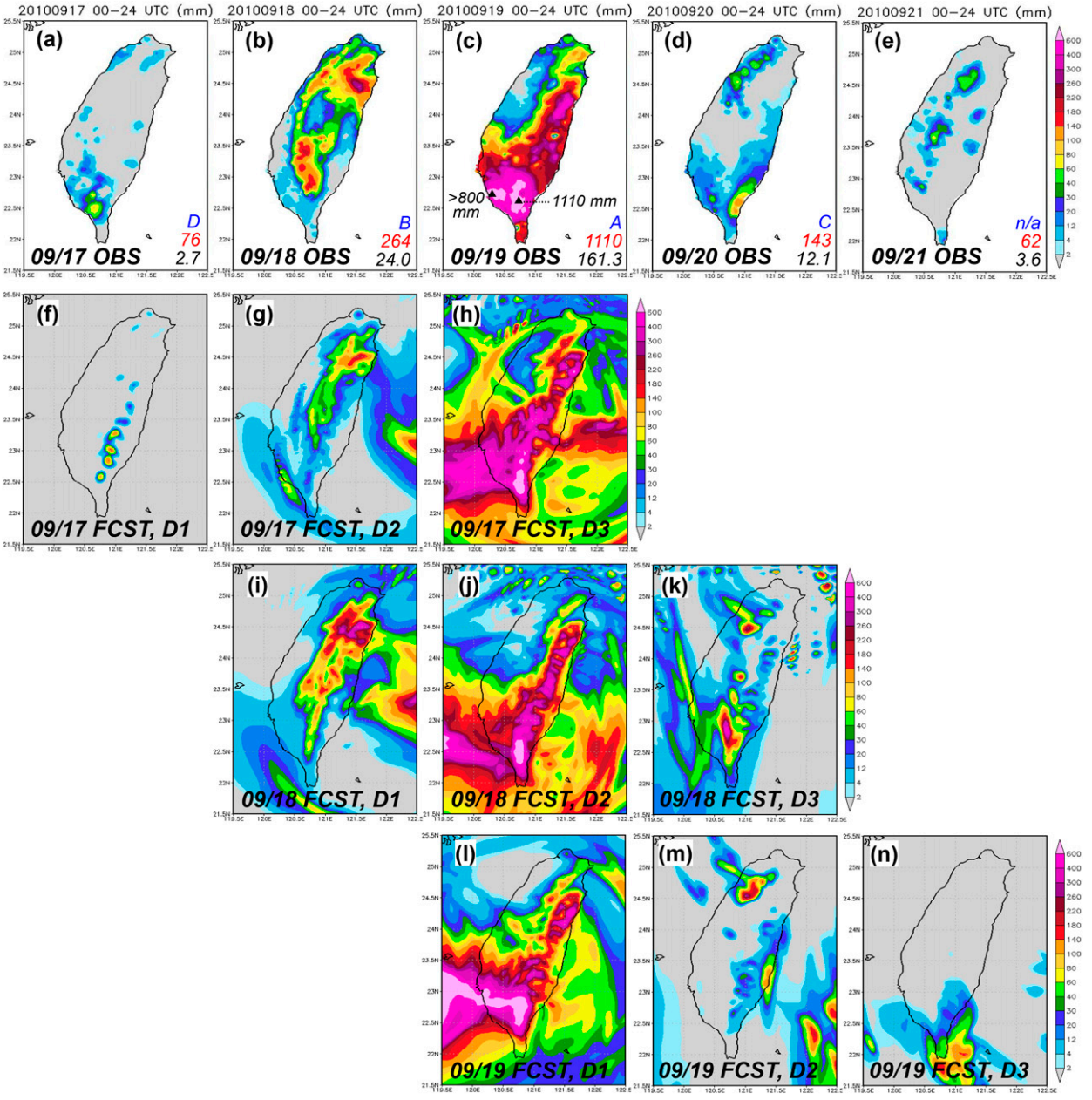

FIG. 3. Observations of total 24-h daily rainfall distributions (mm, 0000–2400 UTC) by rain gauges over Taiwan from (a) 17 Sep to (e) 21 Sep 2010, and corresponding CReSS 24-h QPFs made at 0000 UTC 17 Sep for (f) day 1 (0–24 h), (g) day 2 (24–48 h), and (h) day 3 (48–72 h), during TY Fanapi. (i)–(k) and (l)–(n) As in (f)–(h), but showing 24-h QPFs for days 1–3 made at 0000 UTC of 18 Sep and 19 Sep 2010, respectively. The classification group (see text for details), observed rainfall maximum, and averaged amounts are listed at lower right in (a)–(e) and also marked in (c). The color scales are up to 600 mm for all panels.

are used with no cumulus parameterization (Tsuboki and Sakakibara 2002). Subgrid-scale parameterizations include turbulent mixing in the planetary boundary layer (Mellor and Yamada 1974; Tsuboki and Sakakibara 2007) and surface radiation and momentum/energy fluxes with a substrate model (Kondo 1976; Louis et al. 1981; Segami et al. 1989). Without nesting grids, the main features of the model are similar to those used in Wang et al. (2009, 2013b) and further details can be found in Tsuboki and Sakakibara (2007) (and are available online at http://www.rain.hyarc.nagoya-u.ac.jp/~tsuboki/cress_html/index_cress_jpn.html).

The CReSS model has been used by the author to carry out real-time forecasts for Taiwan since 2006 and,

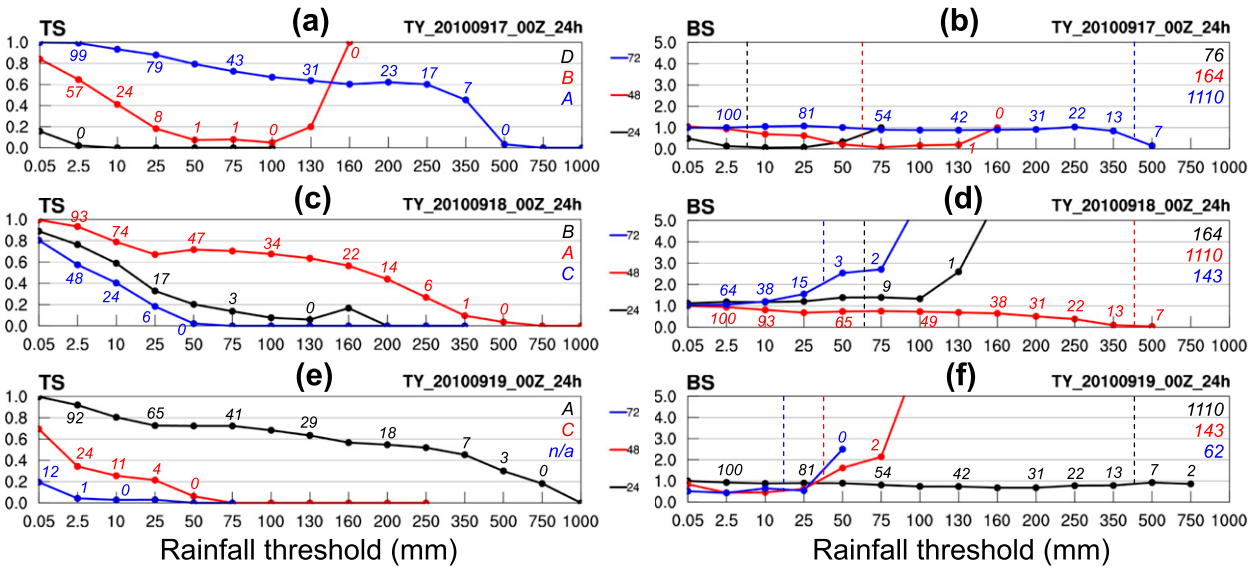

FIG. 4. (a) TS and (b) BS of CReSS 24-h QPFs for day 1 (black), day 2 (red), and day 3 (blue) from the forecast made at 0000 UTC 17 Sep 2010 as a function of threshold (mm). (c),(d) and (e),(f) As in (a),(b), but from the forecasts made at 0000 UTC 18 Sep and 19 Sep 2010, respectively. The hit rate (a/N) and observed base rate (O/N) in percent (%, rounded to integer) are labeled for selected points inside the panels for (left) TS and (right) BS, where the base rate at 10% is marked (vertical dashed line). The (left) classification group and (right) observed maximum 24-h rainfall are also given at the upper-right corner (all using corresponding colors).

at 4-km grid spacing, it produced highly valuable forecasts prior to TY Morakot (2009), the most devastating TC to hit the island in five decades (Wang et al. 2013b; Elsberry et al. 2013). For example, the forecast starting at 0000 UTC 6 August predicted a peak 48-h rainfall of >1900 mm in southern Taiwan, about two-thirds of the observed 4-day total. Since 2010, through participation in the Taiwan Cooperative Precipitation Ensemble Forecast Experiment [TAPEX; http://www.ttfri.narl.org.tw/eng/study02A.html, e.g., Lee et al. (2013)] at the Taiwan Typhoon and Flood Research Institute (TTFRI), a horizontal grid spacing of 2.5 km has been achieved with a domain size of 1080 km × 900 km ($x \times y$), which has been further enlarged to 1500 km × 1200 km since 2012 (both with 40 levels, cf. Fig. 5). By using configurations similar to those for research (e.g., Bryan et al. 2003; Done et al. 2004; Liu et al. 2006; Clark et al. 2007, 2009; Roberts and Lean 2008), high-quality operational forecasts out to 3 days as that shown in Fig. 2 (now to 78 h) are routinely produced every 6 h (at 0000, 0600, 1200, and 1800 UTC, Table 1). In fact, the 2.5-km CReSS is the only cloud-resolving member in TAPEX, where all other members use 5 km in their fine domain (e.g., Lee et al. 2013). Once available, the forecasts are posted in real time at the author's website at http://cressfcst.es.ntnu.edu.tw/ where all past results also reside for verification.

For all real-time CReSS forecasts, the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) analyses and forecasts (Kanamitsu 1989; Kalnay et al. 1990; Moorthi et al. 2001; Kleist et al. 2009), every 6 h on a 1° × 1° latitude/longitude grid at 26 levels were used as initial and boundary conditions (IC/BCs). The GFS forecasts (at 3-h intervals) contain forecast errors, and the track errors in particular are often a major error source in model QPFs for TCs (e.g., Marchok et al. 2007; Wang et al. 2013b; Yamada et al. 2013). At the lower boundary, terrain data on a (1/120)° grid and the NCEP analyzed sea surface temperatures (1° × 1°) are provided for each forecast (Table 1).

b. Data and methodology

Observational data used in this study, mostly provided by the Central Weather Bureau (CWB) of Taiwan, include the best-track, synoptic weather maps, radar reflectivity VMI composites every 30 min, and hourly visible/infrared cloud imageries from the geostationary Multifunctional Transport Satellite (MTSAT) during the TC case periods in 2010–12. For QPF verification, hourly rainfall data from more than 400 automated rain gauges over Taiwan (Hsu 1998) are used (Fig. 1), and all categorical measures of model QPFs are computed at these gauge sites.

To select the TC case periods included for study, the following methods are employed. First, only 24-h QPFs (either 0000–2400 or 1200–1200 UTC) are evaluated since we are mainly concerned about the bulk rainfall

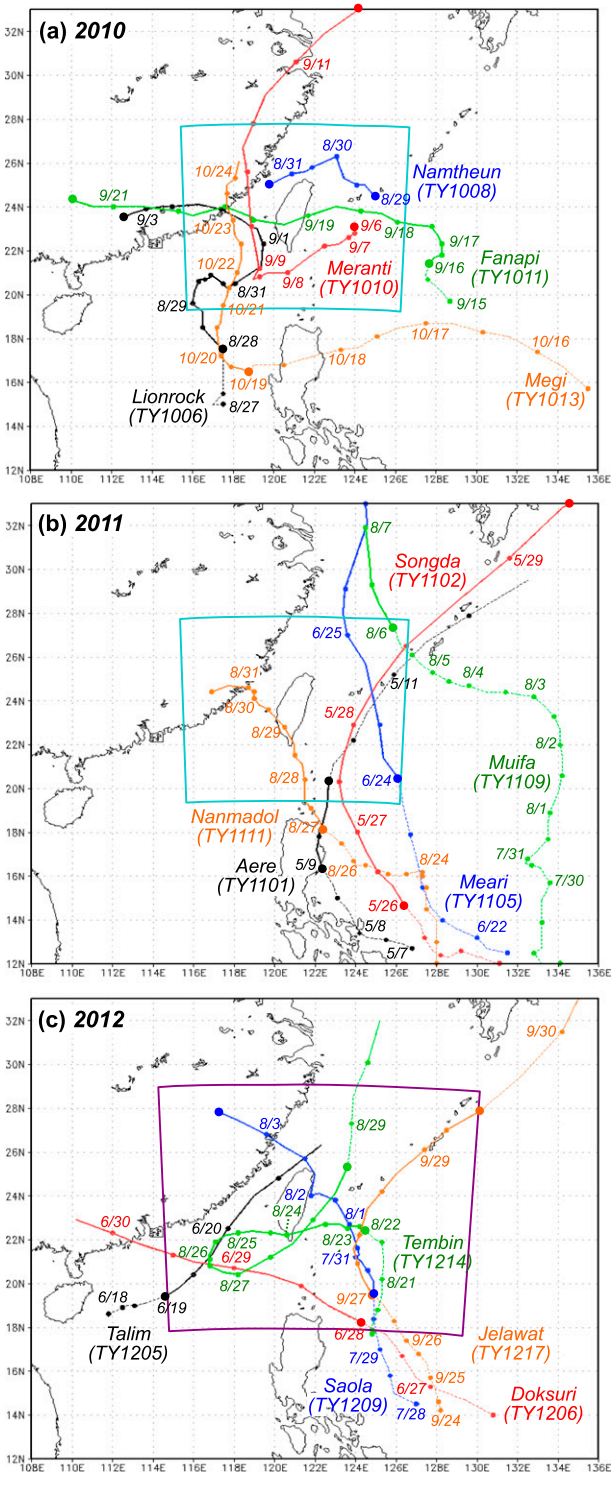

FIG. 5. CWB best tracks (dotted) of typhoon cases in (a) 2010, (b) 2011, and (c) 2012. Typhoon center positions are given every 6 h and marked by solid dots every 12 h (at 0000 and 1200 UTC), and month/date are labeled (at 0000 UTC) as needed. The segments of tracks included in this study are solid and thickened (cf. Table 2), with endpoints enlarged (when applicable), and the forecast domain in 2010–11 (2012) is plotted in cyan (purple) in the corresponding panel.

from the TCs rather than in shorter periods (say, ≤12 h). Using a longer accumulation period (and more tolerance in time), the issue of double penalty on higher-resolution QPFs typical in categorical measures is also less serious (e.g., Ebert and McBride 2000). Second, only sea-warning periods issued by the CWB (including land warnings) are eligible so that the TCs were not too far away from Taiwan. Third, at least part of the rainfall in Taiwan during each 24-h period must be linked to the TC or under its influence such as outer circulation, distant rainbands, or due to interaction with the monsoon. This is determined subjectively using weather maps and radar and satellite loops. Fourth, at least one 24-h segment must be included from each of the 15 typhoons, so that none of them is left out. Using the above four criteria, a total of 99 segments are selected in this study and subsequently classified into four groups (A–D) according to the observed 24-h rainfall amount. To better reflect rainfall over a sizable area and its hazard potential, at least 50 sites (roughly one-eighth of total sites in Taiwan) reaching 100, 50, and 25 mm, and otherwise are used as the criteria for groups A, B, C, and D, respectively. Table 2 lists the 15 TCs, the data periods from each, and the classification result while Fig. 5 shows their tracks and the portions included for analysis. The four groups contain roughly equal number of segments, which will be referred to as "group size" for clarity. Thus, to explore the dependency of QPF skill on rainfall, the case periods from all TCs are classified, and the results for different groups will be presented later in sections 4 and 5.

c. *Skill measures of model QPFs*

In this study, the model QPFs are evaluated over the rain gauge sites using several commonly used measures, including TS and BS. By interpolation of model results onto these sites and assuming equal weights, the TS and BS at any given rainfall threshold can be easily obtained through the entries of the 2 × 2 contingency table (e.g., Schaefer 1990; Wilks 1995; Ebert et al. 2003) as

$$\mathrm{TS} = a/(a + b + c), \quad \text{and} \tag{1}$$

$$\mathrm{BS} = (a + b)/(a + c), \tag{2}$$

where a, b, and c are the counts of points where event occurs in both observation and forecast (hit), in forecast but not observed (false alarm), and in observation but not forecast (miss), respectively, among a total of N points, while d is the count of points in which the event is neither observed nor forecast (correct negative). Thus, $N = a + b + c + d$ and the observed and forecast point counts are $O = a + c$ and $F = a + b$, respectively. Other

TABLE 1. Summary of 2.5-km CReSS configurations, initial and boundary conditions (IC/BCs), and major physical options used for operational 3-day forecasts during the typhoon seasons of 2010–12.

Season	2010 and 2011	2012
Projection	Lambert conformal (center at 120°E, secant at 10° and 40°N)	
Grid spacing (km)	2.5 × 2.5 × 0.2–0.663 (0.5)*	
Domain size (km)	1080 × 900 × 20	1500 × 1200 × 20
Grid dimension (x, y, z)	432 × 360 × 40	600 × 480 × 40
Frequency and range	Every 6 h (at 0000, 0600, 1200, and 1800 UTC), 72 h	
IC/BCs	NCEP GFS analyses/forecasts (1° × 1°, 26 levels)	
Topography and SST	Real at (1/120)° and NCEP analyses at 1° resolution	
Cloud microphysics	Bulk cold-rain scheme (six species)	
PBL parameterization	1.5-order closure with turbulent kinetic energy prediction	
Surface processes	Energy/momentum fluxes and shortwave/longwave radiation	
Soil model	41 levels, every 5 cm to 2 m deep	
No. of cores	120	240

* The vertical grid spacing (Δz) of CReSS is stretched (smallest at the bottom), and the averaged spacing is given in the parentheses. The lowest 10 output levels are at z = 100.0, 323.7, 591.4, 896.4, 1232.4, 1593.8, 1975.6, 2373.3, 2782.9, and 3201.1 m, respectively.

scores used include the probability of detection (POD), false alarm rate (FAR), accuracy (ACC; Mesinger and Black 1992; Hong 2003), equitable threat score (ETS; Schaefer 1990), Peirce skill score (PSS; Mason 2003), and odds ratio (OR; Doswell et al. 1990; Wilks 1995), defined, respectively, as

$$\mathrm{POD} = a/(a+c), \tag{3}$$

$$\mathrm{FAR} = b/(a+b), \tag{4}$$

$$\mathrm{ACC} = (a+d)/N, \tag{5}$$

$$\mathrm{ETS} = (a-R)/(a+b+c-R), \tag{6}$$

$$\mathrm{PSS} = a/(a+c) - b/(b+d) = \mathrm{POD} - b/(b+d), \quad \text{and} \tag{7}$$

$$\mathrm{OR} = (a\times d)/(b\times c), \tag{8}$$

where for ETS in Eq. (6), R is the (expected) number of random hits and $R = O \times F/N$. The value ranges of these measures are between 0 and 1 for TS, POD, FAR, and ACC, $-1/3$ and 1 for ETS (Schaefer 1990), -1 and 1 for PSS, and 0 to ∞ for BS and OR. In all of the above measures, a higher value represents a better skill, except for FAR, which is the opposite and for BS where unity is the ideal value as mentioned.

3. Examples of CReSS model forecasts

In this section, several more forecast examples are used to demonstrate that the dependency of model QPF skill on rainfall amount for typhoons in Taiwan, as shown in section 1b, is a common occurrence rather than an exception. All made at 0000 UTC, these examples include periods of various rainfall amount. For Fanapi, Figs. 3i–k and 3l–n depict the total rainfall distributions for days 1–3 from the forecasts made at 0000 UTC 18 and 19 September 2010 (i.e., 24 and 48 h later than that shown in Figs. 3f–h), respectively, and their TS and BS are plotted in Figs. 4c–f. While both forecasts captured the evolution of daily rainfall over Taiwan quite well (Figs. 3b–e and 3i–n), the TS values are again much higher for 19 September (i.e., the day with the most rain), and tend to be lower for days with less rain (Figs. 4c,e). The BS curves for 19 September from both forecasts are quite stable and indicate reasonably good predictions in rain-area size, with only a slight underforecast from the one made on 18 September (Figs. 4d,f). For other less-rainy days, the BS tends to exhibit much larger variation (say, BS > 2) even at low thresholds when the rain areas are relatively small with $O/N < \sim 10\%$. With the information on O/N and H/N (where $O/N \geq H/N$), one can see that the TS values in Fig. 4 over the range of 75–350 mm for days other than 19 September are all computed with $O/N < 10\%$ and $H/N \leq 3\%$ and may be unstable when O/N approaches 0% (e.g., for day 2 at 160 mm in Fig. 4a). As described, the day-3 QPF made on 17 September is particularly good, where TS = 0.45, $O/N \approx 13\%$, and $H/N \approx 7\%$ at the high threshold of 350 mm. Later, the number of points involved to compute the scores will be referred to as the ''sample size'' exclusively in this paper.

Figure 6 presents other score measures in Eqs. (3)–(8) from the forecast at 0000 UTC 17 September (for Fanapi), together with O/N, F/N, and d/N. Characteristics similar to the TS with clear dependency on rainfall amount also appear in POD, ETS, and PSS, except toward the lowest thresholds for PSS (Figs. 6a,d,e). With the random hit rate (F/N) in R estimated using data from the entire model domain following Wang (2014), the ETS is almost identical to the TS (cf. Fig. 4a). The FAR also shows

TABLE 2. List of the 15 typhoons included in this study, their data period, 24-h segments, and the classification of segments based on rainfall amount. For each typhoon, the segments of 0000–2400 (1200–1200) UTC are listed in the first (second) row, following the order, and classified into four groups of A, B, C, and D (when at least 50 sites reach 100, 50, 25 mm, and less, respectively). The most rainy segment (based on the amount averaged over all sites) of each typhoon is shown in boldface, while case 1 (Lionrock) shared a period with case 2 (Namtheun, denoted by italics) and case 6 (Aere) has only one segment included (underlined).

Name (No.)	Case period	Segment	Classification	No. of segments	Case No.
Lionrock (TY1006)	0000 UTC 28 Aug–1200 UTC 29 Aug	0000–2400 UTC	C	2	1*a*
		1200–1200 UTC	C		
Namtheun (TY1008)	0000 UTC 29 Aug–1200 UTC 31 Aug	0000–2400 UTC	C **A**	4	2
		1200–1200 UTC	B B		
Lionrock (TY1006)	0000 UTC 31 Aug–1200 UTC 3 Sep	0000–2400 UTC	C **A** D	6	1*b*
		1200–1200 UTC	B C D		
Meranti (TY1010)	0000UTC 6 Sep–1200 UTC 11 Sep	0000–2400 UTC	D D C **B** D	10	3
		1200–1200 UTC	D D B C C		
Fanapi (TY1011)	0000 UTC 16 Sep–1200 UTC 21 Sep	0000–2400 UTC	D D B **A** C	10	4
		1200–1200 UTC	D D A A D		
Megi (TY1013)	0000 UTC 19 Oct–1200 UTC 24 Oct	0000–2400 UTC	B B **A** B D	10	5
		1200–1200 UTC	B A A C D		
Aere (TY1101)	0000 UTC 9 May–0000 UTC 10 May	0000–2400 UTC	**D**	1	6
		1200–1200 UTC			
Songda (TY1102)	0000 UTC 26 May–1200 UTC 29 May	0000–2400 UTC	C **B** D	6	7
		1200–1200 UTC	C C D		
Meari (TY1105)	0000 UTC 24 Jun–1200 UTC 26 Jun	0000–2400 UTC	B B	4	8
		1200–1200 UTC	**A** C		
Muifa (TY1109)	0000 UTC 6 Aug–1200 UTC 7 Aug	0000–2400 UTC	D	2	9
		1200–1200 UTC	**D**		
Nanmadol (TY1111)	0000 UTC 27 Aug–1200 UTC 1 Sep	0000–2400 UTC	B A A A C	10	10
		1200–1200 UTC	A **A** A C C		
Talim (TY1205)	0000 UTC 19 Jun–1200 UTC 22 Jun	0000–2400 UTC	B **A** C	6	11
		1200–1200 UTC	A B C		
Doksuri (TY1206)	0000 UTC 28 Jun–1200 UTC 30 Jun	0000–2400 UTC	C D	4	12
		1200–1200 UTC	**C** D		
Saola (TY1209)	0000 UTC 30 Jul–1200 UTC 3 Aug	0000–2400 UTC	B A A A	8	13
		1200–1200 UTC	A A **A** A		
Tembin (TY1214)	0000 UTC 22 Aug–1200 UTC 28 Aug	0000–2400 UTC	C B A C D B	12	14
		1200–1200 UTC	C **A** B D D B		
Jelawat (TY1217)	0000 UTC 27 Sep–1200 UTC 29 Sep	0000–2400 UTC	D **C**	4	15
		1200–1200 UTC	C D		
Total			A: 26, B: 21, C: 26, D: 26	99	

a tendency to be lower (better) when there is more rain, and becomes unstable as F/N approaches zero (Fig. 6b). In ACC, both a and d are valued equally [Eq. (5)] and a high score can be obtained from successful predictions of events or nonevents, and thus the value of day-3 forecast cannot be distinguished (Fig. 6c). On the other hand, OR requires both a and d to be at least not too small (relative to b and c) to register a high score, a condition met by the day-3 QPF across the threshold range of 10–350 mm (Fig. 6f). These OR scores (>20) are much higher than those for days 1–2 (typically ≤ 3), and thus the measure also shows a strong dependency on rainfall amount.

Figures 7 and 8 show the observed and predicted daily rainfall and the TS and BS from the forecasts made at 0000 UTC of 19–21 October for TY Megi (2010). During 19–22 October when Megi was over the South China Sea (cf. Fig. 5a), heavy rainfall occurred over northeastern Taiwan four days in a row (Figs. 7a–e). With a total up to 945 mm and peak intensity reaching 181 mm h^{-1}, however, the rainfall was the heaviest on 21 October (Fig. 7c), and serious landslides occurred along the coastal road in the region that killed 38 people, including tourists on board a bus that was washed off the cliff into the Pacific. Overall, the model captured the general rainfall pattern and its day-to-day variations reasonably well since 19 October (Fig. 7), except the serious overprediction over eastern Taiwan in the forecast made on 20 for 22 October (Fig. 7k) owing to large track errors (in the GFS forecast) where the TC turned northeastward too early and made landfall (cf. Fig. 5a). In the forecast made one day later on 21 October, the track error was reduced and the daily QPF became more accurate again (Figs. 7l–n). In

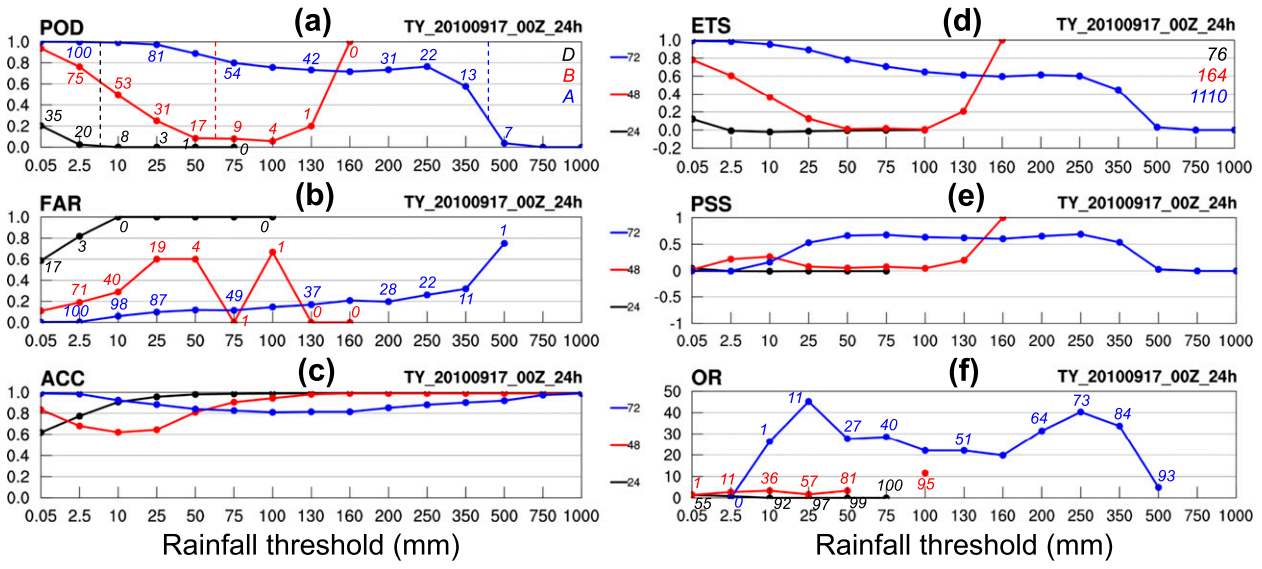

FIG. 6. As in Fig. 4a, but for (a) POD, (b) FAR, (c) ACC, (d) ETS, (e) PSS, and (f) OR of CReSS 24-h QPFs for days 1–3 (black, red, and blue, respectively) from the forecast made at 0000 UTC 17 Sep 2010. The observed base rate (O/N, marked at 10%), forecast base rate (F/N), and rate of correct negatives (d/N) in percent (%, rounded to integer) are labeled for selected points in (a),(b), and (f), respectively, and the classification group and observed maximum 24-h rainfall are also given in (a) and (d).

these three forecasts for Megi, again the TS is the highest (typically by at least 0.2) and the BS more ideal and stable for the day with the most rain (21 October) with nearly no exception (Fig. 8), despite that the differences in observed rainfall are smaller and larger variations in forecast quality also exist (compared to Fanapi). For 21 October, the TS remains above 0.5 up to 350 mm from the QPF made on 19 October (Fig. 8a) and up to 500 mm from that made on 21 October (Fig. 8e).

Three examples of other rainy typhoons are chosen and shown for the day with the most rainfall and the 24-h QPF made at an earlier time for that day (Fig. 9) and the corresponding TS and BS (Fig. 10). For TYs Saola (2012), Nanmadol (2011), and Tembin (2012), the most-rainy days in 0000–2400 UTC were 2, 28, and 24 August, respectively (Figs. 9a–c), and the QPFs made at 0000 UTC of 1 (day 2), 26 (day 3), and 22 August (day 3) all agree closely with the observations (Figs. 9d–f). Again, while the rainfall in successive days can be quite different as indicated by O/N, the TS tends to be higher for the day with the most rain, especially over middle to high thresholds (Figs. 10a,c,e), where the BS is also often more stable and closer to unity (Figs. 10b,d,f). In fact, with peak 24-h rainfall near or over 1000 mm, Fanapi, Megi, and Saola were the three most hazardous TCs to hit Taiwan in 2010–12 (Figs. 3c, 7c, and 9a), causing 2, 38, and 7 deaths and approximately \$150, \$43, and \$40 million (U.S. dollars) in direct losses, respectively (data source: CWB). The other 12 TYs together resulted in 1 casualty and a total loss of about \$40 million (U.S. dollars).

Now, four more examples from cases with moderate to little rainfall are presented in Figs. 11 and 12 for comparison. The dates selected include 9 September for Meranti (2010), 28 September for Jelawat (2012), 28 June for Doksuri (2012), and 6 August for Muifa (2011), all being the most-rainy day (0000–2400 UTC) of each typhoon (Figs. 11a–d). The day-3 QPFs made two days earlier (Figs. 11e–h) are all of reasonable (e.g., for Meranti) to good quality (e.g., for Jelawat). Overall, the TS values (Fig. 12) for periods with moderate rainfall (classified as B or C group) are considerable lower than those seen earlier in more-rainy cases, and progressively so when the rain becomes less and less (e.g., for Doksuri and Muifa) although the model predictions seem fair. Even so, the TS for the most-rainy day often remains higher than those for other days (Fig. 12).

From the ample examples above, a clear tendency for the model to yield higher TS and perform better (out to 3 days) when there is more overall rainfall brought by the TC in Taiwan (and vice versa) is well demonstrated, although exceptions do occur at times as expected. Such a dependency is not limited to TS, but also many categorical measures including POD, FAR, ETS, PSS, and OR (Fig. 6). With O/N and H/N given, one also sees in these examples that although many scores for groups B–D, which are essentially nonhazardous but account for three-fourths of all segments (cf. Table 2), can be computed over the range of 75–500 mm, they are often obtained with small samples (say, $O/N \leq 3\%$ and $H/N \leq 1\%$) especially toward higher thresholds (Figs. 4, 8, 10,

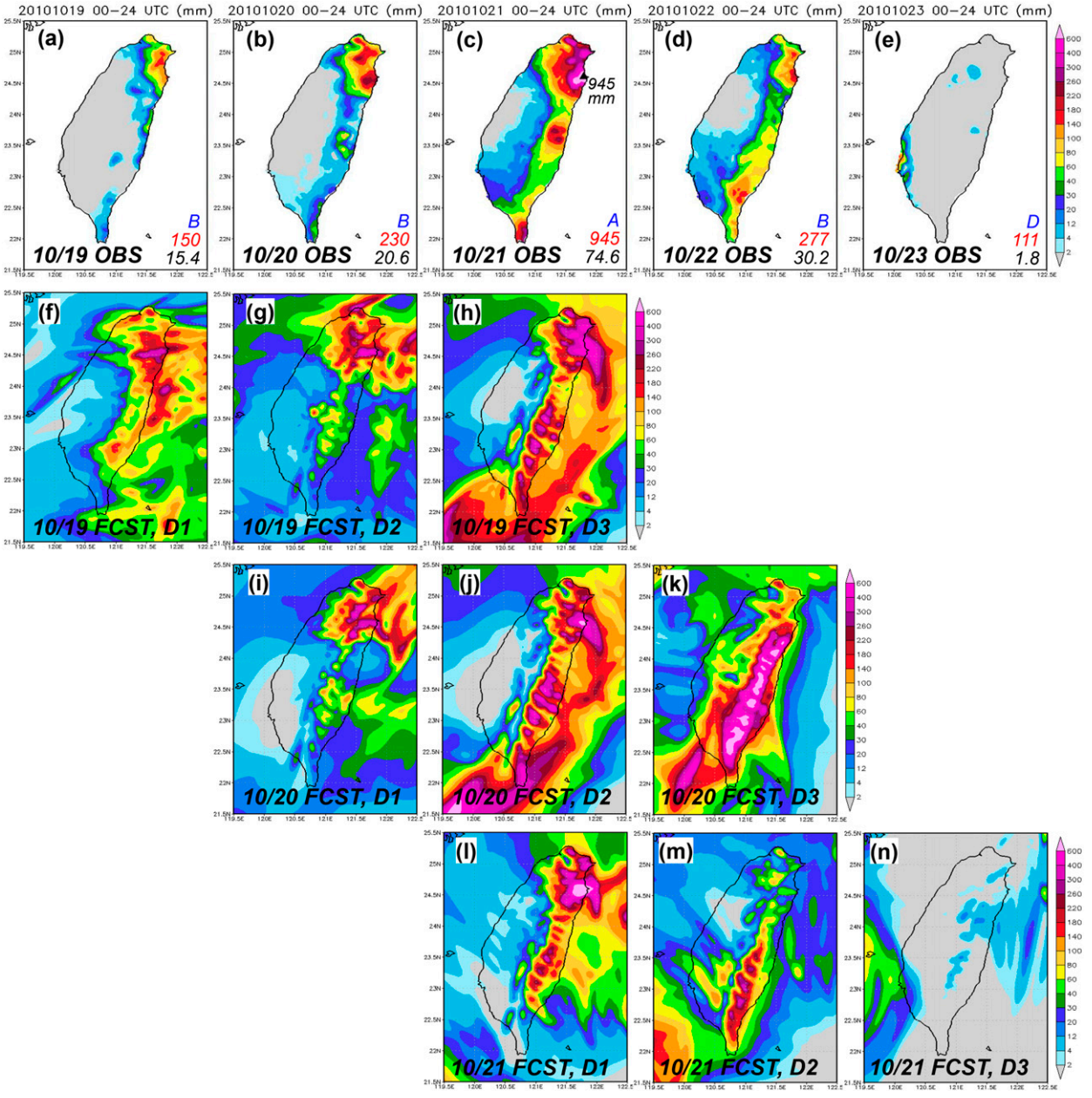

FIG. 7. As in Fig. 3, but showing total daily (0000–2400 UTC) rainfall observation (mm) from (a) 19 Oct to (e) 23 Oct 2010, and CReSS 24-h QPFs for days 1–3 made at 0000 UTC of (f)–(h) 19 Oct, (i)–(k) 20 Oct, and (l)–(n) 21 Oct 2010, respectively, during TY Megi.

and 12), and thus are of low statistical representativeness. While this issue will be discussed later in sections 5 and 6, we further explore the dependency of QPF skill on rainfall amount below.

4. The dependency of model QPF skills on rainfall

Following the method described in section 2b, the averaged TSs from CReSS forecasts starting at both 0000 and 1200 UTC for the target periods in groups A–D, from the most to least rainfall (Table 2), are shown in Fig. 13. The TS curves for day-1 (0–24 h) QPFs (i.e., those starting at the beginning of segments) are shown in Fig. 13a, and those for day-2 (24–48 h) and day-3 (48–72 h) QPFs that started 24 and 48 h earlier are shown in Figs. 13b and 13c, respectively. It is evident that the model indeed performs significantly better (as measured by TS) in forecasting group A when there is more rain, regardless of whether the forecasts start on the same day, or one or two days earlier. For day 1, the

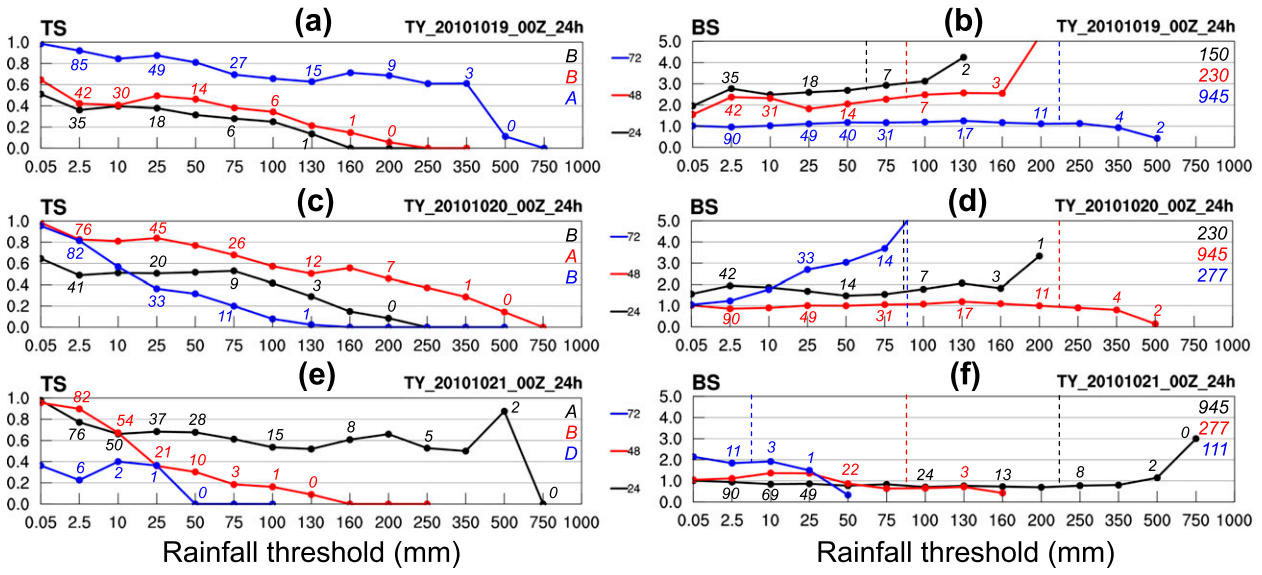

FIG. 8. As in Fig. 4, but for TS and BS of CReSS 24-h QPFs (0000–2400 UTC) from the forecast made at 0000 UTC of (a),(b) 19 Oct; (c),(d) 20 Oct; and (e),(f) 21 Oct 2010, respectively.

TS decreases from 0.9 to 0.52 at 50 mm then gradually to 0.29 at 250 mm, and indicates very good performance (Fig. 13a). The 0–24-h QPFs for group B, however, yield considerably lower TSs than those for group A across all thresholds up to 500 mm, by 0.12–0.28. For groups C and D as the rainfall becomes less and less, progressively lower TSs are yielded, by another 0.07–0.25 up to 130 mm (B minus C), and another 0.05–0.30 up to 50 mm (C minus D). For group D, the TS starts at 0.34 and drops quickly to below 0.1 above 10 mm. The scores from the QPFs made one and two days earlier exhibit similar characteristics (Figs. 13b,c) with very different skills for the four groups, except that the mean TS values tend to be lower to some extent at longer range, and more visible at middle and high thresholds. Even so, the TS of day-3 QPFs for group A still reaches 0.31 and 0.14 at the thresholds of 75 and 200 mm, respectively, significantly higher than the values for group B (Fig. 13c). Thus, although TS is also affected by forecast range, given typical track errors mainly dictated by and in accordance to those of the NCEP GFS (Table 3), its dependency on the overall rainfall amount is by far much stronger.

From Table 2, five typhoons that affected Taiwan longer and produced the most overall rainfall—Fanapi, Megi, Nanmadol, Saola, and Tembin (cases 4, 5, 10, 13, and 14)—are selected and the averaged TSs of QPFs made for the most-rainy 24 h of each TC (i.e., those shown in Figs. 3c, 7c, and 9a–c, all from group A, cf. Table 2) are also plotted in Fig. 13. These curves (labeled as "top 5") represent the model's capability to predict extreme typhoon rainfall in advance when it matters the most (recall that Fanapi, Megi, and Saola were the only three TCs to cause major hazards). Clearly, the model QPF skill for these top cases (group size = 5) is even higher than for group A, by 0.10–0.22 over thresholds of 50–500 mm (Fig. 13a). Higher TSs for the top-5 cases are also found in day-2 and day-3 QPFs (cf. group A), by 0.09–0.26 over 10–250 mm (Fig. 13b) and >0.15 over 25–350 mm (Fig. 13c), respectively. The dependency of QPF skills on rainfall amount is so strong, that the day-3 forecast for the top-5 cases at the threshold of 160 mm has a mean TS (0.36) higher than that of day-1 QPF at the lowest threshold of 0.05 mm for group D (TS = 0.34, Figs. 13a,c). Thus, the model is highly skillful in cases with extreme rainfall, but has much lower and only limited skill to predict weak TC rainfall, which is obviously not important. This result of better QPF performance by CReSS in events with larger magnitude is in sharp contrast to the common perception that numerical models (particular at a grid spacing of ≥10 km) in general are less capable to predict extreme rainfall events (e.g., Buizza et al. 1999; McBride and Ebert 2000; Mullen and Buizza 2001; Grubišić et al. 2005; Zhang et al. 2006; Cuo et al. 2011). The results here should also clarify some of the confusion in interpretation noted previously (e.g., Walser and Schär 2004; Yuan et al. 2005), as the scores tend to be lower toward higher thresholds in forecasts for an individual period but higher for more-rainy periods that can also reach higher thresholds. As shown, these two seemingly inconsistent, or even contradictory, behaviors of scores

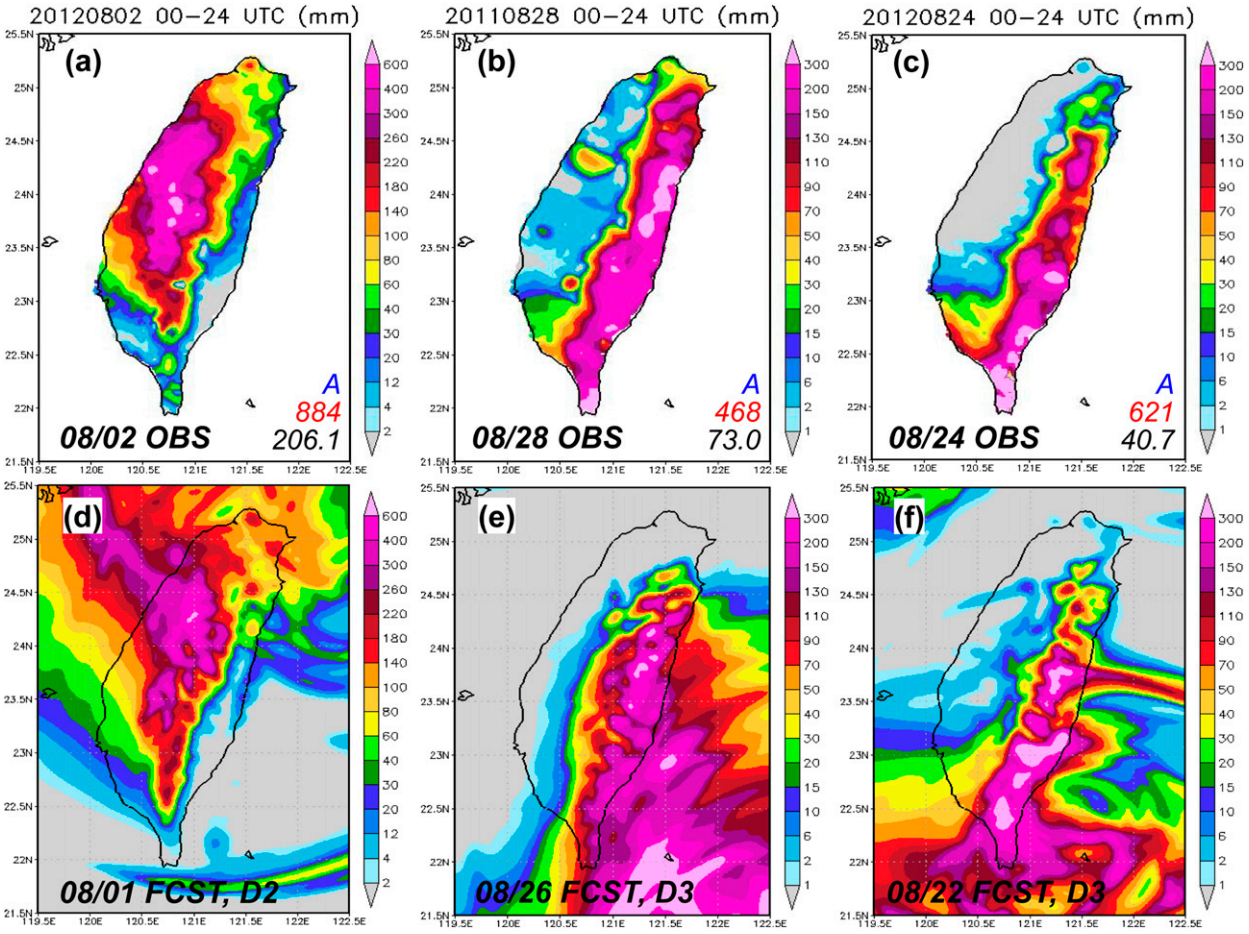

FIG. 9. As in Fig. 3, but showing total daily (0000–2400 UTC) rainfall observation (mm) on (a) 2 Aug 2012, (b) 28 Aug 2011, and (c) 24 Aug 2012, and corresponding CReSS 24-h QPFs made at 0000 UTC of (d) 1 Aug 2012 for day 2 (24–48 h), (e) 26 Aug 2011 for day 3 (48–72 h), and (f) 22 Aug 2012 for day 3 (48–72 h). Note that the color scales are up to 600 mm in (a),(d) but 300 mm in (b),(c) and (e),(f).

are in fact two different facets of the same phenomenon (i.e., the positive correlation between the scores and rain-area sizes). As pointed out in section 1a, the first facet (that the score decreases with threshold) is well known, but the second one is not.

5. The origin of the dependency

As mentioned, the simple linkage between the TS and the rain-area size can give rise to the dependency, since the chance for hits becomes higher when the observed and predicted rain areas are larger. But, is the rain-area size the only factor, or is the model indeed more skillful in predicting larger events? To answer this question, the effects from rain-area size using fixed rainfall thresholds are eliminated by the following method. For each of our 99 total 24-h segments (Table 2), the observed rainfall amounts at all sites over Taiwan are ranked and a new set of thresholds are identified to yield a coverage of 99% (at the lowest threshold), 95%, 90%–10% every 10%, 5%, and 1% (at the highest threshold) of the points. Figure 14 gives an example of the rainfall distribution as Fig. 3c (on 19 September 2010, for Fanapi) but plotted using the new thresholds. Thus, for the same areal coverage, every segment uses its own threshold that yields the specified O/N (say, 30%), and the impact of rain-area size is removed by standardizing according to the observed rainfall.

The averaged TS values for A–D and top-5 groups as in Fig. 13 but using the variable thresholds are plotted in Fig. 15 as functions of rain-area size (O/N), with the mean thresholds given at the bottom. Again, the averaged TS shows a relationship of top-5 > A > B > C > D across nearly all thresholds for QPFs at all ranges of days 1–3 when rain areas are fixed in size, with only a few exceptions at low thresholds or when O/N becomes small (Fig. 15). However, their differences appear smaller at middle and high thresholds, especially among

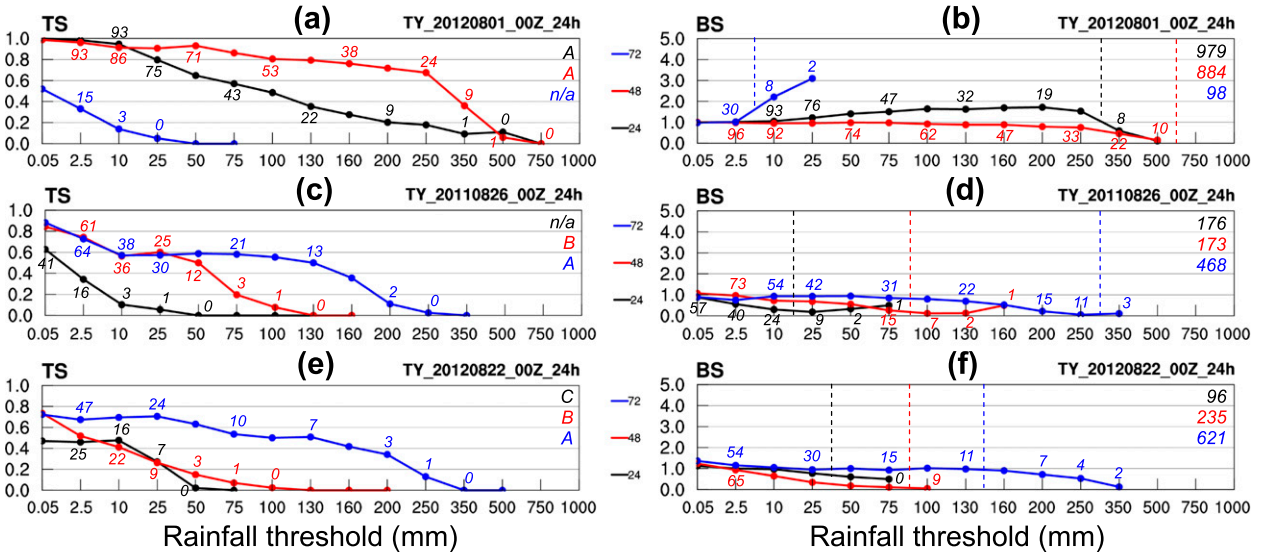

FIG. 10. As in Fig. 4, but for TS and BS of CReSS 24-h QPFs (0000–2400 UTC) from the forecast made at 0000 UTC of (a),(b) 1 Aug 2012; (c),(d) 26 Aug 2011; and (e),(f) 22 Aug 2012, respectively.

top-5, A, and B groups and toward day 3. Still evident in Fig. 15, the dependency, therefore, does not arise solely from the effects of rain-area size, but the CReSS model is indeed more skillful in predicting larger TC rainfall events. Typically with a significant component over the windward slopes (e.g., Chang et al. 1993; Cheung et al. 2008), more rain generally results over Taiwan when a typhoon makes landfall (direct hit), moves slower, and

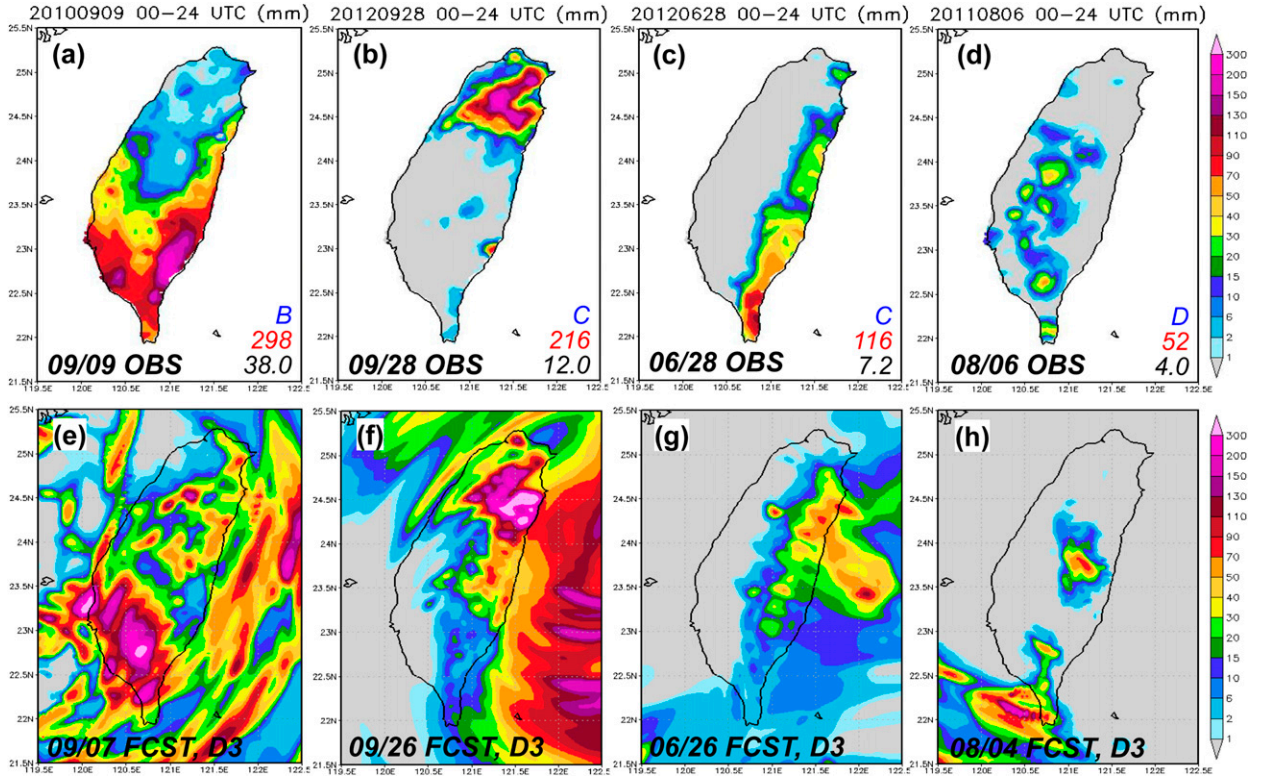

FIG. 11. As in Figs. 9b,e, but showing total daily rainfall observation (mm, 0000–2400 UTC) on (a) 9 Sep 2010, (b) 28 Sep 2012, (c) 28 Jun 2012, and (d) 6 Aug 2011, and corresponding CReSS 24-h QPFs made at 0000 UTC of (e) 7 Sep 2010, (f) 26 Sep 2012, (g) 26 Jun 2012, and (h) 4 Aug 2011, all for day 3, respectively. The color scales are up to 300 mm for all panels.

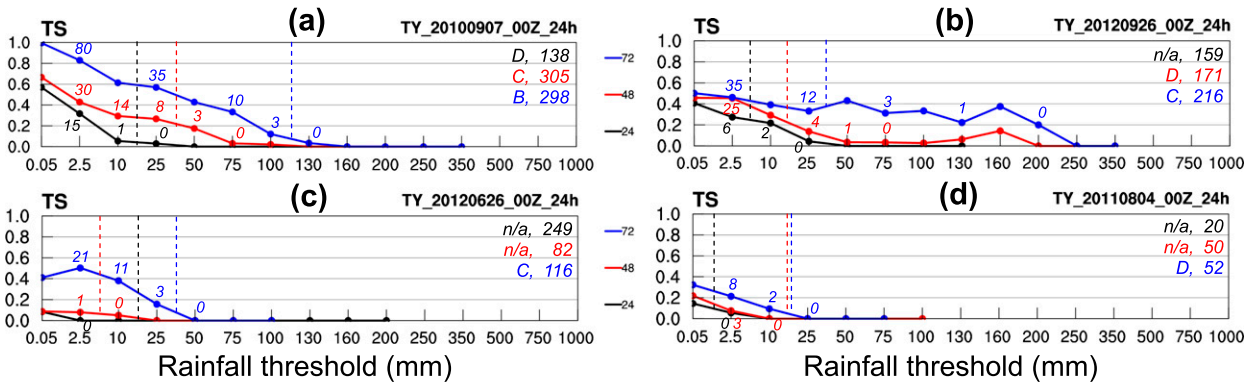

FIG. 12. As in Fig. 4a, but for TS of CReSS 24-h QPFs (0000–2400 UTC) from the forecast made at 0000 UTC (a) 7 Sep 2011, (b) 26 Sep 2012, (c) 26 Jun 2012, and (d) 4 Aug 2011, respectively. The hit rate (a/N, %) are labeled for selected points and observed base rate (O/N) at 10% is also marked (vertical dashed line). The classification group and observed maximum 24-h rainfall are both given.

remains close for a longer period, is larger in size and stronger, and induces or interacts with stronger monsoon flow to increase the moisture supply (e.g., Chien et al. 2008; Wu et al. 2011; Chien and Kuo 2011; Wang et al. 2012; Hsu et al. 2013; Chang et al. 2013). In addition to terrain, these factors (especially the track) are mainly controlled by the structural characteristics of the TC and its environment at the larger scale, and modern-day models such as CReSS clearly can handle these processes properly given reasonable IC/BCs and track errors (cf. Table 3), thus yielding overall higher scores for larger, more hazardous events within a range of 3 days.

6. The implications of the dependency

In earlier sections, a clear dependency of QPF skill, as measured by TS and other scores alike, on rainfall amount for typhoons in Taiwan is demonstrated, and the CReSS model performs better when there is more rain, both among TC cases and among different days (periods) in the same TC. Shown to be very strong, this dependency is a fundamental and vital property of the verification measures of QPFs and has important implications in many ways, especially in the use and verification/evaluation of model QPFs. In this section, these implications are discussed. Below, let us first address what is the model's true ability to predict hazardous and extreme TC rainfalls in Taiwan, and how we can evaluate model QPFs for such events more accurately and effectively.

a. Model performance for hazardous and extreme events

As mentioned, among 15 TCs affecting Taiwan in 2010–12, only 3–5 brought extreme rainfall and were hazardous. If we were to understand the model's capability to forecast such events, the scores can only be averaged for target periods of comparable magnitude (i.e., top cases) as has been done in this study (cf. Figs. 13 and 15), given their strong dependency on rainfall. Needless to say, such evaluations are crucially important as these events are great threats. As listed in Table 4, the mean TS values from day-1 QPFs by the 2.5-km CReSS for the top-5 cases at the thresholds of 25 mm (near 1 in.) and 50, 130, 200, and 350 mm (the CWB heavy-rainfall thresholds) are as high as 0.67, 0.67, 0.58, 0.51, and 0.32, respectively. The day-2 and day-3 QPFs that started 24 and 48 h earlier register mean TSs of 0.75, 0.73, 0.57, 0.42, and 0.17, and 0.64, 0.57, 0.37, 0.33, and 0.22 at the same thresholds, respectively (cf. Fig. 13). These values are strikingly high and indicate that the model not only performs very well in 0–24-h QPFs but is also already skillful about 2–2.5 days prior to the extreme rainfall. When it matters the most, such a lead time is undoubtedly precious and essential for emergency action and hazard reduction. For individual, hazardous TCs, such as Fanapi, Megi, Saola (Figs. 4, 8, and 10a), and Morakot (which was large in size and exhibited strong monsoon interaction) in 2009 (by 4-km CReSS forecasts out to 48 h) (Wang et al. 2012, 2013b; Wang 2014), the TS can be even higher (up to 0.87 at 200 mm and 0.69 at 350 mm, Figs. 13a,b and Table 4), despite the fact that traditional measures tend to double panelize high-resolution models for having the ability to produce concentrated rainfall at high intensity as in the real atmosphere (e.g., Ebert and McBride 2000; Grubišić et al. 2005; Clark et al. 2007, 2009; Roberts and Lean 2008). Such scores are no doubt among the highest in the world and achievable exactly because of the dependency, since Taiwan is one of the most-rainy places when hit by a typhoon. For group A (≥50 sites reaching 100 mm), the corresponding TSs at the same five thresholds from QPFs for days 1–3 are in the range of

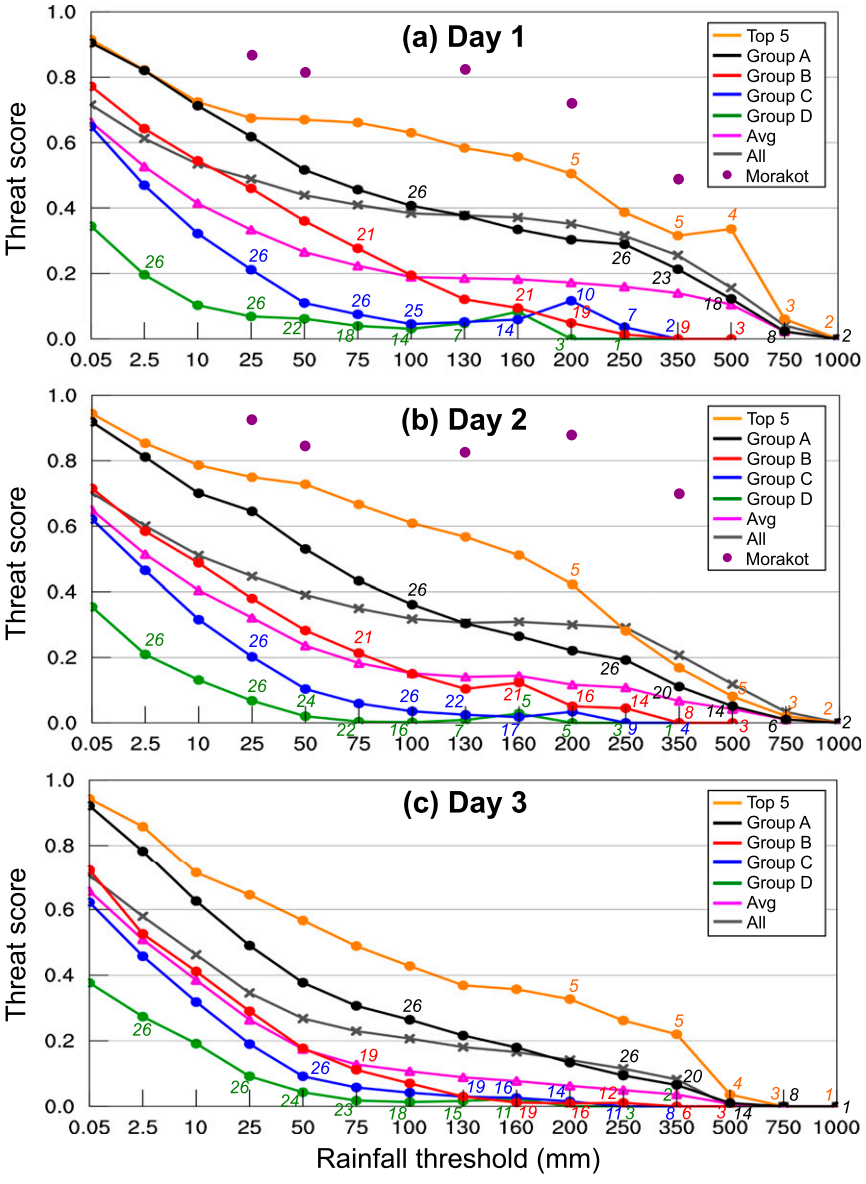

FIG. 13. Averaged TS of CReSS 24-h QPFs from forecasts at 0000 and 1200 UTC for (a) day 1 (0–24 h), (b) day 2 (24–48 h), and (c) day 3 (48–72 h) for groups A–D and the top-5 cases (see legends) as a function of rainfall threshold (mm). Day-1 QPFs are from forecasts starting at the beginning of the 24-h periods of verification, and day-2 and day-3 QPFs are from those starting 24 and 48 h earlier, respectively (i.e., targeting the same periods). The top-5 cases (from group A) are the most-rainy 24-h period from the five most-rainy TCs (one segment from each, Table 2). The group sizes (of available segments) are labeled for selected points. The unweighted (arithmetic) averages (labeled as "Avg") and the TS values obtained by combining all the forecasts into just one contingency table (labeled as "All") are also plotted, and the purple dots depict the forecasts for TY Morakot in 2009 (for 0000–2400 UTC 8 Aug, days 1–2 only, see Table 4).

0.49–0.65, 0.38–0.53, 0.22–0.38, 0.14–0.30, and 0.07–0.21. Although not as high as the top-5 cases, they are already much higher than typical values and long-term averages in many places (e.g., Olson et al. 1995; Mesinger 1996; Ralph et al. 2005; Marchok et al. 2007; Tuleya et al. 2007; Charba and Samplatsky 2011).

b. *Implications in model QPF verification and evaluation*

When evaluating the performance of model QPFs, it is common to average the scores over a long period through many events. Likewise, to assess the model's

TABLE 3. Mean track errors (km) in all forecasts for the case periods in this study within a range of 72 h, given every 12 h since the starting time. The sample counts and the percentages from CReSS and NCEP/GFS (if the TC center is located outside the domain of CReSS) at each range are also given.

Forecast range (h)	0	12	24	36	48	60	72
Mean track error (km)	42.5	56.4	79.5	109.0	130.8	171.9	200.0
Sample counts	120	120	120	119	118	104	89
Sample from CReSS (%)	69.2	69.2	69.2	69.7	69.5	73.1	76.4
Sample from NCEP GFS (%)	30.8	30.8	30.8	30.3	30.5	26.9	23.6

ability for any particular event, often a significant one, the scores are usually averaged through multiple forecasts over several days. Based on this study, the latter practice is in fact *ineffective* (and even misleading at times) and unable to obtain a correct understanding, since the model has a different ability from one day (period) to the next. One can imagine how the key information about the performance of model QPFs would be lost, even near the highest threshold, if the curves in Figs. 4a,c,e were averaged for Fanapi, or Figs. 8a,c,e for Megi, with equal weights.

After examining the nature and characteristics of skill measures in depth, it is evident that up to three types of mistakes are committed if the scores are averaged over multiple forecasts without proper classification. First, when an arithmetic mean is calculated, we assume that the scores for different forecast periods are of the same importance, but this is obviously not true. Taking Fanapi as an example, the only important day was 19 September (Fig. 3c), and it was 21 October for Megi (Fig. 7c). Hence, the forecasts for all other days were not nearly as important and yet the same weight is assigned. Second,

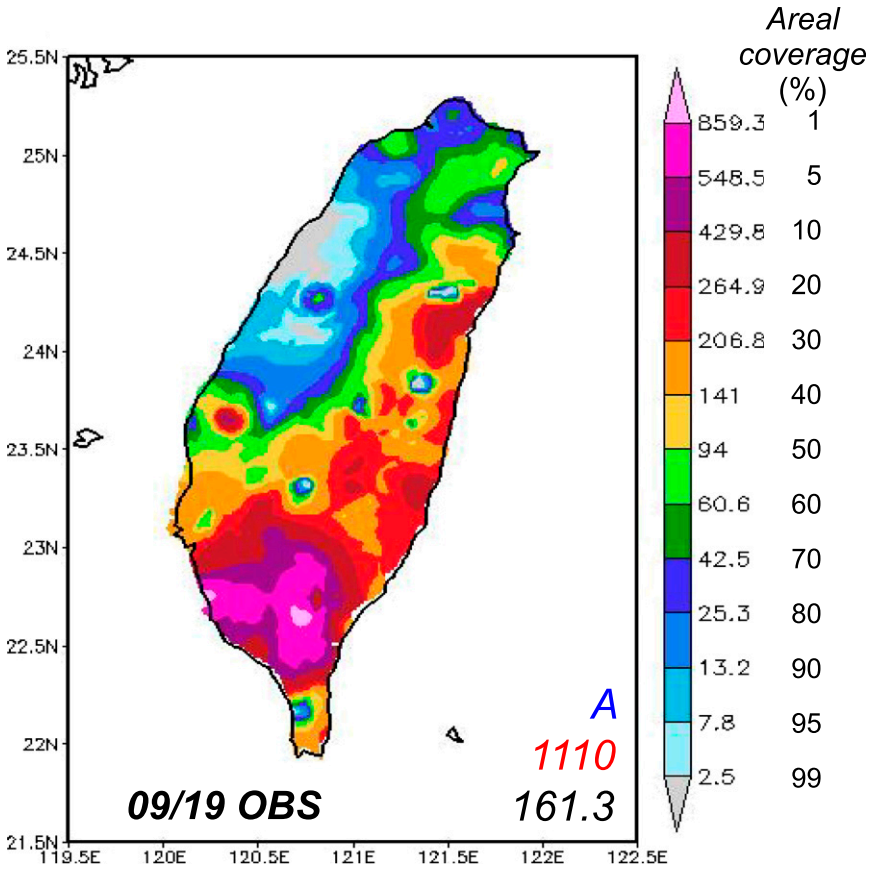

FIG. 14. As in Fig. 3c, but plotted using 13 rainfall thresholds (mm, scales on right) that yield rain areas covering 99%, 95%, 90%–10% (every 10%), 5%, and 1% of all rain gauge sites over Taiwan (as labeled next to the color scale).

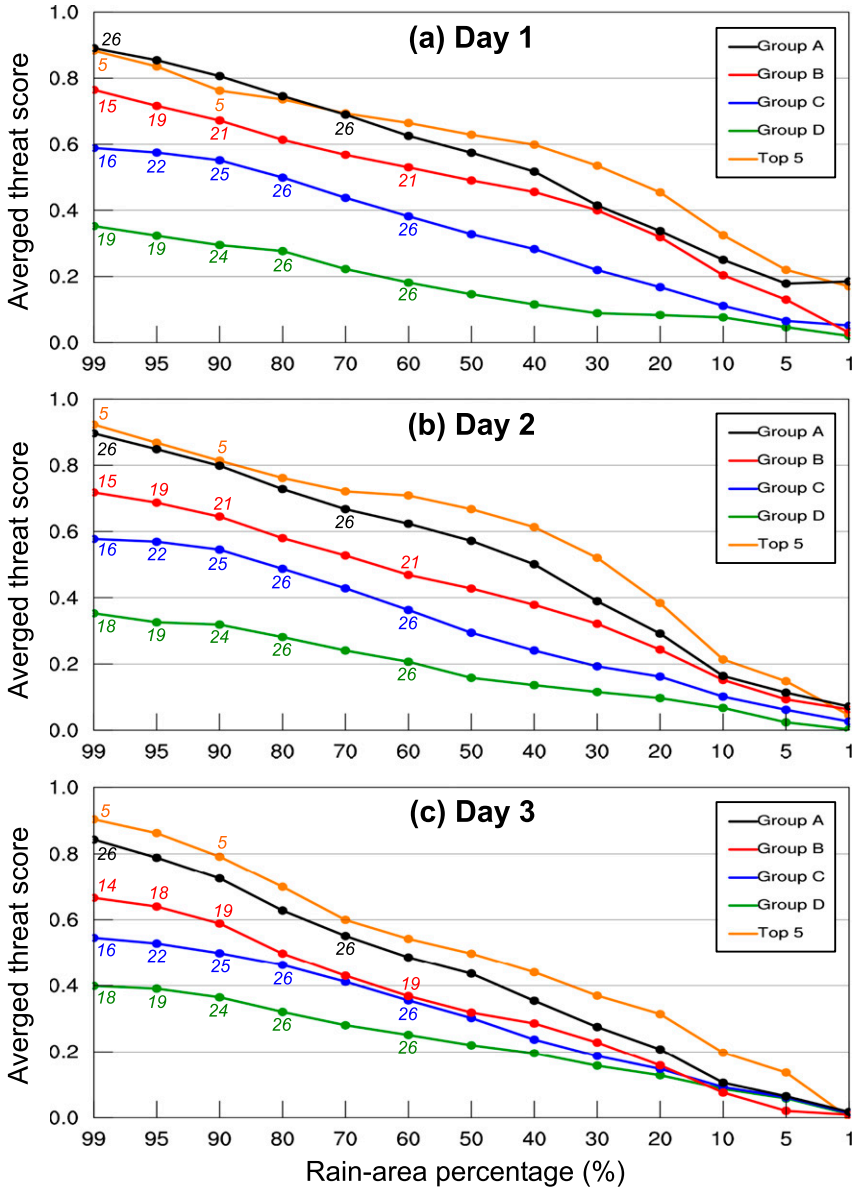

FIG. 15. As in Fig. 13, but for averaged TS for (a) day 1 to (c) day 3 for groups A–D and the top-5 cases at rainfall thresholds covering certain percentages of all sites over Taiwan, from 99%, 95%, and 90% (low threshold) to 10%, 5%, and 1% (high threshold). The averaged rainfall thresholds (mm) for the five groups are also given at the bottom.

an arithmetic mean also implies the same statistical significance in all values, and this is problematic as well. As demonstrated (cf. Figs. 3, 6, 8, 10, and 12), the scores of smaller events are nearly always computed from fewer points, especially toward higher thresholds, and thus are of lower statistical representativeness compared to those of much larger events. Such differences are also evident in Fig. 16, where a significant decrease in

TABLE 4. The averaged TSs from day-1, day-2, and day-3 QPFs by the 2.5-km CReSS model for the top-5 cases (Fanapi, Megi, Nanmadol, Saola, and Tembin, the most-rainy 24-h from each, cf. Table 2) and group A (at least 50 sites ≥ 100 mm, group size = 26) in 2010–12, at the threshold of 25 mm (≈1 in.) and the CWB heavy-rainfall thresholds of 50, 130, 200, and 350 mm [(24 h)$^{-1}$]. Day-1 QPFs are from forecasts starting at the beginning of the 24-h periods of verification, and day-2 and day-3 QPFs are from those starting 24 and 48 h earlier, respectively. Day-1 and day-2 TSs from 4-km CReSS QPFs at real time, verified for (0000–2400 UTC) 8 Aug 2009 (Wang et al. 2013b; Wang 2014), are also given for comparison.

Rainfall threshold		25 mm	50 mm	130 mm	200 mm	350 mm
Morakot (2009)	Day 1 (0–24 h)	0.86	0.81	0.82	0.73	0.48
	Day 2 (24–48 h)	0.93	0.83	0.82	0.87	0.69
Top-5 cases (n = 5)	Day 1 (0–24 h)	0.67	0.67	0.58	0.51	0.32
	Day 2 (24–48 h)	0.75	0.73	0.57	0.42	0.17
	Day 3 (48–72 h)	0.64	0.57	0.37	0.33	0.22
Group A (n = 26)	Day 1 (0–24 h)	0.62	0.52	0.38	0.30	0.21
	Day 2 (24–48 h)	0.65	0.53	0.30	0.22	0.11
	Day 3 (48–72 h)	0.49	0.38	0.22	0.14	0.07

sample sizes exists between groups A and B, and the averaged O/N is often <1%–3% at thresholds above 50–130 mm in groups B–D. For these two reasons, different weights (perhaps based on sample size) should be assigned for different periods to reflect the model's ability to predict precipitation more accurately, if an average must be performed. Applied to the TC dataset here, the unweighted averages as well as the TSs obtained by combining entries from all the forecasts (for the same range) into just one contingency table, which are identical to the averages weighted by the number of points involved (at the given threshold) in each forecast [$a + b + c$ in Eq. (1)] for TS, are also plotted in Fig. 13 for comparison. As expected, the unweighted TS (arithmetic mean) is obscured and cannot reveal the key information across essentially all thresholds. On the other hand, the weighted average does a better job since the smaller events are no longer inflated, but they are still considerably lower than the scores for top events, even at high thresholds except for day 2 (Fig. 13). The third possible mistake is that, the period over which all the forecasts are averaged is often chosen quite arbitrarily, and the rainfall near the two ends may not even be linked to the TC. Even for the segments in groups B–D included in this study, they occupy the majority of case periods but are nonhazardous and not important, so it does not matter if the model has limited ability to predict them. Finally, it is worthy to note that for BS, which can become unstable easily in individual forecasts, using just one contingency table provides good indications for systematic bias, which is quite small for the 2.5-km CReSS (Fig. 17).

Naturally, the long-term averaged scores (e.g., Junker et al. 1992; Mesinger 1996; Buizza et al. 1999; Reynolds 2003; Fritsch and Carbone 2004) represent the model's ability to predict rainfall from a variety of systems over a spectrum of scales. Because of the dependency property, the ability for large (and typically rare) events should be relatively high, but progressively lower for the smaller (and more frequent) ones down to local convection, as the predictability also gradually diminishes. While the model's ability is much better reflected by scores from just one contingency table for the whole period or weighted averages (cf. Fig. 13), proper classification is recommended if a deeper understanding on model QPFs is desired.

c. *Implications in the use of model QPFs*

The dependency of model QPFs on rainfall amount also has important implications to forecasters and other users in how the model forecasts should be perceived when a large event is predicted. The better model performance with higher scores for larger events when the hazard potential is high is probably one of the best news that any forecaster can hear. As also demonstrated in our examples (sections 1b and 3), modern-day high-resolution models like CReSS can often capture the TC rainfall scenario reasonably well (within 72 h), and the predicted overall rainfall is positively correlated with the observed one (Fig. 18 and Table 5). In an operational setting, this means that the forecasters, if they are aware of this dependency property, should trust the model QPF more, not less, when the model predicts a more extreme event, provided of course that the TC evolution (especially in track) in the forecast is reasonable. Prior to Morakot (2009), although an amount of >1900 mm was predicted by CReSS (Wang et al. 2013b, cf. Table 4) and by several other models up to 1365 mm (e.g., Hendricks et al. 2011; Wu et al. 2010), the CWB however initially issued a lower amount (cf. Fig. 15 of Wang et al. 2013b), partly attributable to a lack of confidence in the models. While real-time verification is always essential to judge the likelihood of the scenario in each forecast, a deeper understanding on the dependency will also help resolve

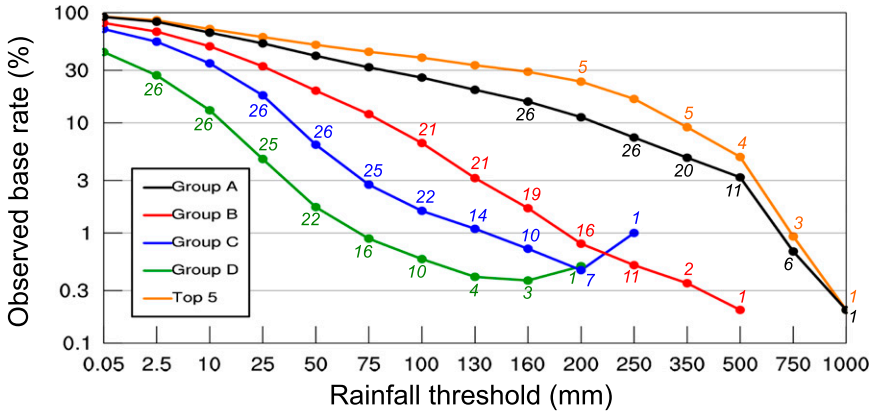

FIG. 16. As in Fig. 13, but showing the observed base rate (O/N) in percent (%) for groups A–D and "top-5 cases" as a function of rainfall threshold (mm). The ordinate is in logarithmic scale, and the group sizes (numbers of segments available) are labeled.

some of the issues related to how forecasts of extreme events should be perceived and used.

In this study, the dependency of model QPFs by the 2.5-km CReSS on rainfall amount arises due both to the correlation between scores and rain-area size, and the fact that the model in general can indeed capture the factors leading to larger events (such as a closer and slower track, and more favorable TC structure and environment). As a fundamental property that can arise from the effects of rain-area size alone, the dependency is expected to exist in all models such as the Weather Research and Forecasting Model (Skamarock et al. 2005) with a similar setup. The dependency should also appear in other rainfall regimes (such as mei-yu) in Taiwan, and hold true around the world even without significant terrain (given an adequate data size). In many regions, however, the dependency is likely less strong than what is shown here, although this remains to be studied and verified in the future.

7. Summary and conclusions

In this study, a strong dependency in the performance of QPFs by the CReSS model at a cloud-resolving grid spacing of 2.5 km on overall rainfall amount (event magnitude), as measured by scores such as the TS, for typhoons in Taiwan is demonstrated and explored through forecasts out to 3 days for all 15 TCs in 2010–12. With higher scores on average when the target periods receive more rain in reality, the model QPFs thus perform the best in the prediction of extreme rainfall events (sections 1b, 3, and 4, cf. Fig. 13). The dependency property arises not only because the "hit area" tends to be larger in proportion when the (observed) rain area is

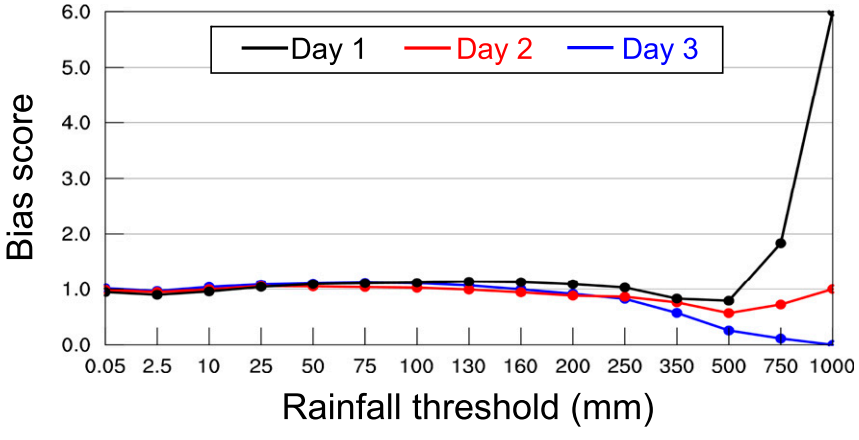

FIG. 17. As in Fig. 4b, but for BS obtained from a single 2×2 contingency table with entries combined from all the forecasts for the 99 target segments in this study (one table for each range of days 1–3, with the sample size $N \approx 40\,000$).

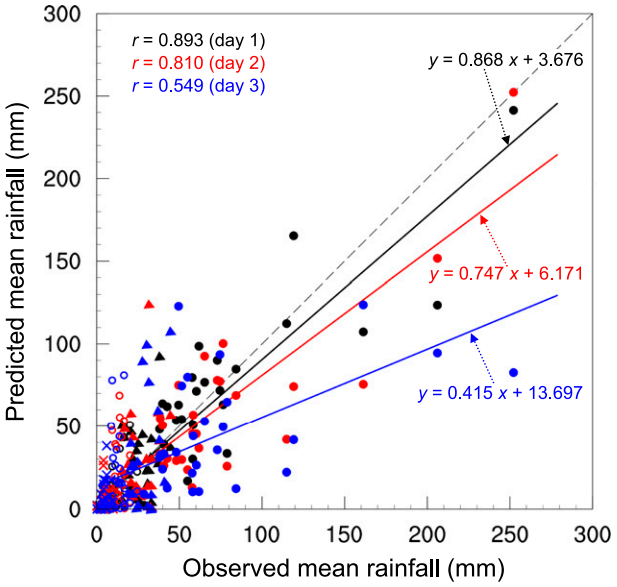

FIG. 18. Scatterplot of averaged 24-h rainfall (among all sites in Taiwan) in the observation vs in model QPFs (for days 1–3) for the 99 segments in this study, their linear regression lines, and correlation coefficients (r). Results for days 1–3 are in black, red, and blue, while data points in groups A–D are plotted as solid dots, solid triangles, open circles, and crosses, respectively.

larger, but also due to the fact that the favorable factors leading to more TC rainfall in Taiwan can be captured with reasonable accuracy by the model (within 72 h), including the track, structure, and environment of the TCs (section 5 and Fig. 15). This phenomenon is thus expected to exist in other rainfall regimes in Taiwan, in other models, and in other regions around the world.

While the dependency itself may be in contrast to the common perception that numerical models are generally less capable of predicting extreme rainfall events, it has important implications on how model QPFs should be evaluated if we wish to understand properly the true ability of models to forecast hazardous events (section 6), and ultimately, how we could improve heavy-rainfall QPFs in the future. For such events, the averaging of scores should only be performed among target periods with comparable overall rainfall, since the inclusion of data from less-rainy periods would lower the score, easily to a level not indicative to the true skill (cf. Fig. 13). In other words, the dependency property must be taken into account in the verification and evaluation of model QPFs, especially for hazardous/extreme events. With proper classification, the day-1 QPFs by the 2.5-km CReSS, targeted for the most-rainy 24 h from the 5 most-rainy TCs (one 24-h period from each) in 2010–12, have averaged TS at thresholds of 25, 50, 130, 200, and 350 mm of 0.67, 0.67, 0.58, 0.51, and 0.32 (sections 5 and 6a), respectively. Starting two days earlier, the day-3 QPFs for the same 24-h target periods yield mean TS of 0.64, 0.57, 0.37, 0.33, and 0.22 at the same thresholds, not much lower than day-1 forecasts. For the extreme case of Morakot (2009), the QPFs from the 4-km CReSS are even higher, with TS at least 0.81 up to 130 mm, 0.73 at 200 mm, and 0.48 at 350 mm (Figs. 13a,b and Table 4). These scores indicate superior model performance in predicting extreme TC rainfall even 2–2.5 days in advance with appreciable lead time, and they are strikingly high precisely because of the dependency property since Taiwan, with its steep topography, can receive extreme rainfall when hit by a typhoon. The top-5 cases are a subset of group A (≥50 sites reaching 100 mm in 24 h), for which the day 1–3 QPFs at the same five thresholds (25–350 mm) have averaged TS in the range of 0.49–0.65,

TABLE 5. Summary of group classification based on QPFs for days 1–3 vs observation using the same criteria in section 2b (top half). The model QPFs are first interpolated onto the rain gauge sites before classification, and the results of correct grouping, or incorrect (off by one, two, or three categories) are also given (bottom half). For day-3 QPFs, two forecasts (those from 0000 and 1200 UTC 17 Oct 2010) are missing due to data lost in real-time operation, so the total segments for day 3 are 97 instead of 99. Italics indicate that the summed values are for days 1 and 2.

	Model QPFs												
	Day 1				Day 2				Day 3				
Observation	A	B	C	D	A	B	C	D	A	B	C	D	Sum
A	21	3	1	1	18	6	2	0	15	5	4	2	*26*
B	9	8	1	3	8	6	5	2	7	4	3	5	*21*
C	2	7	9	8	6	5	6	10	5	5	6	10	*26*
D	0	0	1	25	1	1	1	23	1	5	3	17	*26*
Sum	32	18	12	37	33	17	14	35	28	19	16	34	*99*
Correct			63				53				42		
Off by 1			29				34				33		
Off by 2			6				11				19		
Off by 3			1				1				3		
Total			99				99				97		

0.38–0.53, 0.22–0.38, 0.14–0.30, and 0.07–0.21, already significantly higher than typical long-term averages in summer. For groups B–D with less and less rainfall, the mean TS values are lower and lower (Fig. 13) as the rain areas become smaller and more random in nature (e.g., Walser et al. 2004; Zhang et al. 2006). Shown clearly to concentrate toward the lower-left corner in Fig. 18, all periods in groups B–D are nonetheless unimportant, as the threat for hazards is (almost) nonexistent.

Because the ability of model QPFs can be vastly different for successive days with different rainfall, the common practice to average the scores over a period (typically several days) through multiple forecasts and to attempt to evaluate the model for a given event is inappropriate and not recommended. As demonstrated, such averages would inevitably mask the crucial information (i.e., the skill for the most-rainy period) by lower scores for other periods with less rain (in overall amount but not necessarily in peak value), which are unimportant, not statistically representative, and perhaps not even related to the rainfall system being examined (section 6b). By giving equal weights for all periods in averaging as has usually been done, the results of little importance are artificially blown out of proportion. The outcome under such conditions is basically dictated by the periods included (e.g., Du et al. 1997), and an intermodel comparison would not be fair if different periods are used for them.

Finally, a better model performance with higher scores for more-rainy periods, when the hazard potential is high, also has vital implications for forecasters and the users of model QPFs, especially for emergency action and hazard reduction. If the forecasters/users are aware of this dependency property, they should raise their confidence in model QPFs when a more extreme event is predicted in advance, not the other way around. Of course, in an operational setting, constant real-time verification is always essential for any model forecast to be used wisely and effectively.

Acknowledgments. The author is thankful for the comments from the reviewers that helped improve the content and presentation of this work. The author also thanks assistants Ms. Y.-W. Wang, Mr. T.-C. Lin, Mr. T.-R. Wang, and Ms. D.-J. Kao, and graduate students S.-Y. Huang, C.-S. Chang, B.-K. Chiou, S.-H. Chen, and several others for their help during the course of this study. The CWB is acknowledged for providing the observational data and the radar plots used in Fig. 2, and the TTFRI and the National Center for High-performance Computing (NCHC) for providing the computational resources. This study is jointly supported by the Ministry of Science and Technology of Taiwan under Grants NSC-102-2119-M-003-003, NSC-102-2625-M-003-002, MOST-103-2119-M-003-001-MY2, and MOST-103-2625-M-003-001-MY2.

REFERENCES

Akter, N., and K. Tsuboki, 2012: Numerical simulation of Cyclone Sidr using a cloud-resolving model: Characteristics and formation process of an outer rainband. *Mon. Wea. Rev.,* **140,** 789–810, doi:10.1175/2011MWR3643.1.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.,* **131,** 2394–2416, doi:10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting,* **14,** 168–189, doi:10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Charba, J. P., and F. G. Samplatsky, 2011: High-resolution GFS-based MOS quantitative precipitation forecasts on a 4-km grid. *Mon. Wea. Rev.,* **139,** 39–68, doi:10.1175/2010MWR3224.1.

Chang, C.-P., T.-C. Yeh, and J.-M. Chen, 1993: Effects of terrain on the surface structure of typhoons over Taiwan. *Mon. Wea. Rev.,* **121,** 734–752, doi:10.1175/1520-0493(1993)121<0734:EOTOTS>2.0.CO;2.

——, Y.-T. Yang, and H.-C. Kuo, 2013: Large increasing trend of tropical cyclone rainfall in Taiwan and the roles of terrain. *J. Climate,* **26,** 4138–4147, doi:10.1175/JCLI-D-12-00463.1.

Cheung, K. K. W., L.-R. Huang, and C.-S. Lee, 2008: Characteristics of rainfall during tropical cyclone periods in Taiwan. *Nat. Hazards Earth Syst. Sci.,* **8,** 1463–1474, doi:10.5194/nhess-8-1463-2008.

Chien, F.-C., and H.-C. Kuo, 2011: On the extreme rainfall of Typhoon Morakot (2009). *J. Geophys. Res.,* **116,** D05104, doi:10.1029/2010JD015092.

——, Y.-C. Liu, and C.-S. Lee, 2008: Heavy rainfall and southwesterly flow after the leaving of Typhoon Mindulle (2004) from Taiwan. *J. Meteor. Soc. Japan,* **86,** 17–41, doi:10.2151/jmsj.86.17.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.,* **135,** 3456–3473, doi:10.1175/MWR3467.1.

——, ——, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting,* **24,** 1121–1140, doi:10.1175/2009WAF2222222.1.

Cotton, W. R., G. J. Tripoli, R. M. Rauber, and E. A. Mulvihill, 1986: Numerical simulation of the effects of varying ice crystal nucleation rates and aggregation processes on orographic snowfall. *J. Climate Appl. Meteor.,* **25,** 1658–1680, doi:10.1175/1520-0450(1986)025<1658:NSOTEO>2.0.CO;2.

Cuo, L., T. C. Pagano, and Q. J. Wang, 2011: A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. *J. Hydrometeor.,* **12,** 713–728, doi:10.1175/2011JHM1347.1.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.,* **134,** 1772–1784, doi:10.1175/MWR3145.1.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the weather

research and forecasting (WRF) model. *Atmos. Res. Lett.,* **5,** 110–117, doi:10.1002/asl.72.

Doswell, C. A., R. Davies-Jones, and D. K. Keller, 1990: On summary measures of skills in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585, doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.,* **125,** 2427–2459, doi:10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202, doi:10.1016/S0022-1694(00)00343-7.

——, U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE assessment of short-term quantitative precipitation forecasts (QPFs) from operational numerical weather prediction models. *Bull. Amer. Meteor. Soc.,* **84,** 481–492, doi:10.1175/BAMS-84-4-481.

Elsberry, R., H.-C. Kuo, G.-F. Lin, and E. Shamir, Eds., 2013: Editorial—Special issue of journal of hydrology on typhoon hydrometeorology. *J. Hydrol.,* **506,** 1–3, doi:10.1016/j.jhydrol.2013.10.046.

Fritsch, J. M., and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.,* **85,** 955–965, doi:10.1175/BAMS-85-7-955.

——, and Coauthors, 1998: Quantitative precipitation forecasting: Report of the eighth prospectus development team, U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.,* **79,** 285–299, doi:10.1175/1520-0477(1998)079<0285:QPFROT>2.0.CO;2.

Golding, B. W., 2000: Quantitative precipitation forecasting in the UK. *J. Hydrol.,* **239,** 286–305, doi:10.1016/S0022-1694(00)00354-1.

Grubišić, V., R. K. Vellore, and A. W. Huggins, 2005: Quantitative precipitation forecasting of wintertime storms in the Sierra Nevada: Sensitivity to the microphysical parameterization and horizontal resolution. *Mon. Wea. Rev.,* **133,** 2834–2859, doi:10.1175/MWR3004.1.

Hendricks, E. A., J. R. Moskaitis, Y. Jin, R. M. Hodur, J. D. Doyle, and M. S. Peng, 2011: Prediction and diagnosis of Typhoon Morakot (2009) using the Naval Research Laboratory's mesoscale tropical cyclone model. *Terr. Atmos. Oceanic Sci.,* **22,** 579–594, doi:10.3319/TAO.2011.05.30.01(TM).

Hong, J.-S., 2003: Evaluation of the high-resolution model forecasts over the Taiwan area during GIMEX. *Wea. Forecasting,* **18,** 836–846, doi:10.1175/1520-0434(2003)018<0836:EOTHMF>2.0.CO;2.

Hsu, J., 1998: ARMTS up and running in Taiwan. *Väisälä News,* **146,** 24–26.

Hsu, L.-H., H.-C. Kuo, and R. G. Fovell, 2013: On the geographic asymmetry of typhoon translation speed across the mountainous island of Taiwan. *J. Atmos. Sci.,* **70,** 1006–1022, doi:10.1175/JAS-D-12-0173.1.

Ikawa, M., and K. Saito, 1991: Description of a nonhydrostatic model developed at the Forecast Research Department of the MRI. MRI Tech. Rep. 28, 238 pp.

Johnson, L. E., and B. G. Olsen, 1998: Assessment of quantitative precipitation forecasts. *Wea. Forecasting,* **13,** 75–83, doi:10.1175/1520-0434(1998)013<0075:AOQPF>2.0.CO;2.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley and Sons, 240 pp.

Junker, N. W., J. E. Hoke, B. E. Sullivan, K. F. Brill, and F. J. Hughes, 1992: Seasonal and geographic variations in quantitative precipitation prediction by NMC's Nested-Grid Model and Medium-Range Forecast Model. *Wea. Forecasting,* **7,** 410–429, doi:10.1175/1520-0434(1992)007<0410:SAGVIQ>2.0.CO;2.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428, doi:10.1175/1520-0477(1990)071<1410:GNWPAT>2.0.CO;2.

Kanamitsu, M., 1989: Description of the NMC global data assimilation and forecast system. *Wea. Forecasting,* **4,** 335–342, doi:10.1175/1520-0434(1989)004<0335:DOTNGD>2.0.CO;2.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W. S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP global data assimilation system. *Wea. Forecasting,* **24,** 1691–1705, doi:10.1175/2009WAF2222201.1.

Kondo, J., 1976: Heat balance of the China Sea during the air mass transformation experiment. *J. Meteor. Soc. Japan,* **54,** 382–398.

Lee, C.-S., and Coauthors, 2013: Assessment of sewer flooding model based on ensemble quantitative precipitation forecast. *J. Hydrol.,* **506,** 101–113, doi:10.1016/j.jhydrol.2012.09.053.

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

Liu, A. Q., G. W. K. Moore, K. Tsuboki, and I. A. Renfrew, 2004: A high-resolution simulation of convective roll clouds during a cold-air outbreak. *Geophys. Res. Lett.,* **31,** L03101, doi:10.1029/2003GL018530.

Liu, C., M. W. Moncrieff, J. D. Tuttle, and R. E. Carbone, 2006: Explicit and parameterized episodes of warm-season precipitation over the continental United States. *Adv. Atmos. Sci.,* **23,** 91–105, doi:10.1007/s00376-006-0010-9.

Louis, J. F., M. Tiedtke, and J. F. Geleyn, 1981: A short history of the operational PBL parameterization at ECMWF. *Workshop on Planetary Boundary Layer Parameterization,* Reading, United Kingdom, ECMWF, 59–79.

Maesaka, T., G. W. K. Moore, Q. Liu, and K. Tsuboki, 2006: A simulation of a lake effect snowstorm with a cloud resolving numerical model. *Geophys. Res. Lett.,* **33,** L20813, doi:10.1029/2006GL026638.

Marchok, T., R. Rogers, and R. Tuleya, 2007: Validation schemes for tropical cyclone quantitative precipitation forecasts: Evaluation of operational models for U.S. landfalling cases. *Wea. Forecasting,* **22,** 726–746, doi:10.1175/WAF1024.1.

Marzban, C., and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Wea. Forecasting,* **21,** 824–838, doi:10.1175/WAF948.1.

Mason, I. B., 2003: Binary events. *Forecast Verification—A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley and Sons, 37–76.

McBride, J. L., and E. E. Ebert, 2000: Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Wea. Forecasting,* **15,** 103–121, doi:10.1175/1520-0434(2000)015<0103:VOQPFF>2.0.CO;2.

Mellor, G. L., and T. Yamada, 1974: A hierarchy of turbulent closure models for planetary boundary layers. *J. Atmos. Sci.,* **31,** 1791–1806, doi:10.1175/1520-0469(1974)031<1791:AHOTCM>2.0.CO;2.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade.

Bull. Amer. Meteor. Soc., **77,** 2637–2649, doi:10.1175/1520-0477(1996)077<2637:IIQPFW>2.0.CO;2.

——, and T. L. Black, 1992: On the impact on forecast accuracy of the step-mountain (eta) vs. sigma coordinate. *Meteor. Atmos. Phys.,* **50,** 47–60, doi:10.1007/BF01025504.

Moorthi, S., H. L. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. Tech. Procedures Bull. 484, Office of Meteorology, National Weather Service, 14 pp.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.,* **129,** 638–663, doi:10.1175/1520-0493(2001)129<0638:QPFOTU>2.0.CO;2.

Murakami, M., 1990: Numerical modeling of dynamical and microphysical evolution of an isolated convective cloud—The 19 July 1981 CCOPE cloud. *J. Meteor. Soc. Japan,* **68,** 107–128.

——, T. L. Clark, and W. D. Hall, 1994: Numerical simulations of convective snow clouds over the Sea of Japan: Two-dimensional simulation of mixed layer development and convective snow cloud formation. *J. Meteor. Soc. Japan,* **72,** 43–62.

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting,* **10,** 498–511, doi:10.1175/1520-0434(1995)010<0498:EOYOQP>2.0.CO;2.

Ralph, F. M., and Coauthors, 2005: Improving short-term (0-48 h) cool-season quantitative precipitation forecasting: Recommendations from a USWRP workshop. *Bull. Amer. Meteor. Soc.,* **86,** 1619–1632, doi:10.1175/BAMS-86-11-1619.

——, E. Sukovich, D. Reynolds, M. Dettinger, S. Weagle, W. Clark, and P. J. Neiman, 2010: Assessment of extreme quantitative precipitation forecasts and development of regional extreme event thresholds using data from HMT-2006 and COOP observers. *J. Hydrometeor.,* **11,** 1286–1304, doi:10.1175/2010JHM1232.1.

Reynolds, D., 2003: Value-added quantitative precipitation forecasts. How valuable is the forecaster? *Bull. Amer. Meteor. Soc.,* **84,** 876–878, doi:10.1175/BAMS-84-7-876.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.,* **136,** 78–97, doi:10.1175/2007MWR2123.1.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575, doi:10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Segami, A., K. Kurihara, H. Nakamura, M. Ueno, I. Takano, and Y. Tatsumi, 1989: Operational mesoscale weather prediction with Japan Spectral Model. *J. Meteor. Soc. Japan,* **67,** 907–924.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468 STR, 88 pp.

Tsuboki, K., and A. Sakakibara, 2002: Large-scale parallel computing of cloud resolving storm simulator. *High Performance Computing,* H. P. Zima et al., Eds., Springer, 243–259.

——, and ——, 2007: *Numerical Prediction of High-Impact Weather Systems: The Textbook for the Seventeenth IHP Training Course in 2007.* Hydrospheric Atmospheric Research Center, Nagoya University, and UNESCO, 273 pp.

Tuleya, R. E., M. DeMaria, and R. J. Kuligowski, 2007: Evaluation of GFDL and simple statistical model rainfall forecasts for U.S. landfalling tropical storms. *Wea. Forecasting,* **22,** 56–70, doi:10.1175/WAF972.1.

Walser, A., and C. Schär, 2004: Convection-resolving precipitation forecasting and its predictability in Alpine river catchments. *J. Hydrol.,* **288,** 57–73, doi:10.1016/j.jhydrol.2003.11.035.

——, D. Lüthi, and C. Schär, 2004: Predictability of precipitation in a cloud-resolving model. *Mon. Wea. Rev.,* **132,** 560–577, doi:10.1175/1520-0493(2004)132<0560:POPIAC>2.0.CO;2.

Wang, C.-C., 2014: On the calculation and correction of equitable threat score for model quantitative precipitation forecasts for small verification areas: The example of Taiwan. *Wea. Forecasting,* **29,** 788–798, doi:10.1175/WAF-D-13-00087.1.

——, G. T.-J. Chen, S.-C. Yang, and K. Tsuboki, 2009: Wintertime supercell thunderstorms in a subtropical environment: Numerical simulation. *Mon. Wea. Rev.,* **137,** 2175–2202, doi:10.1175/2008MWR2616.1.

——, ——, and S.-Y. Huang, 2011: Remote trigger of deep convection by cold outflow over the Taiwan Strait in the mei-yu season: A modeling study of the 8 June 2007 case. *Mon. Wea. Rev.,* **139,** 2854–2875, doi:10.1175/2011MWR3613.1.

——, H.-C. Kuo, Y.-H. Chen, H.-L. Huang, C.-H. Chung, and K. Tsuboki, 2012: Effects of asymmetric latent heating on typhoon movement crossing Taiwan: The case of Morakot (2009) with extreme rainfall. *J. Atmos. Sci.,* **69,** 3172–3196, doi:10.1175/JAS-D-11-0346.1.

——, Y.-H. Chen, H.-C. Kuo, and S.-Y. Huang, 2013a: Sensitivity of typhoon track to asymmetric latent heating/rainfall induced by Taiwan topography: A numerical study of Typhoon Fanapi (2010). *J. Geophys. Res. Atmos.,* **118,** 3292–3308, doi:10.1002/jgrd.50351.

——, H.-C. Kuo, T.-C. Yeh, C.-H. Chung, Y.-H. Chen, S.-Y. Huang, Y.-W. Wang, and C.-H. Liu, 2013b: High-resolution quantitative precipitation forecasts and simulations by the Cloud-Resolving Storm Simulator (CReSS) for Typhoon Morakot (2009). *J. Hydrol.,* **506,** 26–41, doi:10.1016/j.jhydrol.2013.02.018.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.,* **136,** 4470–4487, doi:10.1175/2008MWR2415.1.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wu, C.-C., C.-Y. Huang, M.-J. Yang, F.-C. Chien, J.-S. Hong, and T.-H. Yen, 2010: Typhoon Morakot (2009) and a special review on the current status and future challenge of tropical cyclone simulation (in Chinese with English abstract). *Atmos. Sci.,* **38,** 99–134.

Wu, L., J. Liang, and C.-C. Wu, 2011: Monsoonal influence on Typhoon Morakot (2009). Part I: Observational analysis. *J. Atmos. Sci.,* **68,** 2208–2221, doi:10.1175/2011JAS3730.1.

Yamada, H., T. Nasuno, W. Yanase, and M. Satoh, 2013: Westward ''propagating'' typhoons in the tropical western Pacific. Preprints, *Conf. on Mesoscale Convective Systems and High-Impact Weather in East Asia (ICMCS-IX),* Beijing, China, Chinese Academy of Meteorological Sciences, 205.

Yuan, H., S. L. Mullen, X. Gao, S. Sorooshian, J. Du, and H.-M. H. Juang, 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM Ensemble System. *Mon. Wea. Rev.,* **133,** 279–294, doi:10.1175/MWR-2858.1.

Zhang, F., A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting,* **21,** 149–166, doi:10.1175/WAF909.1.