How Well Can the Met Office Unified Model Forecast Tropical Cyclones in the Western North Pacific?

CHRIS J. SHORT AND JON PETCH
Met Office, Exeter, Devon, United Kingdom

(Manuscript received 30 May 2017, in final form 2 November 2017)

ABSTRACT
Convection-permitting numerical weather prediction models are a key tool for forecasting tropical cyclone (TC) intensities, intensity changes, and precipitation. The Met Office has been routinely running a regional (4.4-km grid spacing), explicit convection version of its Unified Model (UM) over the Philippines since August 2014, driven by its operational global model. The principal aim of this study is to assess the performance of this model relative to the driving global model. By evaluating over a year’s worth of operational TC forecasts, it is shown that the Philippines regional model offers clear benefits for TC forecasting compared with the Met Office global model. In particular, it provides much improved predictions for the intensities of strong storms (category 3 and above) and can successfully capture some rapid intensification (RI) events, whereas the global model cannot predict RI at all. The spatial location of rainfall within intense TCs is also more skillfully predicted by the regional model, and the statistical distribution of rain rates is closer to that observed. Although the regional model adds value, notable biases are also identified, highlighting areas for future work to develop and improve the model.

1. Introduction
Tropical cyclones (TCs) are one of the most dangerous natural hazards. The strong winds, heavy rainfall, and storm surges associated with these storms claim many lives and cause enormous damage every year. Reducing their socioeconomic impact requires accurate forecasts of storm track, intensity, size, and precipitation, days prior to landfall.

In the last few decades, the ability of global numerical weather prediction (NWP) models to predict TC tracks has improved dramatically (e.g., Yamaguchi et al. 2017). This is because storm track is primarily governed by the large-scale environment, the representation of which in models has steadily improved through advancements in numerical techniques, observing systems, and data assimilation (DA) methods. For example, the global configuration of the Met Office Unified Model (UM) used for operational NWP at the Met Office now has day 3 TC track errors comparable to day 1 errors in 1996 (Heming 2016).

The rate of improvement of TC intensity forecasts has been slower (e.g., DeMaria et al. 2014 and references therein), because intensity changes are the result of a complex interplay between physical processes operating across a range of length and time scales (e.g., moist convection, radiative transfer, cloud microphysics, fluxes of enthalpy and momentum across the air–sea interface, environmental wind shear, and interaction with synoptic features such as tropical upper-tropospheric troughs). This inherently nonlinear, multiscale problem poses a major challenge for NWP models.

At present, the most advanced operational global NWP models have a horizontal grid length of ~10 km, which is insufficient to resolve the small-scale processes influencing storm development. Resolution studies conducted with regional models suggest a grid spacing of less than 5 km is required, with 1–2 km preferred, to give an accurate representation of TC intensity and structure (Chen et al. 2007; Fierro et al. 2009; Gentry and Lackmann 2010; Gopalakrishnan et al. 2011; Davis et al. 2011; Gopalakrishnan et al. 2012; Jin et al. 2014). At these scales, convection is (partially) resolved, and thus the convection parameterization scheme used in global models is often switched off (or at least restricted in some way), which has been shown to improve precipitation forecasts in a variety of weather regimes (Done et al. 2004; Kain et al. 2006; Weisman et al. 2008; Lean et al. 2008; Weusthoff et al. 2010; Clark et al. 2010).
Several regional modeling systems for TCs are currently operational. The Hurricane Weather Research and Forecast (HWRF) Model (Biswas et al. 2016) is run at the Environmental Modeling Center (EMC) at the National Centers for Environmental Prediction (NCEP), producing real-time forecasts for all global TC basins. Operational TC forecasts for the Atlantic and eastern and western North Pacific basins are also provided by the Coupled Ocean–Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS-TC) model (Doyle et al. 2011), run at the Fleet Numerical Meteorology and Oceanography Center (FNMOC). Both of these models employ storm-following nests, with innermost grid spacings of 2 and 5 km, respectively. The Japan Meteorological Agency (JMA) also runs a 5-km fixed-grid regional model—the JMA Mesoscale Model (Saito et al. 2006)—covering part of the western North Pacific.

Since August 2014, the Met Office has been routinely running a convection-permitting regional configuration of the UM (i.e., with no deep convection scheme) over the Philippines area, nested inside its operational global model. The Philippines is the most exposed country in the world to TCs, with an average of 19 storms occurring in the Philippine area of responsibility (PAR) every year, 9 of which make landfall (Cinco et al. 2016). The model was set up as part of a continuing collaboration between the Met Office and the Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA), designed to provide PAGASA with real-time, detailed forecast information to aid in the preparation of effective warnings during TC events.

The aim of this paper is to assess how well the Philippines regional model is able to predict TCs and consider the added value relative to the driving global model. To do this, a sample constructed from over a year’s worth of operational TC forecasts is analyzed, representing one of the largest collections of convection-permitting TC forecasts in existence [the samples presented in Tallapragada et al. (2015a, 2016) are notable exceptions]. This is the first systematic evaluation of a regional configuration of the UM for TC forecasting. Section 2 describes the numerical models and data sample used in this study, with model evaluation results presented in section 3, focusing on storm intensity, track, and precipitation. Section 4 provides a summary and concluding remarks.

2. Methodology

a. Numerical models

The UM is used operationally at the Met Office to produce global and regional deterministic and ensemble forecasts. It solves the full, deep-atmosphere, nonhydrostatic, Navier–Stokes equations using a semi-implicit, semi-Lagrangian numerical scheme. The dynamical core of the model is described in detail in Wood et al. (2014). Model prognostic fields are discretized on to a regular latitude–longitude grid with Arakawa C-grid staggering (Arakawa and Lamb 1977), while the vertical discretization utilizes Charney–Phillips staggering (Charney and Phillips 1953) and a terrain-following hybrid-height vertical coordinate.

1) GLOBAL MODEL

The configuration of the UM currently used for global NWP at the Met Office is known as Global Atmosphere 6.1 (GA6.1; Walters et al. 2017). A comprehensive set of physics parameterizations are included in GA6.1, which is very briefly summarized in the appendix (Table A1). The reader is referred to Walters et al. (2017) (and references therein) for further details.

The model grid spacings are 0.234° and 0.156° in the zonal and meridional directions, respectively (approximately 26 km × 17 km in the tropics). In the vertical, there are a total of 70 levels up to a fixed model lid 80 km above mean sea level (MSL). Levels are quadratically spaced to give more levels near the surface. The semi-Lagrangian nature of the dynamical solver permits a relatively long model time step of 450 s.

2) PHILIPPINES REGIONAL MODEL

The Philippines regional configuration of the UM is one-way nested inside the global model (i.e., there is no feedback on the global flow fields), which supplies lateral boundary condition (LBC) data at a 3-hourly frequency. The UM LBC framework is described in Davies (2014). Typically, errors arising from LBCs are only a tiny part of the overall model error (Davies 2014). There is no DA in the regional model itself; initial conditions are derived by interpolating global model analyses onto the finer regional grid, taking account of the higher-resolution surface fields (such as orography and land use) of the regional model.

The horizontal grid spacing of the Philippines model is 0.04° (approximately 4.4 km) in both directions. The model domain and topography are shown in Fig. 1. The domain has a large extent to the east of the Philippines so that TCs that develop over the western North Pacific Ocean and travel northwest toward the islands are captured in the high-resolution domain long before making landfall [note that almost half of the TCs occurring in the PAR form within the PAR: Cinco et al. (2016)]. Hereafter, the Philippines regional model will be referred to as PHI.

The vertical-level set used in the PHI model has 80 levels, the spacing of which increases quadratically with...
height up to a fixed lid 38.5 km MSL. This level set has been developed specifically for tropical regions, with improved resolution in the tropical upper troposphere (and hence improved resolution for tropical deep convection) compared to the level set used in the Met Office’s U.K. and European regional models. The model time step is 120 s.

The increased horizontal resolution of the PHI model relative to GA6.1 warrants a different set of physical parameterizations, summarized in the appendix (Table A1). The most important difference between the models is the treatment of convection. As described in Walters et al. (2017), GA6.1 includes a mass-flux convection scheme based on Gregory and Rowntree (1990), but with many extensions. Convection is only partially resolved at a grid length of 4.4 km, so the PHI model employs a modified version of the GA6.1 shallow convection scheme, designed to parameterize the precipitation from small-scale (but not necessarily shallow) showers that cannot be resolved explicitly but which can generate significantly higher accumulations of precipitation than are typically produced with the shallow convection closure used in GA6.1. However, in TC cases, the contribution to the total rainfall from this “gray zone” convection scheme is typically around 1%, meaning convection in the PHI model is essentially explicit (in GA6.1, the convection scheme is responsible for approximately 85% of the total rainfall in TCs).

From a TC modeling perspective, there are three other noteworthy features of the PHI and GA6.1 models. First, both models are atmosphere-only models, with the sea surface temperature (SST) held fixed throughout a forecast. This means there is a limitless source of heat energy for storm intensification and maintenance whereas, in reality, TCs induce negative SST anomalies along their track, reducing the energy available (e.g., Cione and Uhlhorn 2003; D’Asaro et al. 2007). Second, the boundary layer scheme includes a source term representing heating from the dissipation of turbulence based on the theoretical arguments of Bister and Emanuel (1998), known to generate more intense TCs in numerical models (Zhang and Altschuler 1999; Jin et al. 2007). Third, no vortex initialization scheme is implemented in the PHI regional model to improve the initial representation of vortices relative to the coarse global analyses used to initialize the model.

b. Data

1) MODEL EVALUATION PERIOD

The PHI model has been run on a routine basis since August 2014, producing two 5-day forecasts per day (at 0000 and 1200 UTC). For model evaluation, the period from 4 November 2014 to 15 March 2016 is chosen for two reasons. First, both the PHI and global models were essentially unchanged during this time (only technical changes were made, including an upgrade of the UM version and a bug fix in the reconfiguration of global analyses to the regional grid). Second, there are global analyses available for this period that were produced with the new TC initialization scheme described in Heming (2016), in which central pressure estimates from TC warning centers are ingested as part of the global DA cycle. This has been shown to improve global model TC track and intensity forecasts (Heming 2016).

2) TROPICAL CYCLONE TRACKING AND VERIFICATION

TCs are identified in UM output using the Met Office tracking software, the details of which can be found in Heming (2017). In short, observed positions of TCs are matched with local maxima of 850-hPa relative vorticity in the model analysis. A search is then made around each of these points for the nearest local minimum in pressure (at mean sea level; MSL) to determine the storm center.

At each lead time, various tests are applied (e.g., the relative vorticity and pressure values must be greater and less than some prescribed threshold, respectively) to decide if the TC fix is legitimate. If it is, and the corresponding observed maximum wind speed is greater than 31 kt (where 1 kt = 0.51 m s⁻¹), then diagnostic
information is written out, including forecast and observed positions and intensities, positional errors, and the radius at which the maximum wind speed occurs.

3) SAMPLE SIZE

The TC tracker has been applied to operational PHI and global model output over the chosen evaluation period to construct two samples of model TCs. Matching cases were then extracted from the two samples to construct a single, homogeneous sample. A total of 20 storms were tracked in the regional model domain shown in Fig. 1 during this time; the number of storm forecasts is shown as a function of lead time in Fig. 2.

For the purposes of this study, storms are also grouped into two intensity categories: category 3 and above (CAT35), and below category 3 (<CAT35). At a given lead time, the category a storm is assigned to is governed by its observed peak surface (10 m) wind speed at that time. If the wind speed is greater than 95 kt, the storm is assigned to the CAT35 category; otherwise, it is assigned to the <CAT35 category. A storm will typically move between the different intensity categories during a forecast. The dashed and dotted lines in Fig. 2 show the numbers of forecasts for storms in the two intensity groups.

The number of storm forecasts generally decreases with lead time. This is because the TC tracker only follows storms that are observed to exist at the model analysis time. In other words, storms that develop or advect into the domain during the course of a forecast are not tracked. If they were, it is anticipated that the number of cases would be approximately constant with lead time.

3. Evaluation of model performance

a. Intensity forecasts

Figure 3 shows the mean absolute error (MAE) in TC maximum surface wind speed and central pressure as a function of forecast lead time for the PHI and GA6.1 models. The corresponding mean biases are shown in Fig. 4. Error bars are 95% confidence intervals on the mean, and the shaded regions are one standard deviation about the mean. For comparative purposes, errors computed from HWRF (Tallapragada et al. 2015b) and European Centre for Medium-Range Weather Forecasts (ECMWF) operational model forecasts for the same TC cases are also displayed.

The intensity of storms is underestimated in the Met Office global analyses used to initialize UM forecasts, particularly in severe storms (CAT35) where central pressures are ~30 hPa higher (and wind speed values are ~50 kt lower) than observed at T + 0, on average. Given that central pressure estimates from TC warning centers are assimilated as part of the global DA cycle, naively one might expect smaller initial pressure errors than this. However, the pressure estimates are flagged, and subsequently ignored, by the global model DA if the implied increments to the analysis fields are too large (typically when \( O - B \) values are \( \geq 15 \) hPa). This is more likely in intense systems, during periods of rapid intensification, for example. ECMWF global analyses also exhibit a weak intensity bias, although

---

1 The observed wind speeds are 1-min mean surface wind speeds from the JTWC, extracted from the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp et al. 2010).

2 At each lead time, 95% confidence intervals are computed via \( \bar{x} \pm t^{*}s/\sqrt{n} \), where \( \bar{x} \) and \( s \) are the mean and standard deviation of the sample of \( n \) data points at that time, respectively, and \( t^{*} \) is the critical value of the \( t \) distribution with \( n - 1 \) degrees of freedom at a (two tailed) significance level of \( 1 - (95/100) = 0.05 \). Note that this does not account for any serial correlation between successive storm forecasts, meaning that the width of the confidence intervals is likely to be underestimated.

3 Operational HWRF intensity and track forecasts have been downloaded directly online (e.g., https://www.nrlmry.navy.mil/atcf_web/docs/aids/2015). ECMWF forecast data have been retrieved on a 0.125° × 0.125° grid (slightly coarser than the average operational resolution of 9 km), and storms have been identified and tracked using the Met Office tracking software described in section 2.
initial errors in CAT35 storms are smaller than in the UM (especially in central pressure). Recall that initial conditions for the PHI regional model are generated simply by interpolating GA6.1 global analyses, and thus the intensity errors at the analysis time are almost identical. By contrast, the HWRF regional modeling system has much smaller errors in wind speed and central pressure of approximately 5 kt and 5 hPa at T + 0, respectively. This is due to the sophisticated vortex initialization procedure employed in the model (Tallapragada et al. 2015b). Such schemes are used in other operational regional TC modeling systems, for example COAMPS-TC (Hendricks et al. 2011) and ACCESS-TC (Davidson et al. 2014), and have been demonstrated to benefit track and intensity forecasts (e.g., Cha and Wang 2013; Hendricks et al. 2013; Nguyen and Chen 2014). Improving the initial representation of vortices in regional configurations of the UM is thus a top priority for future model development work.

Considering the most intense storms (CAT35), the PHI model undergoes a marked spinup period over the first day or so of the forecast as the model adjusts from the weak initial state. During this time, intensity errors steadily decline and, beyond T + 36, the PHI model provides the closest match to the observed wind speeds of all the models (HWRF has a similar wind speed bias but a larger MAE). In particular, compared with GA6.1,
wind speed predictions are far superior at all lead times, with a reduction in mean absolute wind speed errors of over 50% at $T + 24$ and beyond. This is because the wind speed error in the global model remains approximately constant throughout the forecast, on average, resulting in a systematic weak bias. The ECMWF model generates more intense storms than GA6.1 (recall that it has a grid length approximately a factor of 2 smaller) and, thus, offers improved wind speed predictions (smaller MAE and bias), but it cannot compete with the two regional models.

However, the PHI model overdeepens intense systems, with central pressures dropping below those observed at about $T + 24$, asymptoting to a value approximately 25 hPa too low beyond $T + 48$. Nonetheless, the MAE in central pressure is smaller in the PHI model than GA6.1 up to $T + 60$. Both the HWRF and ECMWF models yield more realistic central pressures than either of the UM configurations for CAT35 storms, especially HWRF, which has a bias close to zero out to $T + 84$.

The PHI model has a tendency to overintensify weaker storms (<CAT35), with maximum surface wind speeds higher (and central pressures lower) than observed, on average. Intensity errors grow monotonically throughout the forecast. The GA6.1 model again underestimates storm intensity at all lead times, but biases are smaller than for strong storms. Predictions are similar to those of the ECMWF model, especially for wind speed. Compared to GA6.1, the PHI model gives a smaller MAE and bias in wind speed up to $T + 30$, but subsequent overintensification leads to larger errors beyond this time. The HWRF regional model also favors overintensification of weaker storms, but to a lesser extent than the PHI model since the growth of errors essentially ceases beyond $T + 48$ or so.
The overintensification of storms by the PHI model could partly be due to the lack of ocean feedback on the atmospheric circulation in the model. In reality, the high wind speeds in TCs mix cooler water from the thermocline with the overlying mixed layer, reducing surface fluxes of moisture and heat to the TC, thus limiting storm intensification (e.g., Bender et al. 1993; Schade and Emanuel 1999; Cione and Uhlhorn 2003; D’Asaro et al. 2007). Another factor could be excessive heating from the dissipation of turbulence in the boundary layer. The parameterization of dissipative heating used in the UM boundary layer scheme is similar to that of Zhang and Altshuler (1999), which is based on the theoretical arguments presented in Bister and Emanuel (1998). Measurements of the turbulent kinetic energy dissipation rate in the TC boundary layer (Zhang 2010) suggest the dissipative heating by Bister and Emanuel (1998) could be significantly higher than observed.

1) WIND–PRESSURE RELATION

It follows from Figs. 4b and 4e that, for strong storms, the dynamical relationship between the pressure and wind fields in the PHI model must be different from that of the Joint Typhoon Warning Center (JTWC) best-track data. To highlight this, Fig. 5 shows scatterplots of maximum surface wind speed and central pressure for the two UM configurations, including all storms in the sample and all forecast lead times (giving a total of 998 data points for each model). The observed wind–pressure relation (WPR) from JTWC is shown for comparison (note that there are as many observational data points as for each model, but there are many identical points that lie on top of each other), along with corresponding relations from the HWRF and ECMWF models.

The ability of the PHI model to generate more realistic intensities than GA6.1 leads to an improved WPR. The PHI model relation is a good match to that observed up to wind speeds of \( \sim 100 \text{ kt} \), but is too steep beyond that point. By contrast, the GA6.1 relation deviates from observations at lower wind speeds (\( \sim 50 \text{ kt} \)) and does not reach the bottom-right corner of the plot, where the most intense systems lie. The ECMWF model relation is similar, again much steeper than observed. The best match to the observations is provided by the HWRF Model, which captures the full range of observed wind speeds, although central pressures are also lower than observed in the most intense systems.

Of course, the observations themselves may be in error in the high-intensity limit. At JTWC, wind speeds are estimated from satellite imagery using the Dvorak technique, and then central pressures are inferred using the WPR described in Knaff and Zehr (2007). As noted by Tallapragada et al. (2016), it is possible that this relation does not fully capture the dynamical constraints between mass and wind fields at high wind speeds.

A possible reason for the difference between the PHI model WPR and the observed relation at high wind speeds could be that surface fluxes are in error in the UM in this regime. The surface transfer scheme used in the PHI model assumes the momentum roughness length over open ocean is given by a modified version of Charnock’s formula (Smith 1988) with a fixed Charnock coefficient of \( \alpha = 0.011 \). The drag coefficient in the model thus increases with increasing wind speed whereas observations suggest it levels off above approximately 64 kt, and may even decrease beyond this point (Powell et al. 2003; Donelan et al. 2004; Black et al. 2007; Holthuijsen et al. 2012; Soloviev et al. 2014). Model predictions for the TC WPR are known to be highly sensitive to surface fluxes (Bao et al. 2002; Bryan 2012; Green and Zhang 2013, 2014). The variation of the drag coefficient with wind speed in HWRF is more consistent with the observations (Tallapragada et al. 2015b), which perhaps explains why it is better able to reproduce the observed WPR at high wind speeds. Another factor could be the increased horizontal resolution of the HWRF Model (2-km inner nest) compared to the PHI model (4.4-km grid length). An investigation of these topics is beyond the scope of the present study.
2) RATE OF INTENSIFICATION

Another measure of how well an NWP model can forecast TCs is whether it can predict the stage of development of a storm (i.e., intensifying, steady state, or decaying). Of particular interest is whether a model can successfully capture episodes of rapid intensification [RI; defined as an increase in maximum surface wind speed of at least 30 kt over a 24-h period, following Kaplan and DeMaria (2003)] since this can have huge impacts if it occurs near the time of landfall. Cases of RI occur more frequently in the western North Pacific than other basins but are still relatively rare (~5%–10%). Forecasting RI is a notoriously challenging problem.

To examine if the PHI and GA6.1 models can predict RI, Fig. 6 shows scatterplots of forecasts against observed 24-h future wind speed changes. All storms and all forecast lead times (up to $T + 96$) are included, giving a total of 1354 data points in each plot.

Data points inside the green squares in Fig. 6 correspond to observed RI events that were correctly captured by the models (hits). Of the 120 observed RI cases, the PHI model predicted 38 of these, giving a probability of detection (POD) of 38/120 = 0.32. The GA6.1 global model does not produce any RI cases at all, giving a POD of zero.

Given the potentially high impact associated with rapidly intensifying storms, the ability of the PHI model to predict genuine RI cases is a key advantage over the global model. The downside is that the model gives too many false alarms (shown by the points in the red rectangle in Fig. 6a), that is, predictions of RI where none was observed. For the PHI model, there are 142 false alarms, giving a false alarm ratio (FAR) of 142/(38 + 142) = 0.79.

Recall that large initialization errors for the most intense storms lead to a pronounced spinup period for the PHI model (see Figs. 4b,e). A natural question to ask is whether this influences the number of RI cases. If the above analysis is repeated, neglecting all CAT35 cases at and before $T + 24$ (of which there are 259, shown by the grayed-out symbols in Fig. 6a), then 57 of the 142 false alarms are removed, and the FAR drops to 0.73 (six of the hits are also removed, lowering the POD to 0.3). The majority of the discarded false alarms (44 out of 57) correspond to cases where storms were observed to be decaying but the model predicts RI (i.e., points lying to the left of the vertical dashed line in Fig. 6a). These typically occur when a weak analysis is followed by a rapid spinup of the model toward the observations. Improvements to the model initialization procedure could potentially eliminate this type of false alarm. Some of the RI false alarm cases lying to the right of the vertical dashed line in Fig. 6a are due to the genuine overintensification of storms at longer lead times, potentially because of the lack of ocean coupling in the model, or excessive dissipative heating in the surface layer.

b. Track forecasts

The Met Office operational global model (GA6.1) is currently one of the leading global models for TC track
forecasts. For example, over the western North Pacific in 2014 the Met Office global model outperformed all other operational global models in terms of day 3 track errors (Heming 2016).

To compare track predictions from the PHI regional model with those from GA6.1, Fig. 7 displays the mean error in storm position relative to observations [as measured by the direct positional error (DPE); e.g., Heming (2017)] as a function of forecast lead time. Corresponding mean track errors derived from HWRF and ECMWF operational forecasts are also shown in Fig. 7.

It is clear that track errors are comparable in the PHI and GA6.1 models; any differences between the mean track errors are not statistically significant. On average, both models are able to forecast the positions of more severe storms (CAT35) better than weaker storms (<CAT35). Although mean track errors relative to observations are similar, storm positions are typically different in the two models. The mean distance between model storm positions (shown by the black lines in Fig. 7) is a significant fraction of the DPE relative to the observations (especially in the first 48 h of the forecast), implying that the steering flow inherited from the driving global model is modified by the convection-permitting model.

Of the various models, the ECMWF model is able to predict the TC track most accurately, on average. HWRF gives similar mean track errors to the two UM configurations out to $T + 48$ or so, but the rate of error growth increases beyond this time, leading to a larger mean track error in the latter stages of the forecast.

c. Precipitation forecasts

Copious rainfall is one of the greatest hazards of landfalling TCs. For example, in the Philippines, heavy rain combined with steep topography often leads to flash floods and landslides, with a high human and financial cost (e.g., Faustino-Eslava et al. 2013). It is clearly important for NWP models to be able to provide accurate rainfall forecasts during TC events.

The PHI model, by virtue of its increased horizontal resolution, offers a much more detailed picture of the rainfall pattern in TCs than GA6.1. To investigate whether this corresponds to improved forecast skill, model predictions for TC rainfall are compared with estimates derived from Global Precipitation Measurement (GPM) satellite (Hou et al. 2014) measurements, specifically the Integrated Multisatellite Retrievals for GPM (IMERG) Final Run product. This combines microwave and infrared (IR) measurements from several satellites, calibrated using surface rain gauge data where available, to generate half-hourly mean precipitation rate estimates on a $0.1^\circ \times 0.1^\circ$ grid extending from $60^\circ$S to $60^\circ$N (Huffman et al. 2015). It is worth bearing in mind that multisatellite precipitation
estimates will be subject to biases (e.g., Habib et al. 2009; Yu et al. 2009; Chen et al. 2013a,b), because rainfall is inferred from path-integrated hydrometeor content (in the case of passive microwave algorithms) or cloud-top temperatures (in the case of IR algorithms).

As demonstrated in section 3b, there can be significant differences between forecast and observed storm positions at long lead times (Fig. 7). In such cases, the location of rainfall associated with a TC would be very different from that observed. To remove this source of bias, a storm-centered approach is adopted here, so that the focus is on how well the models can predict rainfall within TCs.

For every model forecast, at each lead time (i.e., every 6 h) a 9° square, centered on the TC center identified by the tracking algorithm at that time, is extracted from the model output. This is large enough to accommodate a circle of radius 500 km centered on the TC, which is the maximum radius typically used to define TC-related rainfall (e.g., Lonfat et al. 2004). A region of the same size is extracted from corresponding GPM data but using the observed storm position as the center. Any cases where the extracted region extends beyond the boundary of the Philippines regional model domain (Fig. 1) are discarded, reducing the total number of cases in the homogeneous sample from 2031 to 1028.

To facilitate a fair comparison between the models and GPM observations, all data have been upscaled onto a common grid—the GA6.1 grid—using a conservative regridding scheme. Of course, much of the small-scale detail provided by the PHI model will be lost by smoothing onto the coarse global grid. Note that averaging model data onto the finer 0.1° × 0.1° GPM grid instead does not change the results presented in this section in any significant way.

As a final remark, differences in track could lead to very different precipitation patterns when storms are close to land as a result of interaction with topography. However, our results are qualitatively unchanged if the analysis is repeated, discarding all storm-centered boxes that contain any land points (this conservative constraint reduces the number of cases to 311).

1) SPATIAL LOCATION OF RAINFALL WITHIN TROPICAL CYCLONES

A powerful technique for quantifying the spatial accuracy of high-resolution model precipitation forecasts is the fractions skill score (FSS), introduced by Roberts and Lean (2008). The first step in computing the FSS is to apply some threshold to both the forecast field and the verifying observed field (GPM data in this case), both on the same grid. For both fields, the fraction of grid points exceeding the threshold within a specified square area (neighborhood) around each grid point is determined, and the FSS statistic is computed from these fractions [see Roberts and Lean (2008) for a full description of the method]. This process can then be repeated for a range of different size neighborhoods to determine how forecast skill depends on spatial scale.

Thresholds can either be fixed physical thresholds (e.g., 4 mm h⁻¹), or percentile thresholds (e.g., a 90th percentile threshold selects the highest 10% of rain rates). Percentile thresholds are particularly useful for eliminating any forecast bias and assessing just the spatial accuracy of models (Roberts and Lean 2008). This is the approach adopted here, with precipitation biases discussed in the following section.

The FSS for model 6-hourly mean precipitation rates has been computed for each matching pair of storm-centered square regions extracted from model output and GPM data. Figure 8 shows the mean FSS over all cases as a function of neighborhood size, for a 90th percentile threshold. Note that the FSS can take values anywhere between 0 (complete mismatch) and 1 (perfect forecast). At this threshold, the PHI model has a higher skill than the global model at all spatial scales for strong storms (CAT35), being always above the target skill.⁴ For weaker storms (<CAT35), the forecast skill is essentially the same in the two models (no statistically significant differences).

To examine how forecast skill depends on the (percentile) threshold applied, Fig. 9 shows how the neighborhood size at which the target skill is reached varies with threshold. In the most intense systems, it is clear that the convection-permitting model is better able to predict the spatial distribution of rainfall within TCs than the GA6.1 global model at all thresholds. For weaker storms, the PHI model has greater skill than GA6.1 for more widespread rain (75th percentile threshold and below), but there are no statistically significant differences between the models at higher thresholds. Finally, both models are more skillful in predicting the location of rainfall in strong storms than in weak storms, particularly more localized rain (higher percentile thresholds).

---

⁴ The target skill is given by 0.5 + fₘ₀/2, where fₘ₀ is the observed fractional rainfall coverage over the domain (a 9° square in this case). This represents the FSS that would be obtained at the grid scale (neighborhood size of 1) from a forecast with a fraction equal to fₘ₀ at every grid point. This is a useful measure of the smallest scale on which a forecast has skill (Roberts and Lean 2008).
2) RADIAL DISTRIBUTION OF RAINFALL WITHIN TROPICAL CYCLONES

To examine the statistical properties of rainfall within TCs, histograms of 6-hourly mean rain rates are calculated in a series of 25-km-wide annuli around the TC center (out to a maximum radius of 500 km), for each matching pair of storm-centered square regions extracted from model output and GPM data. For each radial bin, a total histogram is formed by summing the histograms from all cases and then normalized. Figure 10 shows the resulting rain-rate histograms as a function of radial distance from the storm center. Overlaid are azimuthally averaged 6-hourly mean rain rates for GPM and the models.

In the innermost regions of storms, the shape of the observed rain-rate distribution is better captured by the PHI model than GA6.1, in both strong (CAT35) and weak (<CAT35) storms. However, the PHI distribution is left skewed relative to the GPM distribution with a peak (mode) at higher rain rates than observed, signifying too much heavy rain in TCs. The GA6.1 distribution is too broad compared to the observations, with too large a contribution from low rain rates and a mode at even higher rain rates than in the PHI distribution.

The GPM distribution of rain rates in the core of intense storms is narrower and more sharply peaked than...
in weaker storms, but with a similar mode. The observed azimuthal-mean rain rate also increases with storm intensity (at least for radii $r \leq 300\,\text{km}$), which is well known (e.g., Rodgers and Adler 1981; Lonfat et al. 2004; Yokoyama and Takayabu 2008). Both models can qualitatively reproduce this dependence on intensity.

As the distance from the storm center increases, the PHI distribution broadens and the peak smoothly shifts toward lower rain rates, with a corresponding drop in the azimuthally averaged rain rate. The GPM distribution displays similar behavior. However, compared to GPM, the distribution broadens too quickly with an increasing radius in the PHI model, which is particularly apparent for the most intense storms.

The shape of the GA6.1 distribution varies with radius in a different way as compared with the GPM observations and the PHI model. At a radius of approximately 125 km for CAT35 storms (75 km for <CAT35 storms), there is a sharp drop in the
mode of the distribution. At radii beyond this size, the mode decreases and the distribution broadens, but at a slower rate than observed, leading to clear differences at large radii.

4. Conclusions
This work has evaluated the performance of a convection-permitting (4.4-km grid length), regional configuration of the Met Office Unified Model (UM) for predicting tropical cyclones (TCs) in the western North Pacific. The main aim has been to assess whether this regional model (referred to as the PHI model) offers benefits for TC forecasting relative to the Met Office global model (GA6.1) used to drive it.

Using a sample of storms constructed from over a year’s worth of operational forecasts, model predictions for TC intensity, intensity changes, track, and precipitation have been compared with observational data. The main findings are as follows:

1) The PHI model gives much better predictions for the surface wind speeds in intense TCs [category 3 and above (CAT35)] than the GA6.1 global model, which cannot generate wind speeds above ~100 kt and thus suffers from a systematic negative wind speed bias. In terms of the central pressure in severe storms, the PHI model also provides a better match to the observations (smaller MAEs) than GA6.1 up to $T + 60$. However, at lead times beyond this, the PHI model tends to overdeepen CAT35 storms (lower central pressures than observed), with a negative bias comparable in size to the positive bias seen in the global model. Weaker systems [below category 3 (<CAT35)] are generally overintensified by the PHI model, with maximum wind speeds higher, and central pressures lower, than observed. The MAEs in wind speed are smaller than in GA6.1 up to $T + 30$, but overintensification results in larger absolute errors beyond this time. At all lead times, absolute central pressure errors are larger in the PHI model than GA6.1.

2) The ability of the PHI model to produce storms with more realistic intensities than GA6.1 leads to an improved wind–pressure relation (WPR) in the high-intensity limit.

3) Episodes of rapid intensification (24-h future wind speed change exceeding 30 kt) can be successfully predicted by the PHI model. The GA6.1 global model cannot capture such events at all. However, the convection-permitting model generates too many false alarms (i.e., rapid intensification predicted but none observed).

4) Tropical cyclone track errors relative to the observations are similar in the PHI and GA6.1 models. Both models are able to forecast the position of intense storms (CAT35) better than weaker storms (<CAT35). Interestingly, storms are typically not collocated in the two models, with the mean distance between model storm positions being a significant fraction of the positional error relative to the observations. This implies that the motion of TCs in the convection-permitting model is not completely controlled by the large-scale flow inherited from the driving global model.

5) Model predictions for the spatial location of rainfall within TCs have been verified against estimates from GPM satellite measurements, using the fraction skill score (FSS). It was found that the PHI model can predict the location of rainfall within storms more skillfully than GA6.1. The skill of both models is greater for strong storms than weak storms.

6) Azimuthally averaged rain rates in the inner regions of TCs are approximately a factor of 2 higher than observed (by GPM) in the PHI model. This bias is also present in the global model but to a lesser extent.

7) In both the PHI and GA6.1 models, the statistical distribution of rain rates in the innermost regions of storms is skewed toward higher rain rates than observed by GPM, indicating too much heavy rain in TCs. However, the shape of the observed distribution is better captured by the convection-permitting model. The radial variation of the rain-rate distribution is also better represented by the PHI model, although the distribution broadens with increasing radius at a rate quicker than observed.

In summary, it has been demonstrated that a convection-permitting configuration of the UM adds value over the Met Office global model for TC forecasting, providing much improved predictions for intensity, intensity changes, and precipitation in severe storms. Nonetheless, several key model biases have been identified that will be the target of future model development work:

1) The intensity of CAT35 storms is significantly underestimated in the initial conditions for the PHI regional model, because they are derived by interpolating Met Office global analyses defined on a relatively coarse grid. To reduce these initial errors, a vortex initialization scheme will be included in the model, similar to other operational
regional TC modeling systems. This should also help mitigate the 24–36-h spinup period of the PHI model seen in intense storms, extending the useful range of the model for TC forecasting.

2) The PHI model has a tendency to overintensify storms, which could be due, at least in part, to the lack of ocean feedback on the atmosphere in the model. Including some form of air–sea coupling in the UM is a long-term goal. Another possibility would be to reduce the heating rate from the dissipation of turbulence in the boundary layer scheme, which may be in excess of that observed in the current parameterization. Preventing the overintensification of storms has the potential to reduce the number of RI false alarms.

3) The PHI model WPR is steeper than observed in the most intense systems, likely because of excessive drag in the model. A new parameterization of air–sea drag in which the drag coefficient is reduced at high wind speeds (motivated by observations) will be implemented in the model.

4) The excessive rain rates seen in the inner regions of model storms may be alleviated by the inclusion of a recently developed scheme for enforcing moisture conservation in semi-Lagrangian models (Zerroukat and Shipway 2017).

TABLE A1. Summary of the global (GA6.1) and regional (PHI) configurations of the Met Office Unified Model used in this study.

<table>
<thead>
<tr>
<th>Model name</th>
<th>GA6.1</th>
<th>PHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid spacing</td>
<td>0.156°N–S × 0.234°E–W</td>
<td>0.04°N–S × 0.04°E–W</td>
</tr>
<tr>
<td>Vertical levels</td>
<td>70 levels, 80-km lid</td>
<td>80 levels, 38.5-km lid</td>
</tr>
<tr>
<td>Time step</td>
<td>450 s</td>
<td>120 s</td>
</tr>
<tr>
<td>Radiation</td>
<td>Two-stream scheme based on Edwards and Slingo (1996) with modifications described in Walters et al. (2017); full radiation calculations once every hour</td>
<td>As in GA6.1, but with full radiation calculations every 30 min and corrections to increments made every 10 min (Manners et al. 2009); surface radiative fluxes are corrected for the mean grid box slope, shading from surrounding terrain, and the effects of sky-view factor (Manners et al. 2012)</td>
</tr>
<tr>
<td>Microphysics</td>
<td>Mixed-phase, single-moment scheme based on Wilson and Ballard (1999) with modifications described in Walters et al. (2017)</td>
<td>Scale-aware, warm rain scheme (Boutle et al. 2014a); additional graupel prognostic included</td>
</tr>
<tr>
<td>Large-scale cloud</td>
<td>Prognostic cloud, prognostic condensate scheme (Wilson et al. 2008) with modifications described in Walters et al. (2017)</td>
<td>Diagnostic cloud scheme (Smith 1990) with cloud area parameterization of Boutle and Morcrette (2010)</td>
</tr>
<tr>
<td>Orographic gravity wave drag</td>
<td>Flow-blocking and mountain wave drag scheme based on Lott and Miller (1997) with modifications described in Walters et al. (2017); form drag parameterized via an effective roughness scheme (Wood and Mason 1993)</td>
<td>Flow-blocking and mountain wave drag scheme (Webster et al. 2003); form drag parameterized as in GA6.1</td>
</tr>
<tr>
<td>Nonorographic gravity wave drag</td>
<td>Spectral gravity wave scheme (Scaife et al. 2002)</td>
<td>Not included</td>
</tr>
<tr>
<td>Boundary layer</td>
<td>1D vertical turbulent mixing scheme (Lock et al. 2000) with modifications described in Walters et al. (2017); frictional heating due to dissipation of turbulence included</td>
<td>“Blended” boundary layer scheme (Boutle et al. 2014b); frictional heating due to dissipation of turbulence included</td>
</tr>
<tr>
<td>Convection</td>
<td>Mass-flux scheme based on Gregory and Rowntree (1990) with modifications described in Walters et al. (2017)</td>
<td>“Gray zone” convection scheme (see text)</td>
</tr>
<tr>
<td>Sea surface</td>
<td>Momentum roughness length parameterized as in Smith (1988) with fixed Charnock coefficient $\alpha = 0.018$; variable thermal roughness length based on surface divergence theory (Edwards 2007); prescribed SSTs</td>
<td>As in GA6.1, but with $\alpha = 0.011$</td>
</tr>
<tr>
<td>Land surface</td>
<td>JULES land surface scheme (Best et al. 2011; Clark et al. 2011) with a single aggregate surface type</td>
<td>JULES with fluxes calculated separately for nine different surface types</td>
</tr>
</tbody>
</table>
APPENDIX

Summary of Model Configurations

Table A1 summarizes the key features of the numerical models used in this study. The reader is referred to the papers cited in the table for full details of the various physics parameterization schemes included in the models.

REFERENCES


FEBRUARY 2018

SH OR T AND P E T C H


