Verification of the Skill of Numerical Weather Prediction Models in Forecasting Rainfall from U.S. Landfalling Tropical Cyclones

BEDA LUITEL[1], GABRIELE VILLARINI[1], GABRIEL A. VECCHI[2]

[1] IIHR-Hydroscience & Engineering, The University of Iowa, Iowa City, Iowa, USA

[2] NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

Manuscript submitted to

Journal of Hydrology

9 February 2016

Revised August 2016

Corresponding author:

Gabriele Villarini, IIHR-Hydroscience & Engineering, The University of Iowa, 306 C. Maxwell Stanley Hydraulics Laboratory, Iowa City, 52242, Iowa, USA. E-mail: gabriele-villarini@uiowa.edu. Tel.: (319) 384-0596

Abstract

The goal of this study is the evaluation of the skill of five state-of-the-art numerical weather prediction (NWP) systems [European Centre for Medium-Range Weather Forecasts (ECMWF), UK Met Office (UKMO), National Centers for Environmental Prediction (NCEP), China Meteorological Administration (CMA), and Canadian Meteorological Center (CMC)] in forecasting rainfall from North Atlantic tropical cyclones (TCs). Analyses focus on 15 North Atlantic TCs that made landfall along the U.S. coast over the 2007-2012 period. As reference data we use gridded rainfall provided by the Climate Prediction Center (CPC). We consider forecast lead-times up to five days. To benchmark the skill of these models, we consider rainfall estimates from one radar-based (Stage IV) and four satellite-based [Tropical Rainfall Measuring Mission - Multi-satellite Precipitation Analysis (TMPA, both real-time and research version); Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN); the CPC MORPHing Technique (CMORPH)] rainfall products. Daily and storm total rainfall fields from each of these remote sensing products are compared to the reference data to obtain information about the range of errors we can expect from "observational data." The skill of the NWP models is quantified: 1) by visual examination of the distribution of the errors in storm total rainfall for the different lead-times, and numerical examination of the first three moments of the error distribution; 2) relative to climatology at the daily scale. Considering these skill metrics, we conclude that the NWP models can provide skillful forecasts of TC rainfall with lead-times up to 48 hours, without a consistently best or worst NWP model.

1. Introduction

North Atlantic tropical cyclones (TCs) are responsible for significant societal and economic impacts. Over the 1900-2005 period, the average annual normalized damage associated with TCs in the continental United States is about $10 billion (value normalized to 2005 monetary value; Pielke et al. 2008). Overall, this damage accounts for close to half (Table 1 in Smith and Katz (2013)) of the total weather and climate disasters over the period of 1981-2011, much more than the damage associated with any other type of weather related disasters.

TCs are associated with multiple hazards, including strong winds, storm surges, heavy rainfall and flooding. While the effects of winds and surge are mostly felt along the coastal areas near the landfall location, heavy rainfall and flooding are responsible for significant damage over much larger areas, even hundreds of kilometers from the coast. More than 50% of the fatalities associated with TCs between 1970 and 2004 were caused by fresh water flooding (http://www.nws.noaa.gov/os/water/ahps/pdfs/InlandFloodBrochure7F.pdf). Over the period 1963-2012, Rappaport (2014) showed that almost 50% of the U.S landfalling TCs have at least one fatality related to rain. Hurricane Ivan (2004) alone accounted for two-thirds of the total flood insurance payments made by the federal government in that year, impacting 23 different states (Czajkowski et al. 2013).

U.S. landfalling TCs are responsible for major flood events over large areas east of the Rocky Mountains, in particular along the eastern and central United States and along the coastal regions on the Gulf of Mexico (Villarini and Smith 2010, 2013, Villarini et al. 2011, 2014). Although precipitation directly associated with TCs is less than 25% of the annual precipitation even in the most affected regions, the impacts can be extremely significant (e.g., Kunkel et al. 2010, Jiang and Zipser 2010, Barlow 2011).

Despite these negative socio-economic impacts, landfalling TCs have also been found to play a significant role as "drought busters" (e.g., Elsberry 2002, Maxwell et al. 2012, 2013, Kam et al. 2013). Torrential rainfall associated with TCs occasionally can have the effect of breaking a prolonged drought by recharging reservoirs and elevating soil moisture. In these situations TC rainfall mitigates one environmental stressor, even as the potential for damage associated with the extreme rainfall, high-speed wind and ocean surge remain (e.g., Kam et al. 2013, Maxwell et al. 2013, Khouakhi and Villarini 2016). Because rainfall associated with TCs has both significant positive and negative impacts on our society, it is critical that we understand how skillful current forecasting systems are in predicting rainfall associated with these storms to help us improve our preparedness and mitigation efforts.

Numerical Weather Prediction (NWP) models provide forecasts of a number of weather-related variables (e.g., precipitation, temperature at different levels) for different lead-times (e.g., Lorenc 1986, Bougeault et al. 2010). However, quantitative information about the skill of NWP models in forecasting TC rainfall is still limited (Marchok et al. 2007, Mohanty et al. 2014). For a skillful prediction of TC rainfall, the models must predict the strength and distribution of the rainfall rate and wind fields together with the track and intensity of the TC system (see Halperin et al. (2013) for a discussion on the genesis forecasting of North Atlantic TCs). Therefore, precipitation forecasts from NWP models in general, and for TCs in particular, are inherently uncertain and subject to three types of error: localization, timing and intensity of precipitation events (e.g., Marchok et al. 2007). In this study, our goal is to evaluate the skill of NWP models in forecasting TC rainfall by quantifying their errors with respect to a reference (rain gauge-based) dataset. Moreover, five additional "observational" (remote sensing-based) datasets are also compared to the reference dataset: the skill of the NWP models in forecasting TC rainfall is

quantified for different lead-times, and discussed and interpreted with respect to the performance of these "observational" products.

In this paper, the description of data and methodology is provided in section 2, followed by results and discussion in section 3. Section 4 summarizes the main points of the study and concludes the paper.

2. Data and Methodology

We use the Climate Prediction Center (CPC) Unified Gauge-Based Analysis of Daily Precipitation over the continental United States. These data represent daily accumulations and are obtained by interpolating rain gage measurements from a number of different networks and sources: the National Oceanic and Atmospheric Administration (NOAA)'s National Climate Data Center (NCDC) daily COOP stations, daily accumulations from hourly precipitation datasets, and the CPC dataset (it includes data from River Forecast Centers and 1st order stations). The spatial resolution is 0.25-decimal degree over the continental United States. There are different quality control steps that are implemented to remove duplicate and overlapping stations, buddy checks are used to eliminate extreme values, and standard deviation checks are used to compare the daily precipitation data against a daily climatology (Higgins et al. 2000). For the North Atlantic TC track information (date, time, latitude and longitude of all recorded storms with a 6-hour resolution) we use the NOAA-Hurricane Research Division's Hurricane Database (HURDAT-2; Landsea and Franklin 2013).

We evaluate the forecast rainfall produced by five state-of-art NWP models: European Centre for Medium-Range Weather Forecasts (ECMWF; Buizza et al. 2007), UK Met Office (UKMO; Bowler et al. 2008), National Centers for Environmental Prediction (NCEP; Toth and Kalnay 1997), China Meteorological Administration (CMA), and Canadian Meteorological

Center (CMC; Houtekamer et al. 2009). Data for NWP models have been archived from the THORPEX Interactive Grand Global Ensemble (TIGGE; Bougeault et al. 2010).

To benchmark the skill of these NWP models, we consider rainfall estimates from five remote sensing products (one ground based radar and four satellite-based rainfall products). Stage IV multi-sensor precipitation dataset is produced by NOAA-NCEP (Lin and Mitchell 2005). It has ~4-km and hourly resolution, and is obtained by merging ground-based radars across the United States, and rain gauge measurements are used to perform bias correction. The four satellite-based rainfall products we use are: Tropical Rainfall Measuring Mission - Multi-satellite Precipitation Analysis [TMPA; both real-time (TMPA_RT) and research version (TMPA_RV); Huffman et al. 2010]; Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN; Sorooshian et al. 2000); CPC MORPHing Technique (CMORPH; Joyce et al. 2004). These satellite-based products have a resolution coarser than Stage IV (3-hourly and 0.25-degree), and TMPA-research version is the only product for which a monthly bias correction with respect to rain gauges is applied.

The rainfall products have different spatial resolutions. Stage IV has the highest spatial resolution (~4km), the four satellite-based remote sensing products and the CPC data have 0.25 × 0.25 degree lat/lon spatial resolution, whereas all the NWP model output we have explored are on a 0.5 degree lat/lon spatial resolution grid (the default resolution in the TIGGE archive). Because our focus is on the NWP models, all the products were regridded into 0.5-degree resolution to establish uniformity in the analysis, with the expectation that both the agreement between observational estimates and the forecasting skill will increase as we coarsen the spatial resolution (e.g., Lavers and Villarini 2013). Results are based only on rainfall over land.

We focus on the evaluation of the precipitation associated with North Atlantic TCs that affected the continental United States over the period 2007–2012, and examine 15 storms that came within 500 km of the coast of the United States. We consider TC-rainfall the rainfall that occurred within a 500-km buffer around the center of circulation of a given storm. We compare the storm total rainfall obtained from the CPC data (our reference) against the rainfall forecasts from the five NWP models. To quantify how close (or far) these forecasts are from the reference data, we also use rainfall estimates from the five remote sensing products to give a range of potentially acceptable results. The availability of these remote sensing products allows us to complement the "absolute" evaluation of the performance of the NWP models with an assessment that is "relative" to what obtained when using "observational" records. For instance, we could use the correlation coefficient as skill metric and obtain a value of 0.4 between CPC and forecast rainfall. While we can interpret this number in an absolute sense (i.e., 0.4 on a scale from -1 to +1), we can also interpret it in a relative sense: the 0.4-correlation value has a different interpretation if the range of values we get from the remote sensing products is between 0.9 and 0.95, rather than between 0.3 and 0.45. Therefore, it will provide us with an additional way of benchmarking the NWP models.

Here we use Hurricane Irene (2011) as an example of our approach and methodology (Figure 1; the results for the other 14 storms are in Supplementary Figures S1-S14). Our approach is to examine the skill of the forecasts starting from the first time the center of circulation of the storm is within 500 km from the U.S. coastline. We will refer to this as the "0-hour lead-time." In the example in **Error! Reference source not found.**, the "0-hour lead-time" represents the storm total rainfall from 27 August 2011 at 12 UTC to 31 August 2011 at 12 UTC. We will refer to the "12-hour lead-time" the forecast for the period from 27 August 2011 at 12

UTC to 31 August 2011 at 12 UTC initialized on 27 August 2011 at 0 UTC. Longer lead-times will follow the same rationale.

We compute rainfall errors with respect to the reference data (CPC) at each 0.5-degree pixel by subtracting CPC rainfall accumulations from the rainfall accumulations obtained from each of the remote sensing products and the NWP models at each 0.5-degree pixel. An examination of the skill of NWP in forecasting storm total TC rainfall is based on both visual and quantitative evaluations. Visual inspection was performed by plotting storm total rainfall and rainfall errors for each of the storms for all the products. Statistical analysis of the rainfall errors was performed by computing the probability density function (pdf) and the first three moments of the error distributions. For each of the statistical measures we used in the process of skill evaluation, we develop envelopes from the five remote sensing products and then use them as the range to quantify and evaluate the skills of the NWP models.

In addition to the analyses focused on the storm total rainfall, we also verify the skill of the NWP models in forecasting TC-daily rainfall. We quantify the accuracy of the forecasts relative to climatology (used as reference) using the mean square error (MSE) skill score SS_{MSE} (e.g., Hashino et al. 2007). The skill of a perfect forecast is equal to 1, with smaller values pointing to a decreasing forecast skill. When the skill value is equal to 0, it means that the forecast accuracy is the same as the one we would have obtained resorting to climatology as our forecast. Accuracy worse than the climatology forecast is represented by negative values.

We can decompose the value of SS_{MSE} into three components (Murphy and Winkler 1992):

$$SS_{MSE} = \rho_{fo}^2 - \left[\rho_{fo} - \frac{\sigma_f}{\sigma_o}\right]^2 - \left[\frac{\mu_f - \mu_o}{\sigma_o}\right]^2 \quad (1)$$

where ρ_{fo} is the correlation coefficient between forecasts observations and quantifies the degree of linear dependence between the two; μ_f and μ_o are the forecast and observation means, respectively; σ_f and σ_o are the forecast and observation standard deviations, respectively.

Based on this decomposition, the correlation coefficient (or its squared counterpart, the coefficient of determination) reflects the forecast accuracy only in the absence of biases, and it represents the potential skill (*PS*). However, without a proper accounting of the potential biases, the forecast skill would be inflated. The second and third terms in the right-hand-side of equation 1 quantify the conditional and unconditional biases, respectively. The former is referred to as slope reliability (*SREL*) and quantifies the departures from the 1:1 line in terms of slope. The unconditional bias (the last term in equation 1) is referred to as the standardized mean error (*SME*). Therefore, the sum of the bias terms *SREL* and *SME* quantifies the differences between potential (*PS*) and actual (*SS*) skill, in a sense representing the "room for improvement" by the forecast system (e.g., Boer et al. 2013, Younas and Tang 2013).

3. Results

We start the evaluation of the skill of the NWP models based on the visual examination of the storm total rainfall fields and on the quantitative analysis of the error characteristics. As an example of the type of analyses we have performed, we focus on the results for Hurricane Irene (2011) (the results for the other 14 storms are in Supplementary Figures S1-S14). According to the NHC report (http://www.nhc.noaa.gov/data/tcr/AL092011_Irene.pdf), Irene caused 41 direct deaths in the United States, among which six fatalities were due to storm surge, 15 were related to high wind and 21 were due to rainfall-induced floods. Based on the National Flood Insurance Program, the NHC reported that the total damage caused by the Irene was $15.8 billion out of which $7.2 billion of the damage was related to inland flooding and surge only (Avila and

Cangialosi 2011, McCallum et al. 2012). Hurricane Irene affected large areas of the eastern United States from North Carolina to Maine (Figure 1).

Figure 2 summarizes the total rainfall accumulation during Hurricane Irene according to the five remote sensing products and the five NWP models (the results for the other 14 storms are in Supplementary Figures S15-S28). These results show that there is a relatively large range of variability in the "observational products" (top row in Figure 2). As a first step, each panel in Figure 2 is visually compared with CPC rainfall accumulation (Figure 1). This visual comparison provides qualitative information about the capability of each of these products in estimating rainfall associated with TCs. Overall, Stage IV is the product that more closely resembles the observational data both in terms of magnitude and location of the areas with the largest rainfall accumulations (see also Villarini et al. (2011)). On the other hand, the satellite-based estimates tend to have smaller rainfall values generally spread over larger areas (this is particularly true for PERSIANN). As mentioned before, we will use this variability in rainfall estimates from observational systems (ground- and space-based sensors) as a way of bounding what we can consider acceptable for the NWP models: we will deem as satisfactory NWP rainfall forecasts that are within the range of outcomes from the remote sensing products.

The TC-rainfall forecasts appear to capture reasonably well the observed storm total rainfall both in location and magnitude. This is particularly true for the shortest lead-time, with the performance decreasing as we increase the lead-time to five days. As mentioned before, this has to do with the fact that we are expecting the models to not only correctly forecast the rainfall fields around the center of circulation of the storms, but also to correctly track these storms (e.g., Marchok et al. 2007). This issue is clear, for instance, for the 5-day lead-time, where the models

were not able to correctly forecast the storm track. Our expectations of the rainfall forecasts from the NWP models are admittedly very high, but necessary to improve our confidence in them.

Furthermore, we compute rainfall errors using an additive formulation by subtracting the storm total rainfall from CPC from each of the remote sensing products and NWP models. The results in Figure 3 are for Hurricane Irene, while those for the other 14 TCs are presented in Figures S29 – S42. Stage IV shows the smallest discrepancies with respect to CPC, likely due to the bias correction performed using rain gauges. On the other hand, the results for the four satellite-based products show areas with consistent over- and under-estimation, similar to what observed for the NWP models. For TMPA and CMORPH, our results are generally consistent with Yu et al. (2009), Villarini et al. (2011), Chen et al. (2013) and Deo et al. (2016), who found an over- (under-) estimation of the TC-rainfall in areas characterized by lower (higher) rain rates. On the other hand, for the forecasting models these patterns can be likely ascribed to errors in the forecast tracks. The magnitude of the differences between CPC and the NWP models increases as we increase the lead-time, with results up to the 2-day lead-time that are comparable to the range of accuracy (both quantity and location wise) of the "observational" products. That is, the NWP models up to the 48-hour lead-time were at least as skillful at predicting rainfall as the least accurate remote sensing product we considered. These findings are generally consistent across all the 15 TCs considered in this study. Among the NWP models, we did not see any one model performing consistently best or worst for all lead-times.

A more quantitative examination of the skill of NWP models is achieved by computing the pdf of the rainfall errors from each product. Figure 4 shows the pdf of the rainfall errors for the ECMWF model for all the storms, together with the results obtained using the satellite products. For a given lead-time, if the pdf from any NWP model and lead-times falls within the

range of variability from the observational products (grey envelope in Figure 4), we can infer that the model at that lead-time was capable of predicting the TC rainfall with a skill comparable to the range of results from our five "observational" products. Taking into account all the challenges associated with the forecasting of TC rainfall (i.e., correctly forecasting both the storm track and the rainfall around the center of circulation of these storms), the results in Figure 4 are rather promising. Overall, the error distributions for these 15 TCs are comparable to the error distributions from the remote sensing products. Moreover, the distribution are highly "peaked" with most of the values concentrated around 0. This is particularly true for the shortest lead-times, with the results for the 5-day lead-time that appear to be much smoother than the others. These results are consistent across all NWP models (Supplementary Figures S43-S46)

In addition to the entire error distribution, we have also computed the first three moments (mean, standard deviation and skewness) of the error distributions, and compared them across NWP models and remote sensing products (Figure 5). A models is considered to have better performance if the expected value of the errors is close to zero (unbiased). In the left column of Figure 5, the expected values of the errors are closer to zero for the ECMWF, followed by UKMO and NCEP. In general, we would consider a "successful" forecast one for which the statistical properties of the error are within what we obtain from the remote sensing products. For instance, in the case of Superstorm Sandy (2012), the average error for UKMO, ECMWF and CMC is close to zero across different lead-times, with error characteristics that are almost better than what we obtain from the remote sensing products. More generally, with the exception of Tropical Storm Lee (2011) and Tropical Storm Debby (2012), the average errors are close to zero (or at least within the uncertainties from the remote sensing products) up to the 2-day lead-time.

The standard deviations of the errors from the NWP models show, in general, more variability than what obtained from the remote sensing products (Figure 5, middle column). The main exception is represented by NCEP, which shows variability that is comparable with that of the observational products. Therefore, these results indicate that the distribution of the errors, while generally unbiased for the NWP models, tends to be flatter than what we would expect from a range of remote sensing products.

Finally, the skewness of the errors from the NWP models is well within the results obtained from the observational data across the different models and lead-times (Figure 5, right column). Depending on the TC, NWP model or lead-time, the error distributions tend to be skewed, in particular negatively skewed. Therefore, our quantitative evaluation of the forecast errors suggests that the NWP models are capable of representing the error structure obtained from remote sensing products, in particular for lead-times up to two days, even though the error distribution tends to be flatter than what observed from the observational datasets.

Up to this point we have focused on the evaluation of the skill of the NWP models in forecasting storm total rainfall. We have also performed analyses quantifying the skill of these models in forecasting TC rainfall at the daily scale based on the decomposition in equation 1 (Figure 6). The potential skill PS of the remote sensing products is generally very high, with values ranging from 0.6 to 0.8 for most of the storms (consistent with Chen et al. (2013), who evaluated the daily TMPA measurements for TCs affecting Australia between 1998 and 2011). The values of PS for the NWP models are lower, in particular at the longer lead times, with the results for the shorter lead-times (~up to 1 day) that are within the bounds from the remote sensing products. This is particularly true for the CMC model, for which the PS values for the 2- and 5-day forecasts are much lower than those for the shorter lead times. Among the five NWP

models, NCEP has overall larger potential skill and weaker dependence on lead-time compared to the other four NWP models. In interpreting these results, it is worth recalling that PS represents the coefficient of determination: this means that a value of 0.25 corresponds to a correlation coefficient of 0.5. Therefore, the overall skill of these models is rather promising.

The second and third columns in Figure 6 quantify conditional and unconditional biases. Overall, the values of SREL and SME are very small for the remote sensing products, indicating that the observed variability with respect to CPC tends to be around the 1:1 line. The results for the NWP models are comparable with the remote sensing products, with small values of SREL and SME. Given the limited biases, the values of the skill score SS and potential skill for the "observational" products are very similar and generally larger than 0.6. On the other hand, the NWP models exhibit positive values of SS in the majority of the cases, in particular for the shorter lead times. This indicates that these models are skillful compared to climatology (used as reference), because of a combination of small biases and good potential skill.

Overall, we have tried to present some general findings that are shared by the remote sensing products and the NWP models across the different storms (e.g., Stage IV is the product that most closely resembles the observations; the skill of the NWP models decreases for longer lead times); however, it is worth highlighting that there was variability across storms as well (see figures in the Supplementary Material). It is likely that some of these differences are due to the different size of the TC rain shield, the amount of rainfall associated with them, and to the skill of the NWP models in forecasting the track of the storms. Future studies based on a larger number of TCs should examine the dependence of the forecast skill on these and other factors.

All the analyses so far have focused on 15 North Atlantic TCs coming at least within 500 km of the U.S. coastline during the 2007-2012. Here we include some results related to the skill of

NWP models in forecasting rainfall associated with a more recent storm, Hurricane Joaquin (2015), which was indirectly responsible for catastrophic flooding in South Carolina (the center of circulation was never less than 500 km from the U.S. coastline). On 1 October 2015 a cut-off low developed over the southeastern United States, pulling moisture from this hurricane, leading to multiple days of heavy rainfall and major flooding in South Carolina (consult Berg (2016) for more details). As shown in Figure 7, large coastal areas of South Carolina experienced rainfall in excess of 13 inches, leading to major flooding in Charleston and Columbia. All of the remote sensing products capture the large rainfall amounts in South Carolina (Figure 8), even though Stage IV is the product that once again performs the best with respect to CPC; the four satellite-based datasets underestimate the rainfall amounts along the U.S. East Coast, with underestimations in excess of 5 inches in South Carolina and overestimations in northern Florida and Alabama (Supplementary Figure S47). The location of the high rainfall amounts was identified by all of the NWP models at the shortest lead times (i.e., up to one day), even though the forecasts at the 2- and 5-day lead times generally did not correctly forecast the areas at the highest risk of high rainfall from this storm.

4. Conclusion

In this study we have examined the skill of five state-of-the-art NWP models [European Centre for Medium-Range Weather Forecasts (ECMWF), UK Met Office (UKMO), National Centers for Environmental Prediction (NCEP), China Meteorological Administration (CMA), and Canadian Meteorological Center (CMC)] in forecasting rainfall associated with 15 U.S. landfalling TCs during the 2007-2012 period. These forecasts with a lead-time up to five days were compared against gridded rain gauge based measurements. We used rainfall estimates from

five remote sensing products as a way of quantifying the fidelity we can expect from different observational datasets. The main findings of this study can be summarized as follows:

- Among the remote sensing products, Stage IV showed the closest resemblance with the CPC dataset, likely because of the use of rain gauge information for bias correction. Among other remote sensing products, PERSIANN consistently underestimated the TC rainfall.
- Overall, the performance of the NWP models was comparable to the difference between remote sensing based products and the reference dataset, particularly at shorter lead-times. While quantitatively the forecasts were comparable to the reference data, a substantial error in the precipitation predictions appears to have been related to predictions of the storm track itself. Track-relative analyses as those performed in Marchok et al. (2007) would alleviate this issue, but were not performed here because of the lack of forecast track.
- We have quantified the accuracy of the daily forecasts, decomposing the mean square error skill score into potential skill, conditional and unconditional biases. The NWP models exhibit small biases and good potential skill, especially at the shortest lead times, leading to an overall positive skill score (i.e., the quality of the forecasts is better than climatology, used as reference).
- While we have tried to draw general conclusions about the skill of the NWP models, it is worth highlighting that there is inter-storm variability. Future studies should examine the dependence of these conclusions on different TC characteristics (e.g., size, translational velocity).
- In interpreting these results, it is important to remember that we are evaluating the performance for a very challenging target: we are asking the NWP models not only to correctly forecast the storm track, but also to correctly characterize the rainfall distribution

around this track. When we consider that the errors from the remote sensing products ("observations") are comparable to the forecast errors, the results from the NWP are rather encouraging. Overall, our findings indicate that skillful forecasting of storm total rainfall associated with North Atlantic TCs making landfall along the U.S. coast is possible up to two days prior to landfall.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants AGS-1262091 and AGS-1262099, and the National Oceanic and Atmospheric Administration Award number NA14OAR4830101 to the Trustees of Princeton University. Gabriele Villarini also acknowledges support from the USACE Institute for Water Resources. The authors thank the five weather prediction centers that provided the data used herein and the TIGGE archive at the ECMWF. The comments and suggestions by the associate editor and two anonymous reviewers are gratefully acknowledged.

References

Avila, L. A., and J. Cangialosi, 2011: Tropical Cyclone Report: Hurricane Irene. National Hurricane Center.

Barlow, M., 2011: Influence of hurricane-related activity on North American extreme precipitation. Geophysical Research Letters, 38.

Berg, R., 2016: Hurricane Joaquin (AL112015), National Hurricane Center Tropical Cyclone Report (http://www.nhc.noaa.gov/data/tcr/AL112015_Joaquin.pdf).

Boer, G. J., V. V. Kharin, and W. J. Merryfield, 2013: Decadal predictability and forecast skill. Climate Dynamics, 41(7-8), 1817–1833.

Bougeault, P., and Coauthors, 2010: The THORPEX interactive grand global ensemble. Bulletin of the American Meteorological Society, 91, 1059-1072.

Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The MOGREPS short-range ensemble prediction system. Quarterly Journal of the Royal Meteorological Society, 134, 703-722.

Buizza, R., J. R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007: The new ECMWF VAREPS (variable resolution ensemble prediction system). Quarterly Journal of the Royal Meteorological Society, 133, 681-695.

Chen, Y., E. E. Ebert, K. J. E. Walsh, and N. E. Davidson, 2013: Evaluation of TMPA 3B42 daily precipitation estimates of tropical cyclone rainfall over Australia, Journal of Geophysical Research, 118, 11,966–11,978, doi:10.1002/2013JD020319.

Czajkowski, J., G. Villarini, E. Michel-Kerjan, and J. A. Smith, 2013: Determining tropical cyclone inland flooding loss on a large scale through a new flood peak ratio-based methodology. Environmental Research Letters, 8, 044056.

Deo, A., K.J.E. Walsh, and A. Peltier, 2016: Evaluation of TMPA 3B42 precipitation estimates during the passage of tropical cyclones over New Caledonia, Theoretical and Applied Climatology, doi: 10.1007/s00704-016-1803-0.

Elsberry, R. L., 2002: Predicting hurricane landfall precipitation: Optimistic and pessimistic views from the symposium on precipitation extremes. Bulletin of the American Meteorological Society, 83, 1333-1339.

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

Hashino, T., A.A. Bradley, and S.S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts, Hydrology and Earth System Sciences Discussions, 3(2), 561-594.

Higgins, R., W. Shi, E. Yarosh, and R. Joyce, 2000: Improved US precipitation quality control system and analysis. NCEP/Climate Prediction Center ATLAS No. 7. Camp Springs,

Maryland. available at http://www. cpc. ncep. noaa. gov/research_papers/ncep_cpc_atlas/7/index. html.

Houtekamer, P., H. L. Mitchell, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. Monthly Weather Review, 137, 2126-2143.

Huffman, G. J., R. F. Adler, D. T. Bolvin, and E. J. Nelkin, 2010: The TRMM multi-satellite precipitation analysis (TMPA). Satellite rainfall applications for surface hydrology, Springer, 3-22.

Jiang, H., and E. J. Zipser, 2010: Contribution of tropical cyclones to the global precipitation from eight seasons of TRMM data: Regional, seasonal, and interannual variations. Journal of climate, 23, 1526-1543.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. Journal of Hydrometeorology, 5, 487-503.

Kam, J., J. Sheffield, X. Yuan, and E. F. Wood, 2013: The influence of Atlantic tropical cyclones on drought over the eastern United States (1980–2007). Journal of Climate, 26, 3067-3086.

Khouakhi, A., and G. Villarini, 2016: Attribution of annual maximum sea levels to tropical cyclones at the global scale. International Journal of Climatology, DOI: 10.1002/joc.4704 (in press).

Kunkel, K.E., D.R. Easterling, D.A.R. Kristovich, B. Gleason, L. Stoecker, and R. Smith, 2010: Recent increase s in U.S. heavy precipitation associated with tropical cyclones. Geophysical Research Letters, 37, L24706,doi:10 .1029/2010 GL045164.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. Monthly Weather Review, 141, 3576-3592.

Lavers, D.A., and G. Villarini, 2013: Were global numerical weather prediction systems capable of forecasting the extreme Colorado rainfall of 9–16 September 2013? Geophysical Research Letters, 40, 6405–6410, doi:10.1002/2013GL058282.

Lin, Y., and K. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. Preprints, 19th Conf. on Hydrology, San Diego, CA, Am. Meteorol. Soc., 1.2.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. Royal Meteorological Society, Quarterly Journal, 112, 1177-1194.

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. Weather and Forecasting, 22, 726–746.

Maxwell, J. T., P. T. Soulé, J. T. Ortegren, and P. A. Knapp, 2012: Drought-busting tropical cyclones in the southeastern Atlantic United States: 1950–2008. Annals of the Association of American Geographers, 102, 259-275.

Maxwell, J. T., J. T. Ortegren, P. A. Knapp, and P. T. Soulé, 2013: Tropical cyclones and drought amelioration in the Gulf and Southeastern coastal United States. Journal of Climate, 26, 8440-8452.

McCallum, B. E., J. A. Painter, and E. R. Frantz, 2012: Monitoring inland storm tide and flooding from Hurricane Irene along the Atlantic Coast of the United States, August 2011: U.S. Geological Survey Open-File Report 2012–1022, 28 p., available at http://pubs.usgs.gov/of/2012/1022/.

Mohanty, U., K. K. Osuri, R. Nadimpalli, and S. Gopalakrishnan, 2014: Uncertainty in rainfall prediction of land-falling tropical cyclones over India: Impact of data assimilation. 3rd International Workshop on Tropical Cyclone Landfall Processes (IWTCLP-III), Jeju, 8-10.

Murphy, A.H., and R.L. Winkler, 1992: Diagnostic verification of probability forecasts, International Journal of Forecasting, 7(4), 435–455.

Pielke Jr, R. A., J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Musulin, 2008: Normalized hurricane damage in the United States: 1900–2005. Natural Hazards Review.

Rappaport, E. N., 2014: Fatalities in the United States from Atlantic tropical cyclones: new data and interpretation. Bulletin of the American Meteorological Society, 95, 341-346.

Smith, A. B., and R. W. Katz, 2013: US billion-dollar weather and climate disasters: data sources, trends, accuracy and biases. Natural hazards, 67, 387-410.

Sorooshian, S., K.-L. Hsu, X. Gao, H. V. Gupta, B. Imam, and D. Braithwaite, 2000: Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. Bulletin of the American Meteorological Society, 81, 2035-2046.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. Monthly Weather Review, 125, 3297-3319.

Villarini, G., and J. A. Smith, 2010: Flood peak distributions for the eastern United States. Water Resources Research, 46.

——, 2013: Flooding in Texas: Examination of temporal changes and impacts of tropical cyclones. JAWRA Journal of the American Water Resources Association, 49, 825-837.

Villarini, G., R. Goska, J. A. Smith, and G. A. Vecchi, 2014: North Atlantic tropical cyclones and US flooding. Bulletin of the American Meteorological Society, 95, 1381-1388.

Villarini, G., J. A. Smith, M. L. Baeck, T. Marchok, and G. A. Vecchi, 2011: Characterization of rainfall distribution and flooding associated with US landfalling tropical cyclones: Analyses of Hurricanes Frances, Ivan, and Jeanne (2004). Journal of Geophysical Research: Atmospheres (1984–2012), 116.

Younas, W., and Y. Tang, 2013: PNA predictability at various time scales. Journal of Climate, 26(22), 9090–9114.

Yu, Z., H. Yu, P. Chen, C. Qian, and C. Yue, 2009: Verification of tropical cyclone–related satellite precipitation estimates in Mainland China, Journal of Applied Meteorology and Climatology, 48, 2227-2241.

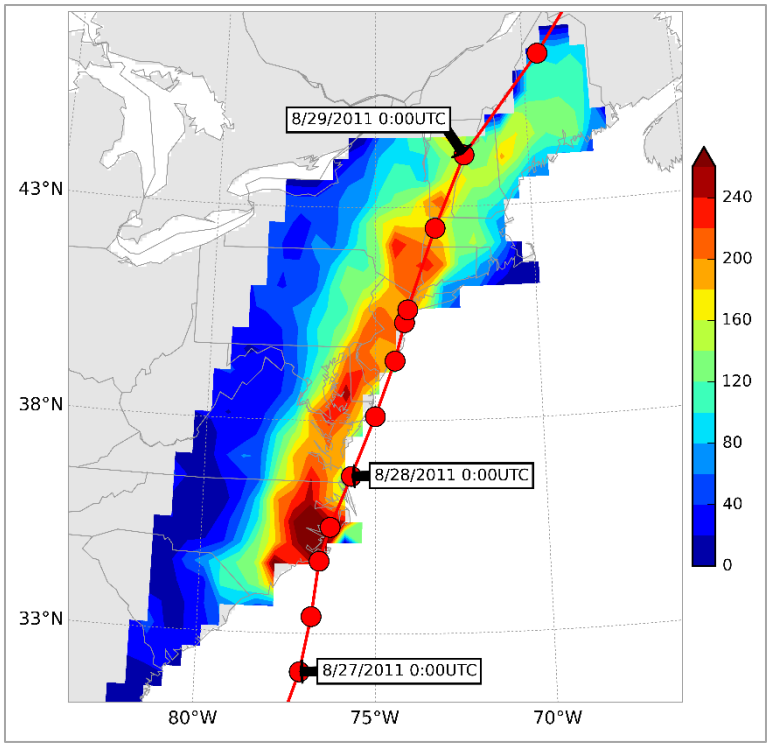

Figure 1: Rainfall accumulation (in mm) for Hurricane Irene (August 26-30 2011) based on the CPC dataset. The red curve with circles represents the storm track.

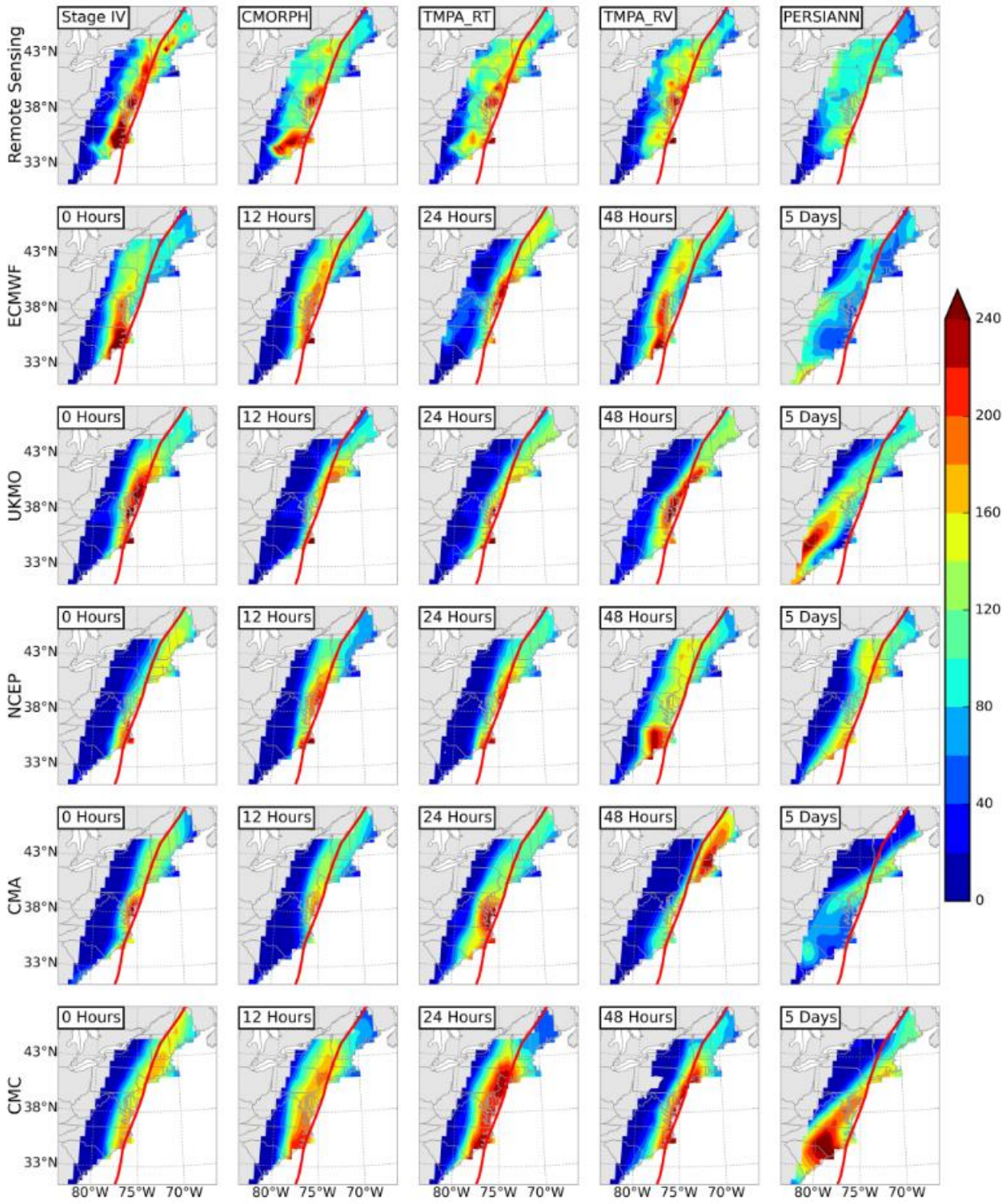

Figure 2: Rainfall accumulation (in mm) for Hurricane Irene (August 26-30 2011) based on remote sensing (top row) and NWP models (second to last rows). For the NWP models, the lead-time increases left to right from 0 hour to 5 days. The red lines represent the storm track.

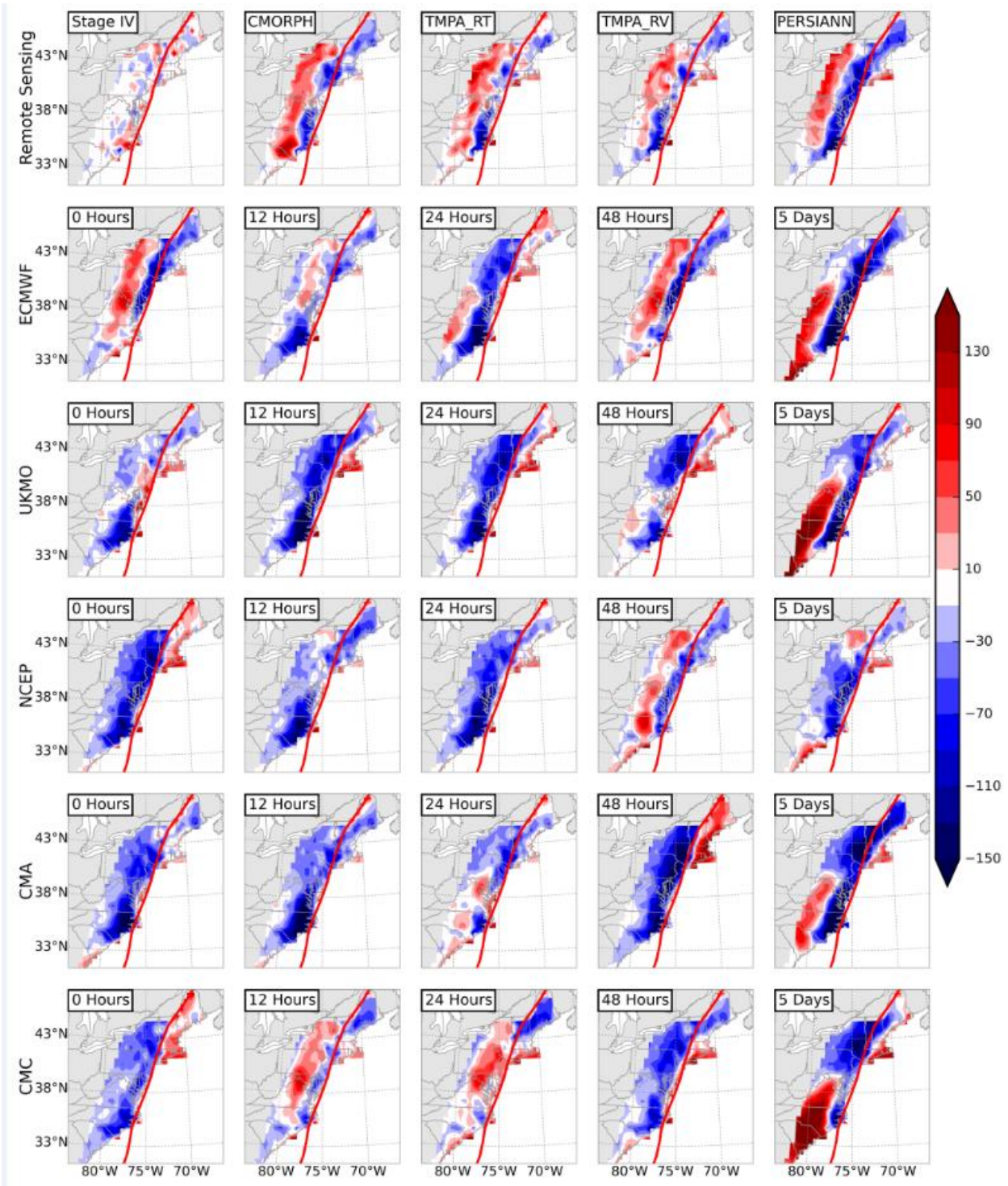

Figure 3: Rainfall errors (in mm) for Hurricane Irene (August 26-30 2011) based on remote sensing (top row) and NWP models (second to last rows) The errors are computed with respect to the CPC data. For the NWP models, the lead-time increases left to right from 0 hour to 5 days. The red lines represent the storm track.

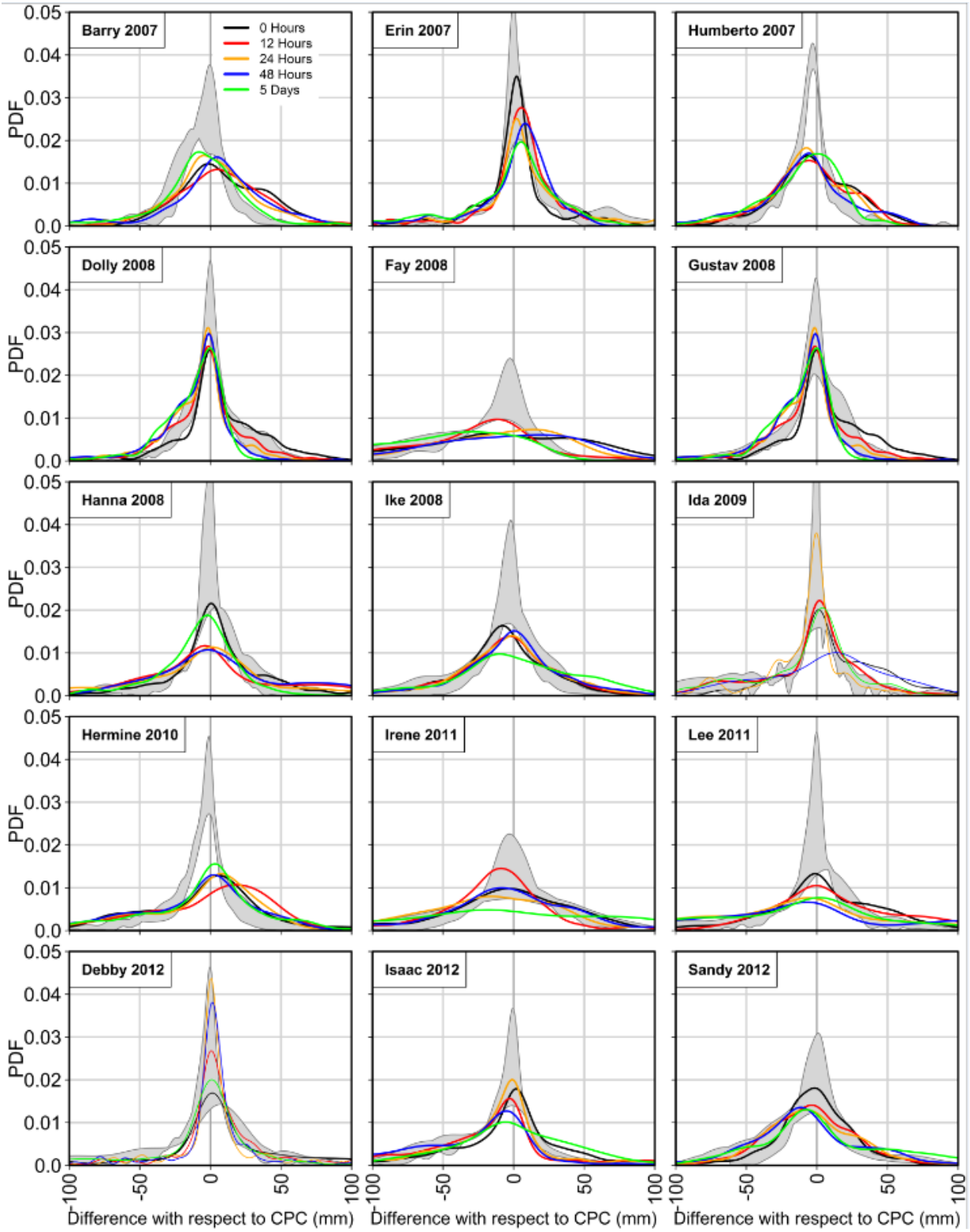

Figure 4: Probability density functions (pdfs) of rainfall errors exhibited by the ECMWF model for different forecast lead-times. The gray shaded envelopes represent the range of pdfs obtained from the rainfall errors of the five remote sensing products.

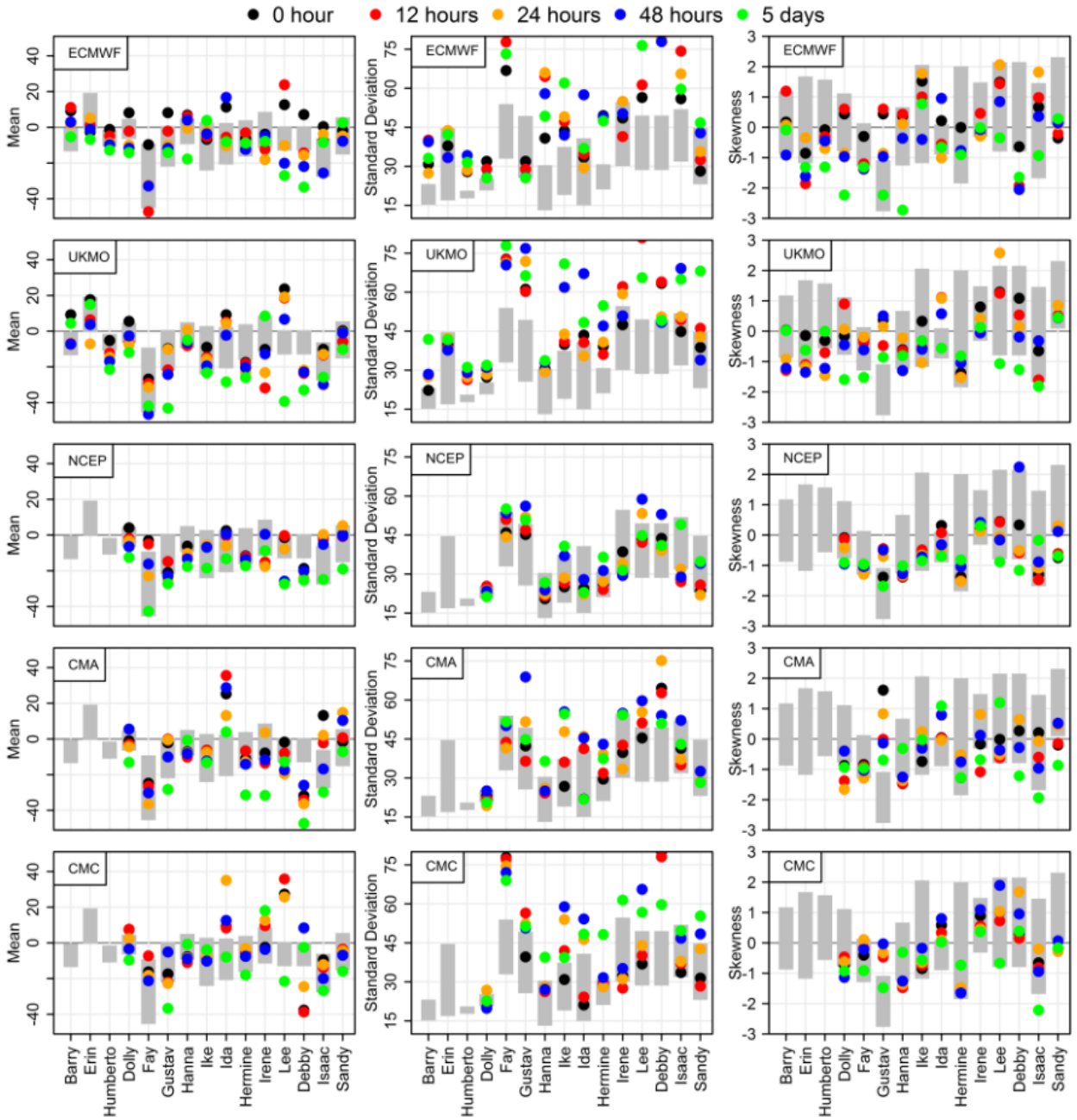

Figure 5: Mean (in mm; left column), standard deviation (in mm; middle column) and skewness (right column) of the rainfall errors for different lead-times and NWP models. The gray rectangular boxes in the background represent the range of values for each statistic based on the five remote sensing products. No forecast information for NCEP, CMA, and CMC was available for Tropical Storm Barry (2007), Tropical Storm Erin (2007) and Hurricane Umberto (2007).

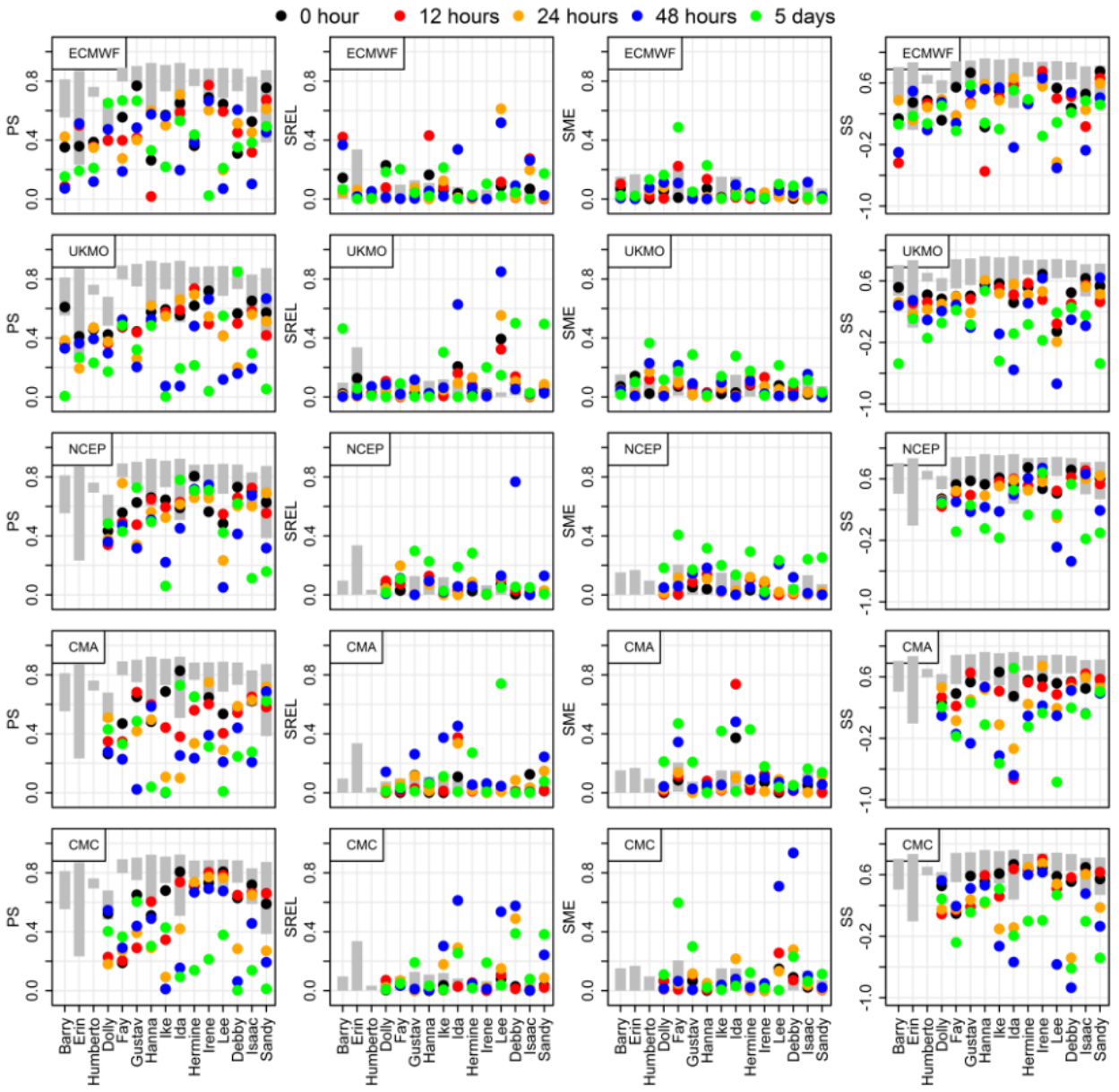

Figure 6: Values of the potential skill (PS; first column), slope reliability (SREL; second column), standardized mean error (SME; third column), and skill score (SS; fourth column). These results are computed with respect to CPC at the daily scale for different lead-times and NWP models. The gray rectangular boxes in the background represent the range of values for each metric based on the five remote sensing products. No forecast information for NCEP, CMA, and CMC was available for Tropical Storm Barry (2007), Tropical Storm Erin (2007) and Hurricane Umberto (2007).

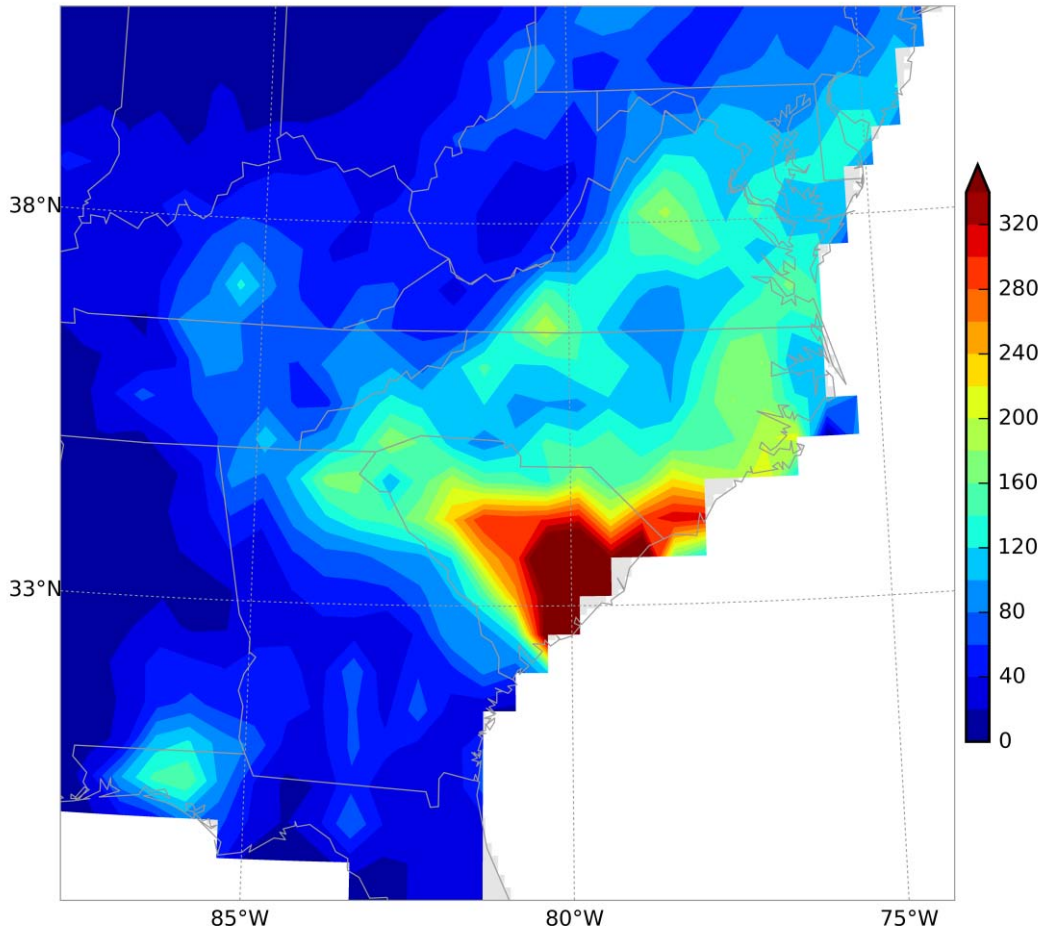

Figure 7: Rainfall accumulation (in mm) for Hurricane Joaquin (September 29 - October 7 2015) based on the CPC dataset.

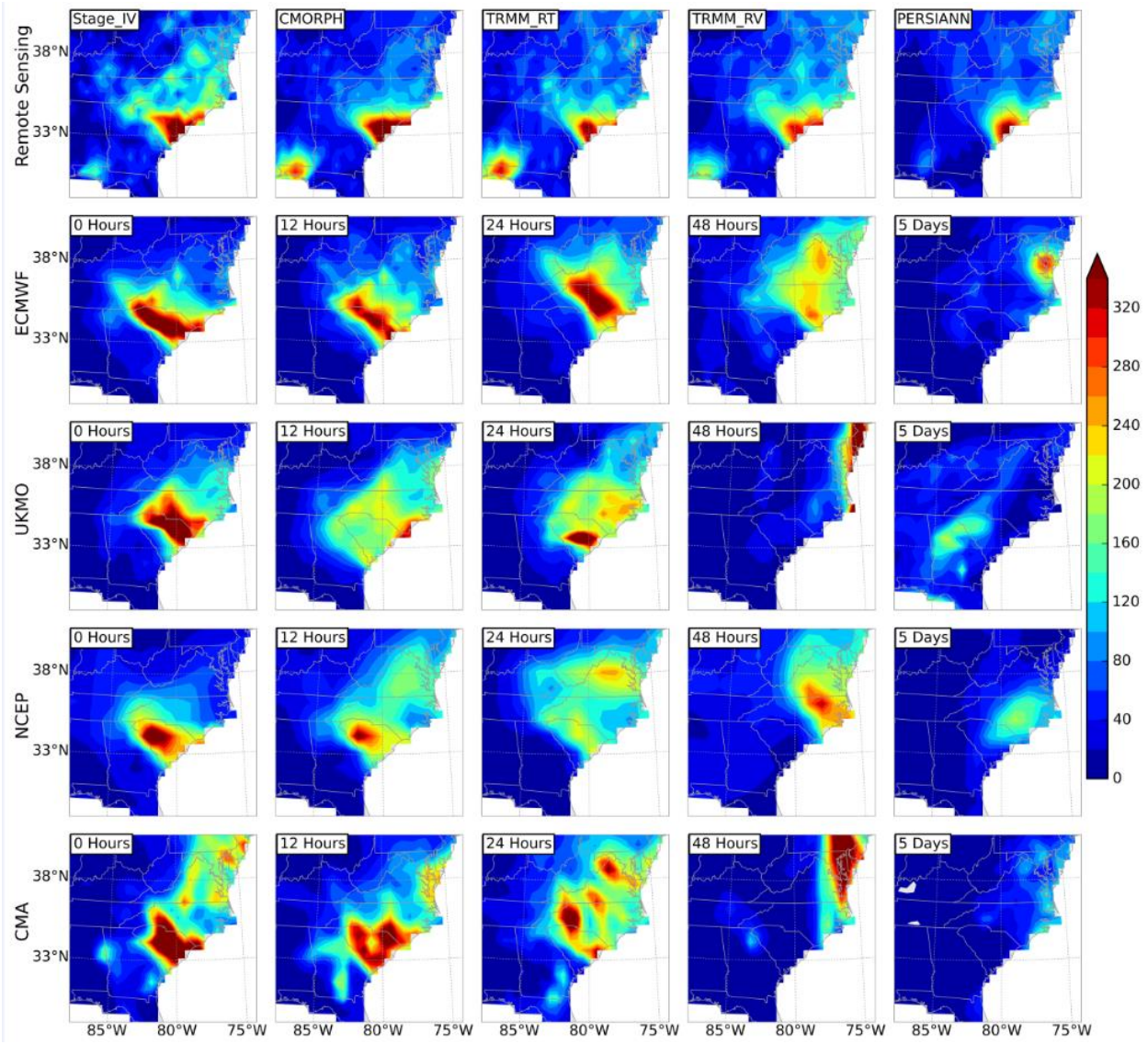

Figure 8: Rainfall accumulation (in mm) for Hurricane Joaquin (September 29 - October 7 2015) based on remote sensing (top row) and NWP models (second to last rows). For the NWP models, the lead-time increases left to right from 0 hour to 5 days.