Evaluating convection-permitting ensemble forecasts of precipitation over Southeast Asia

Article

Accepted Version

Evaluating convection-permitting ensemble forecasts of precipitation over Southeast Asia

Samantha Ferrett[1], Thomas H. A. Frame[1], John Methven[1], Christopher E. Holloway[1], Stuart Webster[2], Thorwald H.M. Stein[1], Carlo Cafaro[1]

[1]Department of Meteorology, University of Reading, Reading, UK

[2]Met Office, Exeter, UK

Corresponding Author:
Samantha Ferrett
s.j.ferrett@reading.ac.uk
Department of Meteorology
Meteorology Building
Whiteknights Road
Earley Gate
Reading, RG6 6ET

Abstract

Forecasting rainfall in the tropics is a major challenge for numerical weather prediction. Convection-permitting (CP) models are intended to enable forecasts of high-impact weather events. Development and operation of these models in the tropics has only just been realised. This study describes and evaluates a suite of recently developed Met Office Unified Model CP ensemble forecasts over three domains in Southeast Asia, covering Malaysia, Indonesia and the Philippines.

Fractions Skill Score is used to assess the spatial scale-dependence of skill in forecasts of precipitation during October 2018 - March 2019. CP forecasts are skilful for 3-hour precipitation accumulations at spatial scales greater than 200 km in all domains during the first day of forecasts. Skill decreases with lead time but varies depending on time of day over Malaysia and Indonesia, due to the importance of the diurnal cycle in driving rainfall in those regions. Skill is largest during daytime when precipitation is over land and is constrained by orography. Comparison of CP ensembles using 2.2, 4.5 and 8.8 km grid spacing and an 8.8km ensemble with parameterised convection reveals that varying resolution has much less effect on ensemble skill and spread than the representation of convection. The parameterised ensemble is less skilful than CP ensembles over Malaysia and Indonesia and more skilful over the Philippines; however, the parameterised ensemble has large drops in skill and spread related to deficiencies in its diurnal cycle representation. All ensembles are under-spread indicating that future model development should focus on this issue.

1. Introduction

The nations of South East Asia are susceptible to devastating impacts of heavy rainfall such as flooding and landslides. Deep convection is central to extreme rainfall intensity in the region (e.g. Matsumoto et al., 2017) and it also plays an active part in the dynamics of the larger scale atmospheric phenomena that dominate in the region. There are many contributing factors to the occurrence of convective rainfall events on a range of spatial and temporal scales from the longest to the shortest such as: the El Nino-Southern Oscillation (Hamada et al. 2012; Villafuerte and Matsumoto 2015; Supari et al. 2018), the Madden-Julian Oscillation (MJO; Wu et al. 2013; Peatman et al. 2014; Xavier et al. 2014; Birch et al. 2016; Vincent and Lane 2018; Lestari et al. 2019), cold surges (Chang et al. 2005; Lim et al. 2017), equatorial waves (Ferrett et al. 2020) and tropical cyclones (Takahashi and Yasunari 2008).

Rainfall variability in many regions of Southeast Asia has a strong diurnal cycle, a result of local land-sea breeze circulations, surface heating during the day and the delayed response of deep convection (e.g. Mori et al., 2004; Yamanaka, 2016). Coarse-grid models (including all current operational global models) rely on convection parameterisations and are known to struggle to capture accurately tropical rainfall features (e.g. Neale & Slingo, 2003; Johnson et al., 2016), such as the diurnal cycle (Yang and Slingo 2001; Love et al. 2011), the propagation of the MJO and equatorial waves (Lin et al. 2006; Holloway et al. 2013; Peatman et al. 2015; Peatman et al. 2018), and other aspects of convection (Qian 2008; Pearson et al. 2010, 2014). Consequently, there has been considerable effort to improve the representation of these processes in models, namely by increasing model resolution so that deep convection can be explicitly simulated by the dynamical core of the model reducing the need for convection parameterisations. In a numerical weather prediction context, there is a trade-off between resolution, model domain size and ensemble size. Resolving deep convection requires a model capable of representing non-hydrostatic, compressible dynamics and a horizontal (and vertical) grid spacing that is much less than the depth of deep convective updrafts (10-20 km). If achieving this resolution requires that the domain is too small, the behaviour of systems developing in the domain will be almost completely specified by the lateral boundary conditions imposed by the parent, lower resolution model, in which the model is nested and there is no benefit in the high-resolution prediction. If the domain is large enough to allow the interior solution of the high-resolution model to deviate from that of its parent, then a compromise must be made on resolution and ensemble size in order to obtain an ensemble forecast in near-real time.

The current state-of-the-art for operational numerical weather prediction is that deep convective motions are only partially resolved; such models have horizontal grid spacing of the order 1-10 km and are described as “convection-permitting” (CP) and the ensembles are small (~10 members). Often shallow moist convection is not resolved and may be parametrized as part of the convective regime of the boundary layer scheme, although the approach taken to turbulence and boundary layer parameterization also depends on model resolution. While CP forecasts are more computationally expensive, they are better able to represent processes that drive convection (Clark et al. 2016). Studies have shown significant improvement in the initiation of convection (Mittermaier et al. 2013; Birch et al., 2014a; 2014b; 2015; Woodhams et al., 2018), the diurnal cycle of convection (Sato et al. 2009; Love et al. 2011; Birch et al. 2015), and large-scale modes that drive convection (Miura et al.

2007; Holloway et al. 2013) in CP models compared to those with parameterised convection.

Forecasting at convective scales is inherently uncertain, even at short lead times (Hohenegger and Schar 2007). This uncertainty can be associated with many things, such as model physics, initial conditions, or boundary conditions. Therefore, ensembles of forecasts are used to account for uncertainty. These convective-scale ensembles have been developed in regions world-wide (e.g. Gebhardt et al. 2011; Golding et al. 2014; Schwartz et al. 2015; Hagelin et al. 2017; Roberts et al. 2019) with obvious improvements in forecast accuracy compared to a single forecast (Hagelin et al. 2017). While there have been many studies of the benefits of using ensemble forecasts to predict the risk of high impact weather in the extra-tropics (Hanley et al. 2013; Bednarczyk and Ancell 2015) and CP ensembles have been shown to add value to forecasts of mesoscale phenomena such as sea breezes (Cafaro et al. 2019), there are fewer studies examining how CP ensembles may benefit forecasts of extreme rainfall in the tropics more widely and Southeast Asia specifically. A few recent studies have focused on CP ensemble forecasts over Singapore and the surrounding region (Porson et al. 2019; Sun et al. 2020). These studies focused on comparing the effect of global models in which CP ensembles are nested on forecast skill and spread (Porson et al. 2019) and comparing objective and subjective evaluation methods of forecasts of squall lines (Sun et al. 2020). Both studies evaluate against observations in relatively small regions around Singapore (approx. 400km x 400km).

In this study CP ensemble forecasts are one-way nested in limited area domains within the operational MOGREPS global ensemble using three horizontal resolutions.

Three domains within the Southeast Asia region are examined, including Peninsular Malaysia, Java and the Philippines. The aim of this study is to quantify the usefulness of CP ensemble forecasts of precipitation in this region, the scale dependence of forecast skill as well as the role of the diurnal cycle in forecast skill.

Descriptions of the CP ensembles and other datasets used for the analysis are provided in Section 2. The methods for evaluation of forecasts using observations are outlined in Section 3. Section 3 also provides details of the construction of a variation of the ensemble forecast, and a persistence forecast, that are used to assess the role of the diurnal cycle in forecast skill. Section 4 provides the results of the study, detailing the spatial scale-dependence in the skill of the forecasts, the role of the diurnal cycle and the spread of the ensembles in relation to mean forecast error as a function of lead time. Results are summarised in Section 5.

2.Data

2.1 Ensemble forecasts

The convection-permitting (CP) ensemble forecasts consist of 18 ensemble members and were created by nesting limited area simulations using the Met Office Unified Model (MetUM) within the 18-member operational Global ensemble of the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G). The CP ensembles were initialised twice daily (00 UTC and 12 UTC) over a period of six months spanning October 2018-March 2019, producing hourly forecast output in three domains corresponding to Malaysia, Indonesia and the Philippines (Figure 1) out to 120 hours, except for 2.2km forecasts which are ran for 60 hours. Only the 00 UTC forecasts are shown here, however analysis has also been carried out for the 12

UTC forecasts with similar results after the initial model spin-up period, albeit displaced by twelve hours.

The forecasts use horizontal grid spacing of 2.2, 4.5 and 8.8 km and are nested in MOGREPS-G that has a horizontal grid spacing of 20km at the equator and 70 vertical levels. The global ensemble initial conditions are derived from the ensemble transform Kalman filter (ETKF) method and a stochastic parameterisation scheme is also used in the global ensemble (Bowler et al., 2008). Each member of the limited area ensembles is obtained by one-way nesting from a MOGREPS-G forecast (dynamical downscaling). No additional stochastic perturbation scheme is used within the CP forecasts. The MetUM dynamical core solves a non-hydrostatic, deep atmosphere equation set using a semi-implicit, semi-Lagrangian time-stepping method and has 80 vertical levels. The science configuration of the dynamics and physics schemes of the atmosphere and land used for the CP simulations in tropical regions, “RAL1-T”, is documented in Bush et al. (2020). RAL1-T is the tropical subversion of RAL1 (Regional Atmosphere and Land configuration). The tropical version is required since the mid-latitude version (RAL1-M) has relatively weak turbulent mixing and stochastic perturbations which causes convection to initiate too early and convective cells to be small in the tropics. In order to account for this, RAL1-T uses the prognostic cloud prognostic condensate (PC2) cloud scheme (Wilson et al. 2008), has an interactive boundary layer free-atmosphere mixing length and has no stochastic boundary layer perturbations.

An 8.8km ensemble with parameterised convection is also included and uses the configuration of the operational global atmosphere version 6 (GA6) documented in Walters et al. (2017). This ensemble is referred to throughout as the GA ensemble.

Before analysis is carried out all forecasts are re-gridded to a common 9 km grid using an area-weighted conservative re-gridding scheme included in the python library “Iris” developed by the Met Office.

2.2. GPM-IMERG

To verify forecasts, rainfall is taken from The Integrated Multi-satellitE Retrievals for GPM (GPM-IMERG; Huffman et al. 2019). The product used is Level 3 half-hourly Final Run Precipitation at a resolution of 0.1° and combines precipitation estimates from GPM constellation satellites (see https://gpm.nasa.gov/missions/GPM/constellation) and Global Precipitation Climatology Centre (GPCC) precipitation rain-gauges. Precipitation estimates from passive microwave radiometers are combined with estimates from infrared data from geostationary weather satellites. Analyses of monthly GPCC gauge accumulations are then used to reduce biases in the multi-satellite monthly averages where available. Results using this dataset are referred to in this paper as “GPM” for simplicity.

Before any analysis takes place, the GPM precipitation field is also converted to an average hourly rain rate and is interpolated from a 0.1 degree grid to a common 9 km grid. Note that this is a slightly higher resolution than the native grid but the area-weighted conservative interpolation scheme maps between staggered grids with similar spacings without affecting integrals over larger scales.. It should also be noted that generally both heavy rainfall and rainfall over ocean tend to be underestimated by GPM (Tan and Duan 2017; Kahn and Maggioni 2019; Sunilkumar 2019; Tan and Santo 2019). Nonetheless, studies suggest that in the Philippines, unless examining very heavy rainfall (99th percentile), GPM captures rainfall relatively well (Sunilkumar 2019). The Singapore diurnal cycle of rainfall is well represented in

GPM (Tan et al. 2019) and Tan and Santo (2018) also concluded that GPM was a reliable precipitation source for a flooding event in Malaysia during 2014-2015. A recent study finds that GPM precipitation is similar to local precipitation for percentiles between the 85th and 95th and concludes that IMERG can be used for forecast evaluation of precipitation up to the 95th percentile (De Silva et al. 2021).

3.Methods

3.1. Fractions Skill Score

The Fractions Skill Score (FSS; Roberts & Lean, 2008) is a metric that compares two gridded fields and measures the degree of correspondence as a function of spatial scale. In order to calculate the FSS the re-gridded (see section 2) forecast and observation fields are converted into binary fields (1 or 0) based on values in each grid cell being above or below a threshold. In this study a threshold of the 95th percentile of rainfall is used, calculated using all grid cells (including zero values) over the re-gridded domains and the six months spanned by the forecasts. The threshold is calculated separately using GPM-IMERG and the ensemble forecast data for the GPM and forecast binary fields respectively. The threshold varies depending on time of day in the case of GPM and depending on lead time in the case of forecasts. Percentiles are used for the threshold choice in calculations (rather than a fixed rain rate threshold) to counteract the influence of intensity bias that effects the frequency of “event” occurrence.

A neighbourhood length (N; number of grid cells) is defined and is used to convert regularly gridded fields into fractions based on how many grid cells within neighbourhoods of size NxN have cell values exceeding the threshold (see figure 2 in Roberts & Lean, 2008). The FSS is given as:

$$FSS_{(N)} = 1 - \frac{MSE_{(N)}}{MSE_{(N)ref}} \tag{1}$$

where

$$MSE_{(N)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left[O_{(N)i,j} - M_{(N)i,j}\right]^2 \tag{2}$$

$$MSE_{(N)\mathrm{ref}} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} O_{(N)i,j}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} M_{(N)i,j}^2\right] \tag{3}$$

Such that $O_{(N)}$ and $M_{(N)}$ are the fields of fractions for neighbourhood length N for observations and models respectively. N_x and N_y are the number of neighbourhoods in the full domain along the longitude and latitude axes respectively.

This calculation can be computed for varying N to obtain FSS as a function of neighbourhood size. The aim is to allow for the fact that the location of features on the smallest scales in CP models, associated with convective updrafts, is highly unpredictable, while the probability that a certain fraction of neighbourhood area is occupied by precipitation cells (reaching an intensity threshold) may be predicted skilfully. Furthermore, in the absence of data assimilation for high resolution observations (such as radar) there is no information in the initial conditions on the smaller scales. The FSS value must lie between 1 and 0 with an FSS of 1 indicating a perfect forecast. FSS is expected to increase with neighbourhood size. A threshold of FSS can be used to determine the minimum spatial scale at which the forecast is considered skilful. This is taken to be 0.5 if observed rainfall frequency is sufficiently low (see Roberts & Lean, 2008).

Using this calculation as a basis, this study examines the ensemble-aggregated FSS (eFSS), the dispersion FSS (dFSS) and the localised FSS (LFSS). Because this study deals with an ensemble, it is required that FSS for all ensemble members is summarised, so to do this the average of $MSE_{(N)}$ and $MSE_{(N)ref}$ is taken over all ensemble members and all forecasts in the six-month period prior to calculation of the FSS using Equ. 1, resulting in the eFSS (Dey et al. 2014).

In order to obtain a measure of ensemble spread in relation to skill the dFSS is also calculated. The method for this is very similar to that of the eFSS except rather than compare ensemble members to observations they are compared to the control member of the ensemble (Rezacova et al. 2009; Dey et al. 2014). This provides a metric that demonstrates how similar precipitation patterns are between members of the ensemble. It is possible that model bias can influence metrics of spread (e.g. Wang et. al 2018). However, this FSS-based metric uses a percentile threshold, therefore accounting for precipitation intensity biases as part of the calculation.

Finally, a localised FSS (Woodhams et al. 2018) is used to determine regions in the domain that have particularly high or low values of skill relative to the full domain. For this metric the $MSE_{(N)}$ and $MSE_{(N)ref}$ are calculated for each neighbourhood (using the same domain-wide threshold precipitation rate, calculated from a percentile of the data from all forecasts at all grid points, as described above) but are not averaged over the domain prior to FSS calculation.

3.2. Persistence Forecast and ‘Shifted Forecast’

Forecast skill should be benchmarked against some more simple climatological or persistence-based reference forecast, since a useful forecast has to provide greater skill than these basic methods. The FSS is calculated for three variations of forecast over Oct 2018-Mar 2019. The three variations are:

- The standard CP ensemble forecasts.

- A ‘shifted ensemble forecast’, where forecast time stamps (for all lead times) are shifted by adding one day such that the forecast is verified against observations that occur a day later than the actual forecast verification times.
- A persistence forecast based on forward extrapolation from observations.

The ‘shifted forecast’ tests how much potential predictability comes from the similarity of the observed diurnal cycle from one day to the next (e.g., the pattern of precipitation over mountains and coasts). If the standard ensemble were no more skilful than the shifted forecast, it would imply that the model gives little useful information beyond its representation of the diurnal cycle in that flow regime and season.

The persistence forecast uses no model data but instead uses hourly GPM precipitation from the day prior to forecast initialisation and replicates this for every 24-hour period of the forecast. If the observed precipitation at each location were dominated by its diurnal cycle then the persistence forecast would be almost perfect, while the ‘shifted forecast’ would have the same skill as the standard ensemble and comparison of both these forecasts with observations would reflect model bias. However, longer timescale variability reduces the skill in the persistence forecast.

4. Results

4.1. Rainfall climatology and 4.5 km forecast bias

GPM 95th percentile rainfall amounts in the Malaysia and Indonesia domains have a pronounced diurnal variation (dashed lines in Fig. 2b). Each domain has 2 peaks in rainfall; one around 0800 UTC and one around 2000 UTC. Malaysia and Philippines local time is UTC+8 hours and Indonesia local time is UTC+7 hours so these times correspond to local early evening when rainfall is mostly over land and local early morning when rainfall is mostly over ocean (Mori et al. 2004). The morning peak is larger in the Malaysia domain than in the Indonesia domain. This is because rainfall off the northwest coast of Sumatra is relatively strong during early morning. The Philippines also has a variation throughout the day at similar times, but it is not as pronounced as that in the other two domains (Fig. 2b). Only the 95th percentile is shown here, but the diurnal variation is consistent across other intensities, such as average rainfall (not shown).

Percentiles are also calculated for forecasts. Note here that there is no averaging involved in this calculation; the percentiles are calculated over all ensemble members, all available forecasts and all grid points in the domain. For three-hourly accumulations, the forecasts tend to underestimate the 95th percentile rainfall in comparison to GPM (Fig. 2b; Table 1). A similar result was found by Woodhams et al. (2018) examining the 95th percentile rainfall in East Africa forecasts. In contrast, the 95th percentile and mean of 24-hourly accumulations (Fig 2a; c) are overestimated, as is the 3-hourly 99th percentile (Table 1). Woodhams et al. (2018) find that in Africa, at higher percentiles, such as 99%, the CP model has higher values than observed. This is consistent with other studies that find that CP models in the extra-tropics overestimate rainfall amounts and persistence at very high intensities, but underestimate rainfall at lower intensities (e.g. Kendon et al., 2012). However, bias varies with lead time and CP models tend to have large amounts of rainfall after spin-up. Note also that GPM is known to underestimate heavy rainfall events (Tan and Duan 2017; Sunilkumar 2019) which will impact the extent to which the model shows ‘bias’ in comparison, though De Silva et. al (2021) find that 95th percentile rainfall over the Maritime Continent in GPM is suitable for forecast verification.

Forecast precipitation rapidly increases during the initial hours of the forecast as the model spins up, then declines with lead time (Fig. 2b). Variations with the diurnal cycle are captured in the forecasts and peaks tend to occur at the correct times of day, however the difference between the morning and evening peaks in rainfall is larger than in GPM for the Malaysia and Indonesia domains. This suggests that rainfall over the ocean is not as well captured in the model, or is overestimated in GPM, or a combination of both. Since both heavy rainfall and rainfall over ocean tend to be underestimated by GPM (Tan and Duan 2017; Kahn and Maggioni 2019; Sunilkumar 2019) this suggests the bias largely lies with the ensemble.

The 95th percentile of precipitation in Malaysia forecasts also shows a relatively large decrease following the first day of forecasts (Fig. 2). Such abrupt declines between day 1 and day 3 are less evident in the Indonesia and Philippines domains (Fig. 2). This visible drop in forecast precipitation for Malaysia may be a result of the spatial distribution of precipitation in the Malaysia domain between ocean and land. Indeed, when performing the percentile analysis for only land points and only ocean points this decline is visible for the ocean rainfall but less so for the land rainfall (not shown). This may be a result of differing drivers of rainfall in the regions. In the Malaysia domain there is strong rainfall off the east coast of Sumatra, mainly driven by convergence because of Sumatran orography (Wu et al. 2009). Rainfall in the Indonesia domain is also largely driven by convergence lines between land masses, though Wu et al. (2009) suggest that differences in orography of nearby land masses may explain differences in ocean rainfall across Southeast Asia, such as between the east coast of Sumatra and the west coast of Borneo. Therefore, the representation of orography may be a large factor in diurnal cycle rainfall biases.

Other aspects of the ensemble may also contribute to biases, such as initial conditions, or the dynamical configuration and parameterizations of the parent ensemble. The drop in rainfall is also found by Dipankar et al. (2020) in similar CP forecasts covering Singapore and surrounding regions, such as Peninsular Malaysia and Sumatra (SINGV). Dipankar et al. (2020) find that rainfall in a version of the model driven by the global UM declines over ocean following 24 hours. This was not the case in a version of the model driven by the ECMWF operational deterministic forecast. It is suggested that this is primarily a result of the lateral boundary conditions and a dry bias in global UM forecasts. Further analysis would be required to determine the precise causes of bias here, and why they differ between regions. Examining the spatial pattern of extreme precipitation in subsequent figures will highlight the differences between the domains further. At longer lead times rainfall in Malaysia and Indonesia domains continues to decrease, whereas Philippines rainfall remains well captured at all lead times.

In GPM the highest values of 95th percentile rainfall are over ocean, particularly over the west coast of Sumatra (Fig. 3a;b). On the first day of forecasts the largest precipitation amounts occur over mountainous regions of Sumatra and Java (Fig. 3d;e), in contrast to GPM. There is some heavier rainfall off the west coast of Sumatra where the observed rainfall peaks, but they are smaller than the land rainfall amounts. The 95th percentile of rainfall in the Philippines domain is well replicated with comparable rainfall amounts to GPM in the south of the Philippines and the north-east region of Borneo (Fig. 3f). Rainfall over mountains is heavier in the forecasts than observed, explaining the discrepancies in the diurnal variations of rainfall (Fig. 2a; b).

As lead time increases the 95th percentile of rainfall off the west coast of Sumatra decreases (Fig 3g). Rainfall over the ocean here decreases at a faster rate than that over land between lead day 1 and lead day 3, explaining the large decrease in the rainfall shown in Fig. 2b after lead day 1. The rainfall in the ocean to the north of Java and in the Strait of Malacca does not show this decrease and even increases slightly. As mentioned previously it seems likely there is a meteorological phenomenon in this area that is more accurately captured than equivalent phenomena occurring west of Sumatra and results in this differing behaviour. One possibility is convergence lines in the Strait of Malacca resulting from land-sea breezes (Weller et al. 2017; Mohd Nor et al. 2020) that are relatively well represented and persist in the forecast. At longer lead times rainfall over mountainous regions continues to decrease (Fig. 3j;k). Rainfall amounts over the Philippines show less change as lead time increases (Fig 3l), as is also indicated in Fig. 2.

4.2. Skill of 4.5 km ensemble forecasts of daily precipitation accumulations

The spatial scales over which forecasts of 24-hourly accumulated rainfall can be considered skilful is assessed using the ensemble-aggregated Fractions Skill Score (eFSS) to determine the smallest scale for useful forecast information from each of the ensembles. Forecast skill of 24-hourly precipitation exceeds the skilful threshold of 0.5 (red line in Fig. 4) at spatial scales greater than around 150 km on the first day of forecasts for all three domains. Skill in the Philippines domain is slightly higher and so forecasts are considered skilful at spatial scales greater than around 50 km. Skill decreases as lead time increases such that by the end of the 5-day forecasts only spatial scales exceeding 350 km are considered skilful. It should be noted that here the ensemble-average FSS is examined for comparison with ensemble spread, not a metric based on the probabilistic output of the FSS. The skill of the probabilistic ensemble output is slightly higher (not shown) and so the probabilistic forecasts will be able to be displayed usefully at slightly smaller scales than these (approximately 50-100km smaller depending on lead time and region).

While this is useful information for forecasting daily rainfall totals, variations in rainfall with the diurnal cycle and between land and sea mean that forecast skill is likely to vary with both location and time of day, particularly in Malaysia and Indonesia. To further understand contributors to forecast skill it is important to also examine diurnal variations in forecast skill, as well as spatial variations in skill.

4.3. The role of the diurnal cycle in 4.5 km forecast skill

As mentioned previously, rainfall has a strong diurnal cycle during the day and can be defined by two peaks: one during local early morning when rainfall is mostly over the ocean (not shown), and one during local early evening when rainfall is over land (Fig. 5). In GPM during early evening the highest values of rainfall are over Sumatra and Borneo (Fig. 5a; b). The rainfall over Sumatra tends to be located around orographic features: mountains run down the west side of Sumatra. There are also large amounts of rainfall just off the west coast of Sumatra. There are smaller amounts of rainfall over Peninsular Malaysia, Java and the southern Philippines (Fig 5a;b;c).

For forecasts, twelve hours after forecast initialisation the spatial pattern of 95th percentile precipitation is similar to that in GPM but with varying amounts. Precipitation in Sumatra is most strong in the northwest (Fig. 5d), unlike GPM where larger amounts of precipitation run the full length of the island. Forecasts do not capture the precipitation off the west coast of Sumatra (Fig. 5d), though they still have some precipitation over the ocean, similar to GPM. Three days into the forecast precipitation over the ocean is decreased (Fig. 5g) but the precipitation over land is still relatively strong. Examining early morning in the forecasts also shows that precipitation over the ocean around Sumatra decreases following the first day forecast (not shown), explaining the large drop in Malaysia domain precipitation shown in Fig. 2a. By the final day of the forecast precipitation amounts have slightly decreased but the spatial pattern remains the same (Fig. 5j;k;l)

To examine how forecast skill depends on the diurnal cycle in the region, the skill of the three ensembles is examined. The shading in Fig. 6 demonstrates eFSS for the standard ensemble forecast, but the eFSS=0.5 “skilful threshold” contours are shown for all three forecast variations described in Section 3. Forecasts with skill exceeding 0.5, regions to the right of the colored lines in Fig. 6, are considered skilful for the given spatial scale and lead time. Forecasts with lines further to the left in the figure are therefore more skilful at smaller spatial scales.

Forecast skill of three-hourly precipitation is strongly tied to the diurnal cycle for the Malaysia forecasts (Fig. 6a) and the Indonesia forecasts (Fig. 6b). The Philippines forecast skill shows some link to the diurnal cycle but this is less pronounced than that of the other two domains (Fig. 6c). Skill tends to be largest in the daytime when precipitation is over land and smallest at night when precipitation is offshore. It is likely that this is due to precipitation that is constrained by topography, and therefore more predictable, during the day. For the first day forecasts are considered skilful at spatial scales greater than approximately 200 km (Fig. 6). After the first day forecast skill begins to decrease as lead time increases. On day 5 of the forecasts there is skill on spatial scales greater than around 400 km. These scales are comparable to analysis by Dey et al. (2014) that showed skill on scales greater than 200 km for forecasts of extreme rainfall in the UK using a similar nested 2.2 km ensemble. Woodhams et al. (2018) also find that a deterministic CP model forecast of extreme rainfall over East Africa is skilful at spatial scales around 275-300 km.

The reliance of skill on time of day leads to a question: how much of forecast skill is driven simply by diurnal variations? The ensemble forecast is compared with the other two forecast variations described in section 3.2 in order to gain an idea of the role of persisting weather in the forecast skill, and how much value is added by the dynamical ensemble forecast. It is found that the standard ensemble forecast is more skilful than the persistence forecast for all three domains, indicating that the ensemble forecast contains more information about the weather occurring in the future than can be inferred from local knowledge of the diurnal cycle observed on the day before making the forecast.

The ‘shifted forecast’ (see section 3.2) tests how much predictability comes from the characteristics of the diurnal cycle in the current flow regime and season. The standard forecast is substantially more skilful than the shifted forecast in all three domains (Fig. 6). This implies that substantial skill in the precipitation forecast is associated with phenomena with multi-day timescales, or that the forecast is skilfully predicting day-to-day variation in the characteristics of the diurnal cycle which are conditional on the large-scale environment. It is expected that as lead time increases the standard ensemble forecast skill must tend towards the skill of the shifted forecast as the model forecast increasingly becomes no better than a forecast of a different day.

Interestingly, for Malaysia (Fig. 6a) and Indonesia (Fig. 6b) the shifted forecast is more skilful than the persistence forecast at all lead times beyond the initial 3 hours.

During the first 24 hours the difference is small. Moving into day 2, the persistence forecast becomes substantially worse than the shifted forecast, implying that the precipitation field has multi-day variability that is captured by the model. This is particularly true for the Philippines (Fig. 6c), which has a large drop in the skill of the persistence forecast following day 1 relative to the model. In conclusion, even at five-days lead time, the ensemble forecast is still more skilful than a forecast based on persistence, as well as the shifted forecasts, that mainly capture the diurnal cycle.

Skill is increased when precipitation is located over land, perhaps because it is more constrained by orography. To examine this further, a Localised Fractions Skill Score (LFSS, see section 3.1) can be calculated (Fig. 7). It should be noted that while the red line in Fig. 7 indicates a threshold of 0.5 in keeping with Fig. 6, this should not be considered a threshold of ‘skilfulness’ in this context. Rather, the LFSS is a tool to understand the distribution of skill across the full domain, in relation to the eFSS.

During early evening higher skill tends to be located over land where the forecast has most precipitation, such as in the northwest region of Sumatra (Fig. 7a) and over Java (Fig. 7b). Comparing the spatial patterns of skill to the 95th percentile of precipitation (Fig. 3) shows that the spatial patterns are similar in these regions. There is much less skill over ocean for the Malaysia and Indonesia domains (Fig. 8a;b). Peaks in skill over ocean also tend to be lower than those over land (Fig. 8). This supports the idea that most skill comes from precipitation that is spatially constrained, and results in the diurnal variation of skill as precipitation moves from ocean to land.

There are also regions of high skill that have lower 95th percentile rainfall in both GPM and forecasts, such as the north of the Philippines (Fig. 7c). This skill is likely tied to synoptic-scale variability, i.e. tropical cyclones. This region of skill is also relatively independent of the diurnal cycle and is still present at other times of day (not shown), supporting this hypothesis.

As lead time increases, skill decreases rapidly but is still highest in mountainous regions and in the north of the Philippines (Fig. 7d-i). Localised skill analysis suggests that the diurnal cycle is more important for skill in Malaysia and Indonesia since rainfall is more spatially constrained by orographic features when over land. For the Philippines, skill is more likely to come from synoptic-scale systems that are not as dependent on time of day.

4.4. Ensemble spread-skill relationship for 4.5 km forecasts

The dispersion Fractions Skill Score (dFSS; Dey et al. 2014) is used to assess spatial differences between ensemble members, or "ensemble spread", in relation to forecast skill (see section 3.1). If the ensemble is "well calibrated" then the dFSS and eFSS should be the same. If the dFSS is smaller than the eFSS, then the ensemble is over-spread (under-confident) and if the dFSS is larger than the eFSS, then the ensemble is under-spread (over-confident). For Malaysia forecasts the dFSS is larger than eFSS (Fig. 9a;b) at all lead times and spatial scales, indicating that ensemble members are too similar to one another compared to the difference between the forecasts and observations. This is also true for Indonesia and the Philippines (Fig. 9c-f). Ensembles are under-spread in relation to the skill, and displacement errors in forecast rainfall features are typically larger than the differences between ensemble members. This is consistent with previous studies that find MetUM convective-scale ensembles to be under-spread (Porson et al. 2020, Cafaro et al. 2020), and being under-spread is a common error for CP ensembles generally (Schwartz et al. 2014; Beck et al. 2016; Raynaud and Bouttier 2017). Problems with spread can be linked to model errors (e.g.

Stensrud et al. 2000) and initial conditions and lateral boundary conditions from the parent ensemble. However, finding the underlying cause in this case would require further study.

The spread also varies with the diurnal cycle, consistent with the variation in ensemble skill. Ensemble spread is smallest (larger dFSS) during local evening, with dFSS peaking around 8-11pm (lead 12-15), particularly at longer lead times (Fig. 9a;c;e). Consistent with the skill results discussed previously, this is when precipitation is typically over land and is more spatially constrained, and therefore more similar between ensemble members. When precipitation is over ocean there is slightly more ensemble spread, indicated by reduced dFSS.

4.5. The role of resolution and convection parameterisation in forecast skill and spread

The 4.5 km ensembles examined in the earlier sections are part of a larger set of nested ensembles that include ensembles with a 2.2 and 8.8 km horizontal grid spacing, but the same model levels (see section 2.1), and provide an opportunity to examine the role of resolution in forecast skill. Analysis of an 8.8 km ensemble with parameterised convection (GA ensemble) is also included to examine how much skill is gained from partially resolving convection with the dynamical core. Resolution tends to play a fairly minor role in forecast skill for all three regions (Fig. 10). This is also the case when using a higher rainfall percentile threshold, such as the 99th percentile (not shown). 2.2 km CP, 4.5 km CP and 8.8 km CP ensembles all have similar variation in skill as a function of lead time and in this measure the 2.2 and 4.5 km forecasts are barely distinguishable. This was found to be true evaluating across the range of neighbourhood scales from grid-scale to 288 km (results are only shown for 144 km, the smallest scale for which the eFSS exceeds 0.5 for all domains in the first day). During early evening, when skill in the 4.5 km ensemble (blue line in Fig. 10) is highest, the 8.8 km CP ensemble (red line) has less skill in the Malaysia domain than the 2.2 km and 4.5 km ensembles (Fig. 10a), but in general the differences are small. A possible reason for this is that resolution may be less important at longer lead times where large-scale conditions dominate, and so resolution variations have little effect on skill.

There is a larger difference in the skill between the CP ensembles and the GA ensemble (green line in Fig. 10), though this is region dependent. For the Malaysia domain the GA ensemble is less skilful across all lead times, except for the first few hours of forecasts, during the spin-up period (Fig. 10a). Parameterised convection has less effect on skill in the Indonesia domain, though skill peaks earlier in the day compared to the CP ensemble skill. (Fig. 10b). A note to make here is that the diurnal cycle of precipitation in Malaysia and Indonesia is not as well captured by the GA ensemble (not shown); the observed peaks in rainfall tend to occur a few hours earlier in the GA ensemble, around 6 UTC (early afternoon local time) and 18UTC (early morning local time). The afternoon skill peak therefore coincides with a time when rainfall is increasing in observations and decreasing in the GA ensemble. This peak in skill in the Indonesia domain likely reflects a time when rainfall patterns match better than other times of day, but not for the correct reasons; rainfall in the GA ensemble does not move over ocean during night in the same way as in observations and the CP ensembles (not shown). Since rainfall in the Malaysia and Indonesia domains is largely driven by diurnal variations it is expected that the skill of the GA ensemble is less than the CP ensembles that capture diurnal variations more accurately.

While there is little difference between the skill in the CP ensembles it is important to note that rainfall is reduced in lower resolution 8.8 km CP ensembles (Table 1). Peaks also tend to occur slightly later in the day in these ensembles, particularly for the Malaysia domain, with the early morning peak hardly being captured at all (not shown). This is indicative of larger errors in the diurnal cycle of precipitation in the lower resolution ensembles, despite little difference being shown in the FSS measure of spatial skill. It is therefore important not to consider all CP ensembles to have equal value for operational forecasting despite similar FSS values.

In the Philippines domain there is almost no difference in skill between the CP ensembles at all resolutions (Fig. 10c). However, following the first 12 hours of the forecast the GA ensemble has slightly more skill than the CP ensembles, in contrast to the other two regions. A possible reason for this is that precipitation patterns associated with certain, likely larger scale, phenomena are represented more accurately in the GA ensemble. The CP forecasts spin up at initialisation and this may explain the higher skill of the GA model since less spin up may occur from the initial conditions. This is a surprising result since previous studies find that drivers of Philippines rainfall, such as tropical cyclones, can be represented more accurately by CP models (Bousquet et al. 2020). A full analysis of the differences in the representation drivers of rainfall between CP and GA ensembles is outside of the scope of this study, but this result certainly highlights an area of future research.

Resolution also has a small impact on ensemble spread (Fig 9d-f). All three CP ensembles are under-spread to a similar degree at the 144 km neighbourhood scale. Note here that ensemble spread is shown over a sub-region with boundaries 2 degrees away from the 2.2 km boundaries to avoid a larger influence of lateral boundary conditions on the 2.2 km ensemble than the larger domain ensembles. When performing analysis over the full 2.2 km ensemble domain the spread of the 2.2 km ensemble is smaller than that of the 4.5 km and 8.8 km CP ensembles (not shown) as a result of the role of lateral boundary conditions on ensemble spread. Ensemble spread in the GA ensemble is relatively similar to the CP ensembles at most times of day but reduces during early afternoon (local time; approximately 4 UTC) when observed rainfall is moving from over ocean to over land (Fig. 10d; e). This suggests that the ensemble spread in the spatial patterns of rainfall is much less than the CP ensemble as a result of the parameterisations used. In the Philippines, where rainfall is not as strongly driven by the diurnal cycle, ensemble spread is similar in the GA ensemble and the CP ensembles, particularly during the first two to three days of the forecast.

5. Summary and Conclusions

Convection-permitting (CP) ensemble forecasts in the Tropics have, until now, been relatively uncommon. Recently, the UK Met Office have developed such systems over Singapore (Porson et al. 2019) and East Africa (Cafaro et al. 2020). Model development has now extended to further domains in Southeast Asia, covering Malaysia, Indonesia and the Philippines. In this study the skill of newly-developed CP ensembles over these regions has been examined using the Fractions Skill Score (FSS), with a particular focus on the role of the diurnal cycle and quantifying the spread of the ensemble in relation to skill.

The skill of forecasts of precipitation in the Malaysia and Indonesia domains are strongly linked to the diurnal cycle, peaking during local early evening when rainfall is over land and dropping in local early morning when rainfall is over ocean (Fig. 6).

The diurnal variation suggests that orographic and coastal features play a role in skill, such that precipitation over mountainous regions is spatially constrained and therefore better captured by the model. Land-sea contrast is also important to the characteristics of the rainfall. Examination of maps of the Localised Fractions Skill Score (LFSS) confirms this, as the largest skill tends to be over mountainous regions and skill is much lower over the ocean (Fig. 7; Fig. 8). The diurnal cycle is less strong in the Philippines domain, suggesting that other modes of variability drive precipitation there. Skill decreases as lead time increases, and forecasts are skilful at spatial scales greater than 400 km for all lead times and all domains, providing a suggested scale for future forecast display.

The large role played by the diurnal cycle motivates the question as to how much of the skill is potentially predictable simply by representing well the diurnal cycle in the region. Vogel et al. (2018) found that forecasts of precipitation in West Africa that were based on climatological precipitation outperformed operational ensemble forecasts that used parameterised convection schemes. Here, comparisons are made to persistence forecasts to examine how the CP ensembles perform relative to a forecast from recently observed rainfall. In this case all three ensembles have much higher skill than a persistence forecast, suggesting that there is significant added value in the CP ensemble compared to a forecast based on observed rainfall from the previous day. Skill for a ‘shifted forecast’, in which the forecast date stamps are shifted one day later and therefore verified against observations from one day after the true verification date, is also presented to determine how much of the model skill is related to simulation of the diurnal cycle. The skill of the shifted forecast is larger than the persistence forecast, but less than the standard ensemble forecast, for lead times longer than one day. This suggests that increased CP ensemble forecast skill is associated with phenomena characterised by multi-day timescales that are captured by the model, highlighting the usefulness of the CP ensemble in this region.

Variation of horizontal grid spacing, across 2.2, 4.5 and 8.8 km, plays a fairly small role in ensemble skill assessed using the Fractions Skill score over the full range of neighbourhood scales. These results agree with previous studies that also find 1-2km resolution forecasts rainfall in the United States to perform as well as 4km forecasts (Kain et al. 2008; Schwarz et al 2009; Loken et al. 2017). However, other studies found that increased resolution resulted in higher forecast skill for United States rainfall (Schwarz et al. 2017; Shwartz and Sobash 2019). Schwartz and Sobash (2019) suggest that this could be because of higher quality initial conditions than in previous studies. Given the computational cost of higher resolution forecasts this is a result to be considered in future forecast model development focussing on this region

The 8.8 km GA ensemble using a parameterised convection scheme has less skill compared to the CP ensembles, but only in the regions where rainfall patterns are strongly driven by the diurnal cycle, i.e. Malaysia and Indonesia. In the Philippines region, CP ensembles have slightly lower skill than the GA ensemble suggesting that CP forecasts add limited value when forecasting spatial patterns of rainfall in the Philippines region. This is likely due to the higher latitude location of the Philippines and hence stronger synoptic scale forcing of rainfall and weaker dependence on local thermal contrasts in forcing convective systems. Also, there are no additional high-resolution observations assimilated into the CP forecasting system and the initial conditions for each forecast member are obtained simply by interpolating the parent global member onto a finer grid. The only benefit that could come from the higher resolution is an improvement in the representation of the lower boundary and weather system dynamics. The higher resolution forecasts need to spin up and this may explain the apparently higher skill of the 8.8 km model with parameterised convection, rather than explicit. The model dynamics with parameterisation is more similar to the global model parent and therefore less spin up may occur from the initial conditions. An important note and caveat to make here is that while resolution plays a small role in skill measured using the FSS there are many other reasons why a higher-resolution model is preferable. Other aspects of rainfall, aside from spatial pattern, such as intensity and timing, are of importance and should be considered when choosing models for an operational forecast.

The spread of the ensemble is examined using the dispersion Fractions Skill Score (dFSS), which compares ensemble member pairs to obtain a measure of the spatial differences within the ensemble. Spread also varies with the diurnal cycle such that there is less ensemble spread when precipitation is over land, again supporting the idea that convection is more constrained by orography at that time (Fig. 9). The ensemble is under-dispersive. Spread is 59%, 61% and 33% less than ensemble mean forecast error on average over 4.5 km forecasts for Malaysia, Indonesia, and Philippines domains respectively, confirming that these ensembles suffer from a persistent drawback of CP ensembles. While using percentiles for the threshold choice in calculations (rather than a fixed amount threshold) somewhat counteracts the influence of rainfall intensity bias, there are further possible causes of reduced ensemble spread, such as the influence of initial and lateral boundary conditions and model errors. Studies have found various ways to improve ensemble spread, such as combining ensembles (Beck et al. 2016) and time-lagging ensemble member forecasts (Porson et al. 2020). Porson et al. (2019) assessed the difference in skill and spread in an ensemble covering Singapore nested in two different global ensembles. They find that spread is sensitive to initial conditions at the beginning of the forecast and lateral boundary conditions towards the end of the forecast. While beyond the scope of the present study, investigating why the ensembles analysed here are under-spread would be an interesting topic of future research. Assimilation of high resolution data to generate a higher resolution regional analysis would be expected to improve initial conditions for the CP ensembles, introducing spread associated with observed mesoscale features and acting to reduce artificial forecast spin up effects which may reduce skill. However, assimilation in the convection-dominated regime in the tropics is challenging, due to convective instability and the high variance on small spatial and temporal scales, and prototype systems assimilating satellite radiance data are currently under development in the Southeast Asia region (Heng et al. 2020).

Despite the lack of sufficient spread in the CP ensembles shown here, the forecasts have significant skill beyond that shown by a persistence forecast based on observed climatology. Aside from improvements to ensemble spread, future work with these forecasts should focus on sources of skill in the three different regions, including the role of large-scale conditions in forecast skill. Understanding drivers of forecast skill will lead to improved forecasts of high-impact weather in the region and should therefore be a high priority in future.

Acknowledgements

This work and SF, THAF, CEH, JM and THMS were supported by the Weather and Climate Science for Services Partnership (WCSSP) Southeast Asia as part of the Newton Fund. We thank three anonymous reviewers for insightful comments and suggestions on this work. The GPM-IMERG data were provided by the NASA

Goddard Space Flight Center's Precipitation Measurement Missions Science Team and Precipitation Processing System, which develop and compute GPM-IMERG as a contribution to GPM, and archived at the NASA GES DISC.

References

Beck, J., F. Bouttier, L. Wiegand, C. Gebhardt, C. Eagle, and N. Roberts, 2016: Development and verification of two convection-allowing multi-model ensembles over Western Europe. *Q. J. R. Meteorol. Soc.*, **142**, 2808–2826, https://doi.org/10.1002/qj.2870.

Bednarczyk, C. N., and B. C. Ancell, 2015: Ensemble sensitivity analysis applied to a southern plains convective event. *Mon. Weather Rev.*, **143**, 230–249, https://doi.org/10.1175/MWR-D-13-00321.1.

Birch, C. E., J. H. Marsham, D. J. Parker, and C. M. Taylor, 2014a: The scale dependence and structure of convergence fields preceding the initiation of deep convection. *Geophys. Res. Lett.*, **41**, 4769–4776, https://doi.org/10.1002/2014GL060493.

——, D. J. Parker, J. H. Marsham, D. Copsey, and L. Garcia-Carreras, 2014b: A seamless assessment of the role of convection in the water cycle of the West African Monsoon. *J. Geophys. Res. Atmos.*, **119**, 2890–2912, https://doi.org/10.1002/2013JD020887.

——, M. J. Roberts, L. Garcia-Carreras, D. Ackerley, M. J. Reeder, A. P. Lock, and R. Schiemann, 2015: Sea-breeze dynamics and convection initiation: The influence of convective parameterization in weather and climate model biases. *J. Clim.*, **28**, 8093–8108, https://doi.org/10.1175/JCLI-D-14-00850.1.

——, and Coauthors, 2016: Scale Interactions between the MJO and the Western

Maritime Continent. *J. Clim.*, **29**, 2471–2492, https://doi.org/10.1175/JCLI-D-15-0557.1.

Bousquet, O, D. Barbary, S. Bielli, and Coauthors. 2020 An evaluation of tropical cyclone forecast in the Southwest Indian Ocean basin with AROME-Indian Ocean convection-permitting numerical weather predicting system. *Atmos Sci Lett.* **21**:e950. https://doi.org/10.1002/asl2.950

Bowler N, A. Arribas, K. Mylne, K. Robertson, S. Beare. 2008. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc*. **134**:703–72

Bush, M., and Coauthors. 2020: The first Met Office Unified Model–JULES Regional Atmosphere and Land configuration, RAL1, *Geosci. Model Dev.*, **13,** 1999–2029, https://doi.org/10.5194/gmd-13-1999-2020.

Cafaro, C., T. H. A. Frame, J. Methven, N. Roberts,. and J. Bröcker, 2019. The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble. Q. J. R. Meteorol. Soc., **145**(721), 1780-1798.

Cafaro, C., and Coauthors, 2020: Do convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical East Africa ?, *Wea. Forecasting*, under review.

Chang, C. P., P. A. Harr, and H. J. Chen, 2005: Synoptic disturbances over the equatorial South China Sea and western maritime continent during boreal winter. *Mon. Weather Rev.*, **133**, 489–503, https://doi.org/10.1175/MWR-2868.1.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016:

Convection-permitting models: A step-change in rainfall forecasting. *Meteorol. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

De Silva, N., and Coauthors, 2021: Validation of GPM IMERG extreme precipitation in the Maritime Continent by station and radar data (In Preparation)

Dey, S. R. A., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Weather Rev.*, **142**, 4091–4107, https://doi.org/10.1175/MWR-D-14-00172.1.

Dipankar, A., and coauthors, 2020: SINGV: A convective-scale weather forecast model for Singapore. *Q J R Meteorol Soc*. 1– 16. https://doi.org/10.1002/qj.3895

Ferrett, S., G. Yang, S. J. Woolnough, J. Methven, K. Hodges, and C. E. Holloway, 2020: Linking extreme precipitation in Southeast Asia to equatorial waves. *Q. J. R. Meteorol. Soc.*, **146**, 665–684, https://doi.org/10.1002/qj.3699.

Gebhardt C., S.E. Theis, M. Paulat, Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, *Atmospheric Research*, **100**, 168-177, https://doi.org/10.1016/j.atmosres.2010.12.008.

Golding, B. W., and Coauthors, 2014: Forecasting Capabilities for the London 2012 Olympics. *Bull. Amer. Meteor. Soc.*, **95**, 883–896, https://doi.org/10.1175/BAMS-D-13-00102.1.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Q.J.R. Meteorol. Soc.*, **143**: 2846-2861. doi:10.1002/qj.3135

Hamada, J.-I., S. Mori, H. Kubota, M. D. Yamanaka, U. Haryoko, S. Lestari, R. Sulistyowati, and F. Syamsudin, 2012: Interannual Rainfall Variability over Northwestern Jawa and its Relation to the Indian Ocean Dipole and El Ni^|^ntilde;o-Southern Oscillation Events. *SOLA*, **8**, 69–72, https://doi.org/10.2151/sola.2012-018.

Hanley, K. E., D. J. Kirshbaum, N. M. Roberts, and G. Leoncini, 2013: Sensitivities of a Squall Line over Central Europe in a Convective-Scale Ensemble. *Mon. Weather Rev.*, **141**, 112–133, https://doi.org/10.1175/MWR-D-12-00013.1.

Heng, B. C. P, and Coauthors, 2020: SINGV-DA: A data assimilation system for convective-scale numerical weather prediction over Singapore. *Q J R Meteorol Soc.*, **146**, 1923– 1938, https://doi.org/10.1002/qj.3774

Hohenegger, C., and C. Schar, 2007: Atmospheric Predictability at Synoptic Versus Cloud-Resolving Scales. *Bull. Amer. Meteor. Soc*., **88**, 1783–1794, https://doi.org/10.1175/BAMS-88-11-1783.

Holloway, C. E., S. J. Woolnough, and G. M. S. Lister, 2013: The effects of explicit versus parameterized convection on the MJO in a large-domain high-resolution tropical case study. Part I: Characterization of large-scale organization and propagation. *J. Atmos. Sci.*, **70**, 1342–1369, https://doi.org/10.1175/JAS-D-12-0227.1.

Huffman, G. J., and Coauthors, 2019: *NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG), Algorithm Theoretical Basis Document (ATBD), Version 06*. 38 pp. https://gpm.nasa.gov/sites/default/files/document_files/IMERG_ATBD_V06.pdf

Johnson, S. J., and Coauthors, 2016: The resolution sensitivity of the South Asian monsoon and Indo-Pacific in a global 0.35° AGCM. *Clim. Dyn.*, **46**, 807–831, https://doi.org/10.1007/s00382-015-2614-1.

Kain, J. S., Weiss, S. J., Bright, D. R., Baldwin, M. E., Levit, J. J., Carbin, G. W., Schwartz, C. S., Weisman, M. L., Droegemeier, K. K., Weber, D. B., & Thomas, K. W. 2008: Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP, *Weather and Forecasting*, **23**(5), 931-952

Khan, S. and V. Maggioni, 2019: Assessment of Level-3 Gridded Global Precipitation Mission (GPM) Products Over Oceans. *Remote Sens*, **11**, 255.

Kendon, E. J., N. M. Roberts, C. A. Senior, and M. J. Roberts, 2012: Realism of rainfall in a very high-resolution regional climate model. *J. Clim.*, **25**, 5791–5806, https://doi.org/10.1175/JCLI-D-11-00562.1.

Lestari, S., A. King, C. Vincent, D. Karoly, and A. Protat, 2019: Seasonal dependence of rainfall extremes in and around Jakarta, Indonesia. *Weather Clim. Extrem.*, **24**, 100202, https://doi.org/10.1016/j.wace.2019.100202.

Lim, S. Y., C. Marzin, P. Xavier, C. P. Chang, B. Timbal, S. Yee Lim, C. Marzin, and P. Xavier, 2017: Impacts of Boreal Winter Monsoon Cold Surges and the Interaction with MJO on Southeast Asia Rainfall. *J. Clim.*, **30**, 4267–4281, https://doi.org/10.1175/JCLI-D-16-0546.1.

Lin, J. L., and Coauthors, 2006: Tropical intraseasonal variability in 14 IPCC AR4 climate models. Part I: Convective signals. *J. Clim.*, **19**, 2665–2690, https://doi.org/10.1175/JCLI3735.1.

Loken, E. D., Clark, A. J., Xue, M., and Kong, F. 2017: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble, *Weather and Forecasting*, **32**(4), 1403-1421

Love, B. S., A. J. Matthews, and G. M. S. Lister, 2011: The diurnal cycle of precipitation over the Maritime Continent in a high-resolution atmospheric model. *Q. J. R. Meteorol. Soc.*, **137**, 934–947, https://doi.org/10.1002/qj.809.

Matsumoto, J., and Coauthors, 2017: An overview of the Asian Monsoon Years 2007-2012 (AMY) and multi-scale interactions in the extreme rainfall events over the Indonesian maritime continent. *World Scientific Series on Asia-Pacific Weather and Climate*, Vol. Volume 9 of, World Scientific Publishing Co. Pte Ltd, 365–385.

Met Office, 2010-2020: Iris: A Python library for analysing and visualising meteorological and oceanographic data sets; v2.4, http://scitools.org.uk/

Miura, H., M. Satoh, T. Nasuno, A. T. Noda, and K. Oouchi, 2007: A Madden-Julian oscillation event realistically simulated by a global cloud-resolving model. *Science (80-.).*, **318**, 1763–1765, https://doi.org/10.1126/science.1148443.

Mittermaier, M., N. Roberts, and S. A. Thompson, 2013; A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteor. Appl.* **20,** 176-186, https://doi.org/10.1002/met.296

Mohd Nor, M. F. F., C. E. Holloway, and P. M. Inness, 2020: The Role of Local Orography on the Development of a Severe Rainfall Event over Western Peninsular Malaysia: A Case Study. Mon. Wea. Rev., 148, 2191–2209, https://doi.org/10.1175/MWR-D-18-0413.1.

Mori, S., H. Jun-Ichi, M. D. Yamanaka, N. Okamoto, F. Murata, N. Sakurai, and H. Hashiguchi, 2004: *Diurnal Land-Sea Rainfall Peak Migration over Sumatera Island, Indonesian Maritime Continent, Observed by TRMM Satellite and Intensive Rawinsonde Soundings*. American Meteorological Society, 2021–2039 pp.

Neale, R., and J. Slingo, 2003: *The Maritime Continent and Its Role in the Global Climate: A GCM Study*. American Meteorological Society, 834–848 pp.

Pearson, K. J., R. J. Hogan, R. P. Allan, G. M. S. Lister, and C. E. Holloway, 2010: Evaluation of the model representation of the evolution of convective systems using satellite observations of outgoing longwave radiation. *J. Geophys. Res.*, **115**, D20206, https://doi.org/10.1029/2010JD014265.

——, G. M. S. Lister, C. E. Birch, R. P. Allan, R. J. Hogan, and S. J. Woolnough, 2014: Modelling the diurnal cycle of tropical convection across the 'grey zone.' *Q. J. R. Meteorol. Soc.*, **140**, 491–499, https://doi.org/10.1002/qj.2145.

Peatman, S. C., A. J. Matthews, and D. P. Stevens, 2014: Propagation of the Madden-Julian Oscillation through the Maritime Continent and scale interaction with the diurnal cycle of precipitation. *Q. J. R. Meteorol. Soc.*, **140**, 814–825, https://doi.org/10.1002/qj.2161.

Peatman, S. C., Methven, J. and Woolnough, S. J., 2018: Isolating the effects of moisture entrainment on convectively coupled equatorial waves in an aquaplanet GCM. *J. Atmos. Sci.*, **75**, 3139-3157. https://doi.org/10.1175/JAS-D-18-0098.1.

——, ——, and ——, 2015: Propagation of the Madden–Julian Oscillation and scale interaction with the diurnal cycle in a high-resolution GCM. *Clim. Dyn.*, **45**, 2901–2918, https://doi.org/10.1007/s00382-015-2513-5.

Porson, A. N., S. Hagelin, D. F. A. Boyd, N. M. Roberts, R. North, S. Webster, and J. C. Lo, 2019: Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore. *Q. J. R. Meteorol. Soc.*, **145**, 3004–3022, https://doi.org/10.1002/qj.3601.

——, and Coauthors, 2020: Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Q. J. R. Meteorol. Soc.*, qj.3844, https://doi.org/10.1002/qj.3844.

Qian, J. H., 2008: Why precipitation is mostly concentrated over islands in the maritime continent. *J. Atmos. Sci.*, **65**, 1428–1441, https://doi.org/10.1175/2007JAS2422.1.

Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Q. J. R. Meteorol. Soc.*, **143**, 3037–3047, https://doi.org/10.1002/qj.3159.

Rezacova, D., P. Zacharov, and Z. Sokol, 2009: Uncertainty in the area-related QPF for heavy convective precipitation. *Atmos. Res.*, **93**, 238–246, https://doi.org/10.1016/j.atmosres.2008.12.005.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245-1258.

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

Sato, T., H. Miura, M. Satoh, Y. N. Takayabu, and Y. Wang, 2009: Diurnal cycle of precipitation in the tropics simulated in a global cloud-resolving model. *J. Clim.*, **22**, 4809–4826, https://doi.org/10.1175/2009JCLI2890.1.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., & Coniglio, M. C. 2009: Next-Day Convection-Allowing WRF Model Guidance: A Second Look at 2-km versus 4-km Grid Spacing, *Monthly Weather Review*, **137**(10), 3351-3372

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble kalman filter. *Weather Forecast.*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's Experimental Real-Time Convection-Allowing Ensemble Prediction System. *Wea. Forecasting*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

Schwartz, C. S., G. S. Romine, K. R. Fossell, A. Sobash, and M. L. Weisman. 2017: Toward 1-km Ensemble Forecasts over Large Domains, *Monthly Weather Review*, **145**(8), 2943-2969

Schwartz, C. S., and R. A. Sobash 2019: Revisiting Sensitivity to Horizontal Grid Spacing in Convection-Allowing Models over the Central and Eastern United States, *Monthly Weather Review*, **147**(12), 4411-4435

Stensrud, D. J., J. Bao, and T. T. Warner, 2000: Using Initial Condition and Model Physics Perturbations in Short-Range Ensemble Simulations of Mesoscale Convective Systems, *Monthly Weather Review*, **128**(7), 2077-2107

Sun, X., and Coauthors, 2020: A subjective and objective evaluation of model forecasts of sumatra squall events. *Weather Forecast.*, **35**, 489–506, https://doi.org/10.1175/WAF-D-19-0187.1.

Supari, F. Tangang, E. Salimun, E. Aldrian, A. Sopaheluwakan, and L. Juneng, 2018: ENSO modulation of seasonal rainfall and extremes in Indonesia. *Clim. Dyn.*, **51**, 2559–2580, https://doi.org/10.1007/s00382-017-4028-8.

Takahashi, H. G., and T. Yasunari, 2008: Decreasing Trend in Rainfall over Indochina during the Late Summer Monsoon: Impact of Tropical Cyclones, *J. Meteorol. Soc. Jpn.*, **86**, 429-438

Tan, M.L. and Z. Duan, 2017: Assessment of GPM and TRMM Precipitation Products over Singapore. *Remote Sens.*, **9**, 720.

Tan, M.L. and H. Santo, 2018: Comparison of GPM IMERG, TMPA 3B42 and PERSIANN-CDR satellite precipitation products over Malaysia. *Atmos. Res.*, **202**, 63-76.

Tan, J., Huffman, G. J., Bolvin, D. T., & Nelkin, E. J. 2019: Diurnal cycle of IMERG V06 precipitation. *Geophysical Research Letters*, **46**, 13584– 13592.

https://doi.org/10.1029/2019GL085395

Villafuerte, M. Q., and J. Matsumoto, 2015: Significant influences of global mean temperature and ENSO on extreme rainfall in Southeast Asia. *J. Clim.*, **28**, 1905–1919, https://doi.org/10.1175/JCLI-D-14-00531.1.

Vincent, C. L., and T. P. Lane, 2018: Mesoscale variation in diabatic heating around Sumatra, and its modulation with the Madden-Julian oscillation. *Mon. Weather Rev.*, **146**, 2599–2614, https://doi.org/10.1175/MWR-D-17-0392.1.

Vogel, P., P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting, 2018: Skill of global raw and postprocessed ensemble predictions of rainfall over Northern Tropical Africa. *Weather Forecast.*, **33**, 369–388, https://doi.org/10.1175/WAF-D-17-0127.1.

Wang, J., Chen, J., Du, J., Zhang, Y., Xia, Y., & Deng, G., 2018: Sensitivity of Ensemble Forecast Verification to Model Bias, *Mon. Weather Rev.*, **146**, 781-796

Wilson, D. R., A. C. Bushell, A. M. Kerr-Munslow, J. D. Price, and C. J. Morcrette, 2008: PC2: A prognostic cloud fraction and condensation scheme. I: Scheme description, *Q. J. Roy. Meteor. Soc.*, **134**, 2093–2107, https://doi.org/10.1002/qj.333

Woodhams, B. J., C. E. Birch, J. H. Marsham, C. L. Bain, N. M. Roberts, and D. F. A. Boyd, 2018: What is the added value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Mon. Weather Rev.*, **146**, 2757–2780, https://doi.org/10.1175/MWR-D-17-0396.1.

Wu, P., M. Hara, J. Hamada, M. D. Yamanaka, and F. Kimura, 2009: Why a Large Amount of Rain Falls over the Sea in the Vicinity of Western Sumatra Island during Nighttime. *J. Appl. Meteor. Climatol*., **48**, 1345–1361, https://doi.org/10.1175/2009JAMC2052.1.

Wu, P., A. A. Arbain, S. Mori, J. Hamada, M. Hattori, F. Syamsudin, and M. D. Yamanaka, 2013: The Effects of an Active Phase of the Madden-Julian Oscillation on the Extreme Precipitation Event over Western Java Island in January 2013. *SOLA*, **9**, 79–83, https://doi.org/10.2151/sola.2013-018.

Xavier, P., R. Rahmat, W. K. Cheong, and E. Wallace, 2014: Influence of Madden-Julian Oscillation on Southeast Asia rainfall extremes: Observations and predictability. *Geophys. Res. Lett.*, **41**, 4406–4412, https://doi.org/10.1002/2014GL060241.

Yamanaka, M. D., 2016: Physical climatology of Indonesian maritime continent: An outline to comprehend observational studies. *Atmos. Res.*, **178**–**179**, 231–259, https://doi.org/10.1016/j.atmosres.2016.03.017.

Yang, G.-Y., and J. Slingo, 2001: *The Diurnal Cycle in the Tropics*. American Meteorological Society, 784–801 pp.

Region	GPM/model resolution	95^{th} pc	99^{th} pc
Malaysia	GPM	1.85	5.62
	2.2	1.49	7.13
	4.5	0.78	6.45
	8.8	0.62	6.04
	8.8 (GA)	1.41	2.78
Indonesia	GPM	1.35	4.76
	2.2	0.52	5.69
	4.5	0.48	5.87
	8.8	0.12	4.21
	8.8 (GA)	1.23	2.57
Philippines	GPM	0.34	2.66
	2.2	0.18	3.21
	4.5	0.27	4.34
	8.8	0.11	2.93
	8.8 (GA)	0.58	2.19

Table 1: *The average over all lead times of 95^{th} and 99^{th} percentiles for 3-hourly accumulations of GPM and 2.2 km CP, 4.5 km CP, 8.8km CP, 8.8km GA forecasts for all three regions (using 2.2 km domains in Fig. 1). All values are converted to rain rate in mm hr^{-1}.*

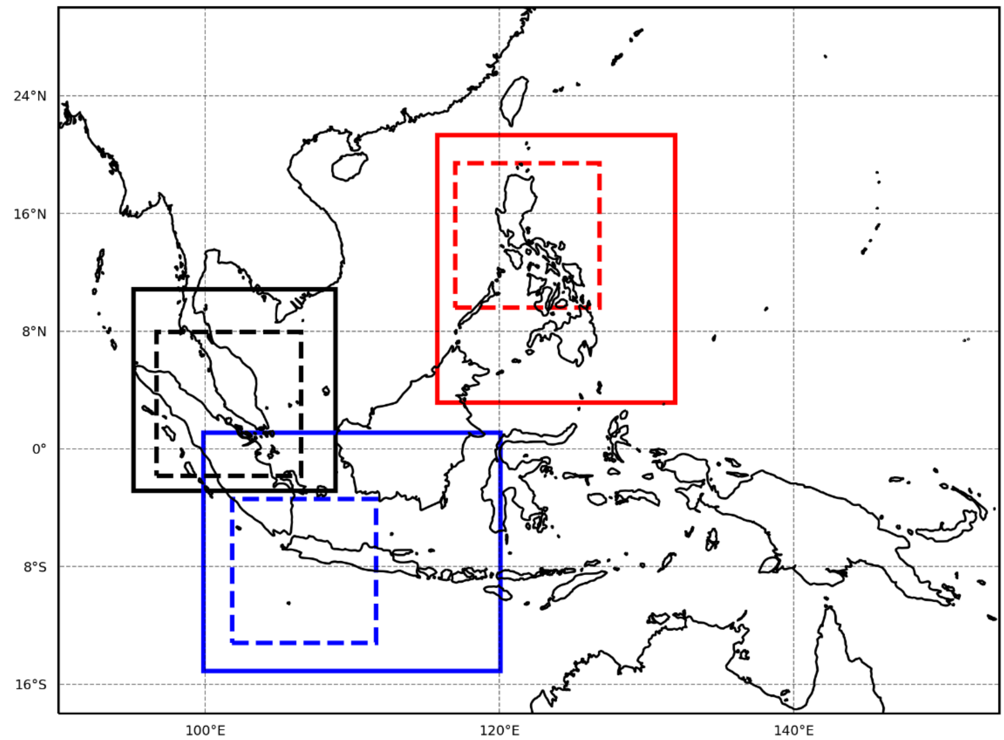

Figure 1: Malaysia (black boxes), Indonesia (blue boxes) and Philippines (red boxes) forecast domains for 2.2 km forecasts (dashed lines) and 4.5 km forecasts (solid lines). The full domain shows the 8.8 km domain. For analysis, the 8.8 km forecasts are subset to the relevant smaller regional domains.

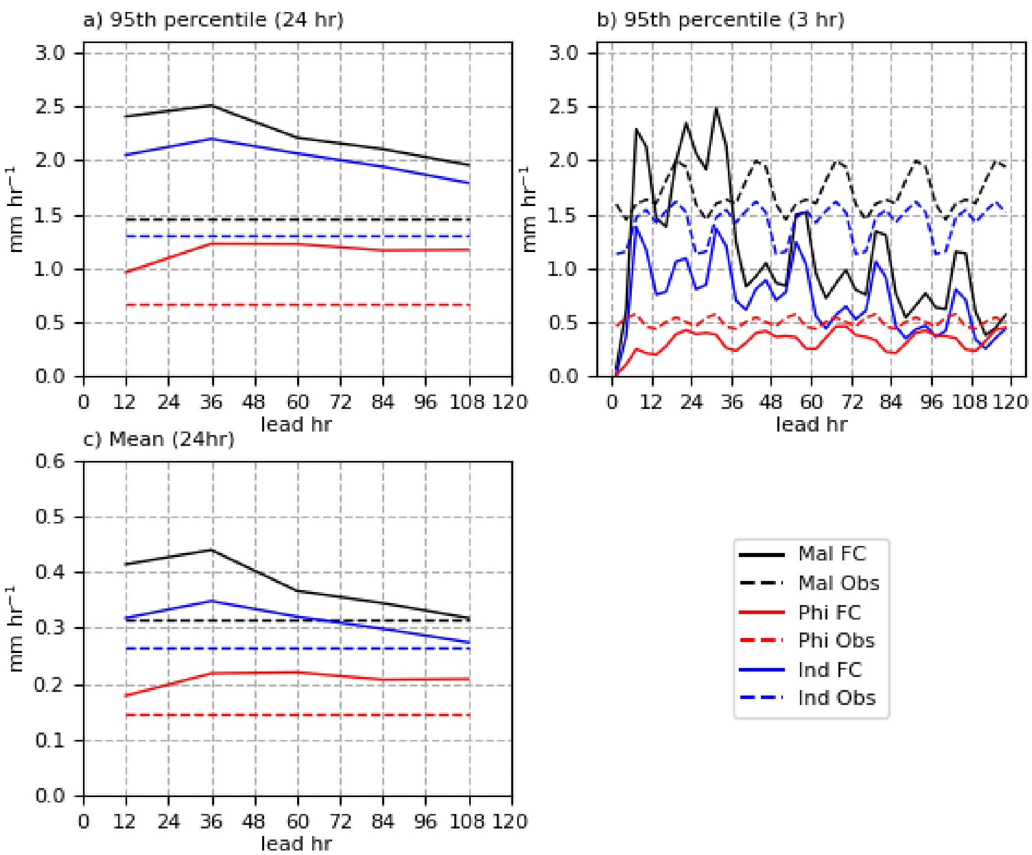

Figure 2: GPM-IMERG (dashed lines) and 4.5 km forecast (solid lines) 95th percentile for a) 24-hourly accumulations, b) 3-hourly accumulations and c) mean of 24-hourly precipitation accumulations as a function of forecast lead time. All values are converted to rain rate in mm hr^{-1}. Values shown here are the thresholds used for FSS calculation (see section 3). For forecasts the percentile are calculated over all ensemble members, all forecasts and all gridpoints in the domain.

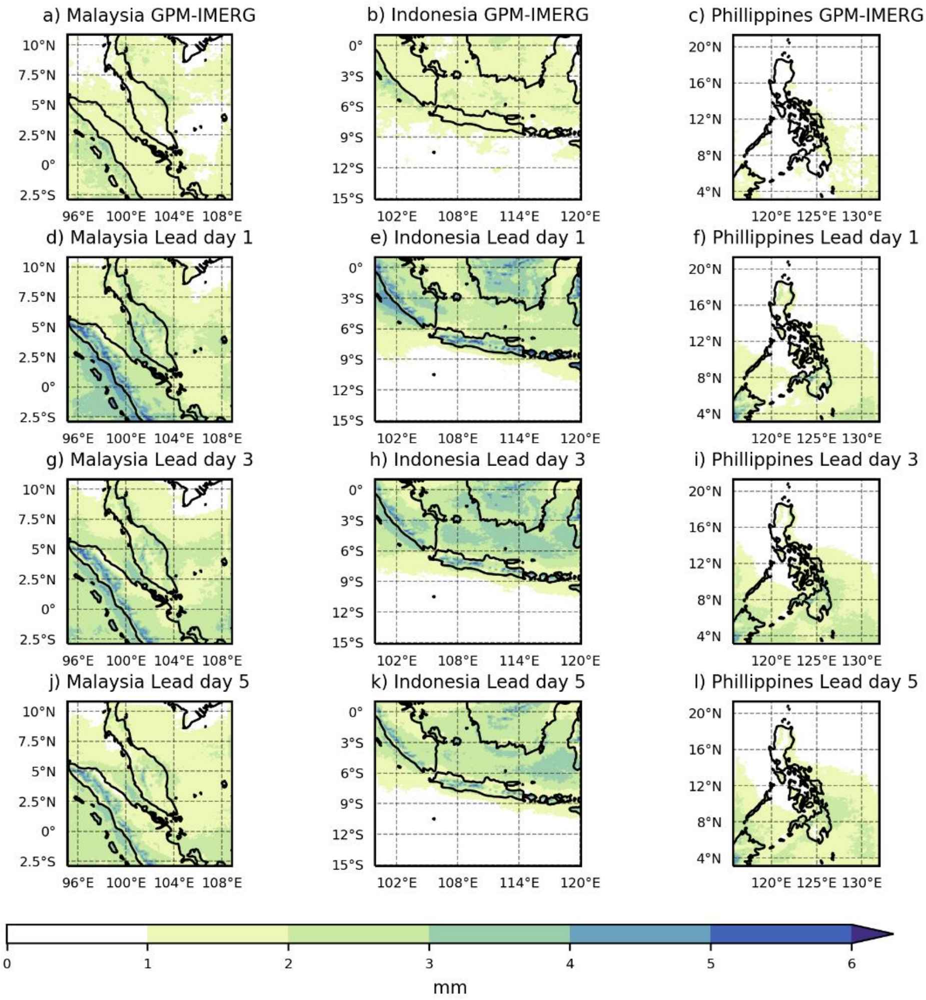

Figure 3: GPM-IMERG and 4.5 km forecast 95th percentile precipitation for Malaysia (left panels), Indonesia (centre panels) and Philippines (right panels). The 95th percentile is calculated for 24-hourly accumulations and is displayed as mm hr^{-1}. Forecast precipitation is shown for lead day 1 (hours 00-24), lead day 3 (hours 48-72) and lead day 5 (hours 96-120).

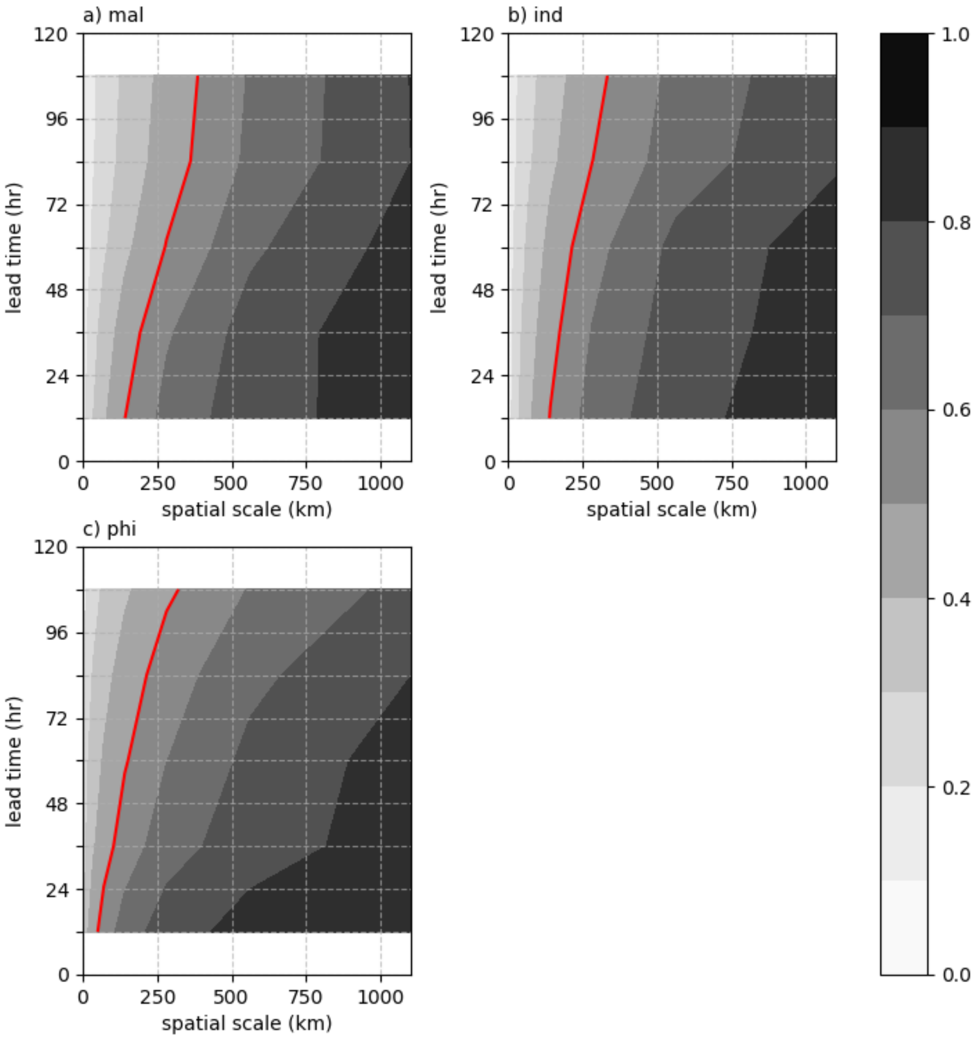

Figure 4: Fractions Skill Score (eFSS) of 24-hourly accumulated precipitation exceeding 95th percentile aggregated over all 4.5 km forecasts in Oct 2018-Mar 2019 as function of spatial scale (x-axis) and lead time (y-axis) for a) Malaysia model domain, b) Indonesia model domain, and c) Philippines model domain. The red line shows the eFSS=0.5 "skilful" contour.

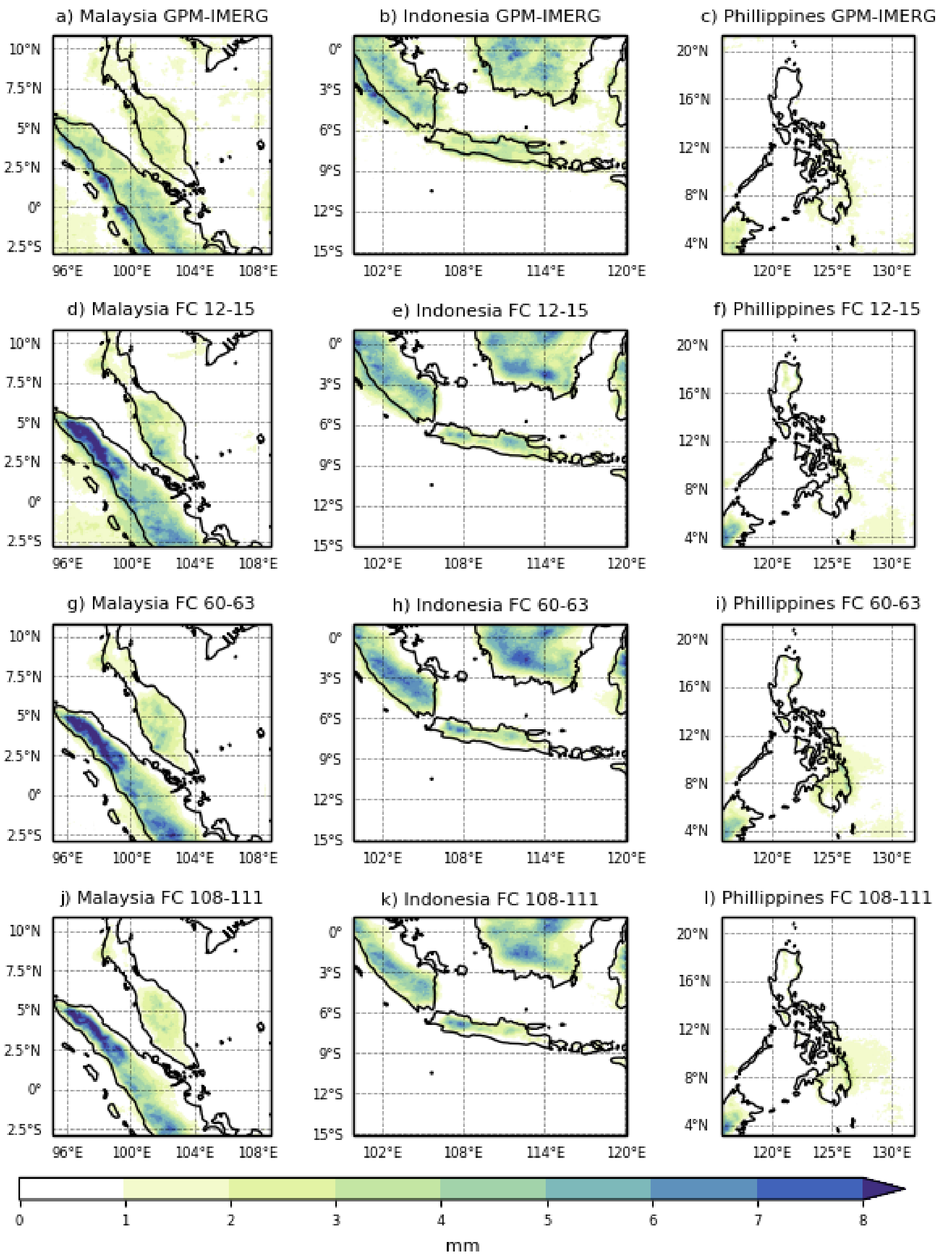

Figure 5: GPM-IMERG and 4.5 km forecast 95th percentile precipitation for Malaysia (left panels), Indonesia (centre panels) and Philippines (right panels). The 95th percentile is calculated for 3-hourly accumulations at 12-15UTC (local evening) and is displayed as mm hr^{-1}. Forecast precipitation is shown for lead day 1 (hours 12-15), lead day 3 (hours 60-63) and lead day 5 (hours 108-111).

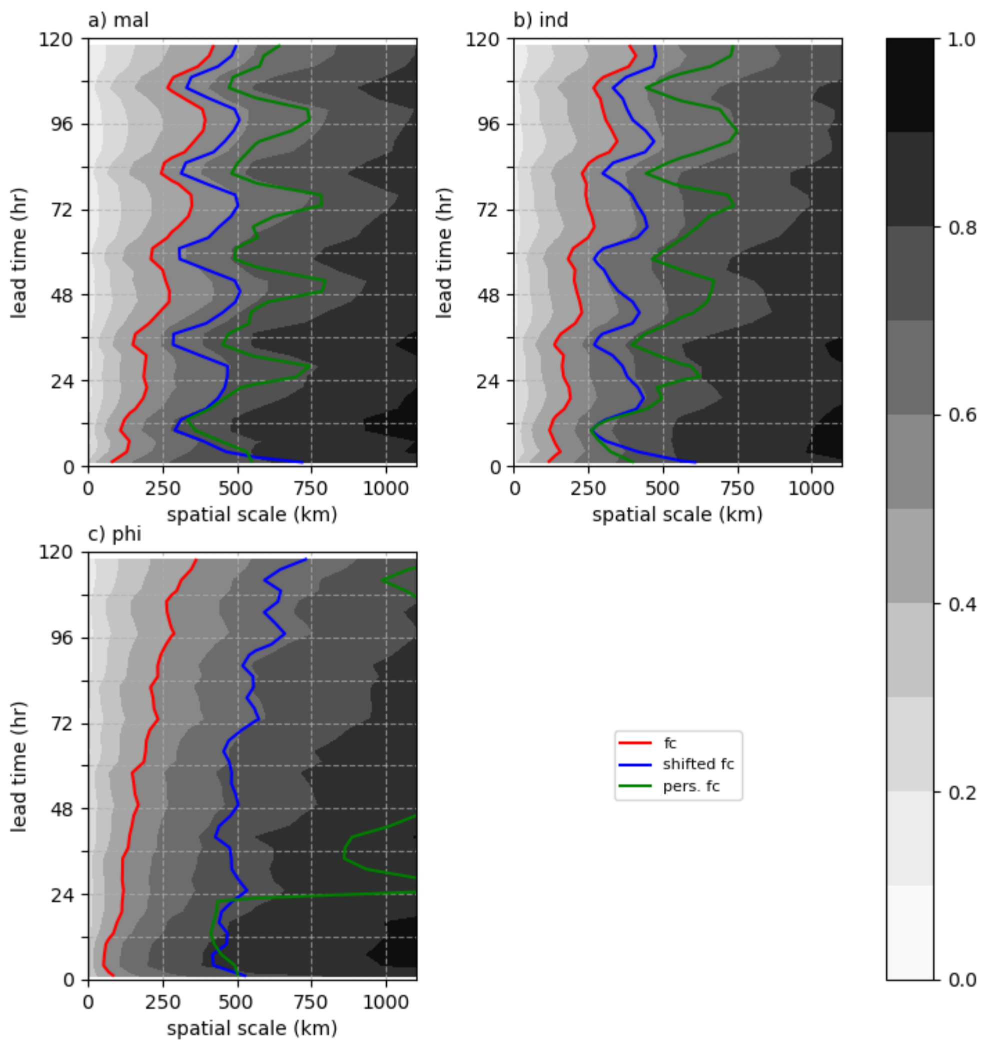

Figure 6: eFSS of 3-hourly accumulated precipitation exceeding 95th percentile aggregated over all 4.5 km forecasts in Oct 2018-Mar 2019 as a function of spatial scale (x-axis) and lead time (y-axis) for a) Malaysia model domain, b) Indonesia model domain, and c) Philippines model domain. The red line shows the eFSS=0.5 "skilful" contour for 3-hourly forecasts. The blue line shows the eFSS=0.5 contour for 3-hourly "shifted forecasts" where the forecast time-stamp has been shifted a day ahead of observations. The green line shows the FSS=0.5 contour for a persistence forecast such that lead time corresponds to the lag of the persistence forecast (see methods).

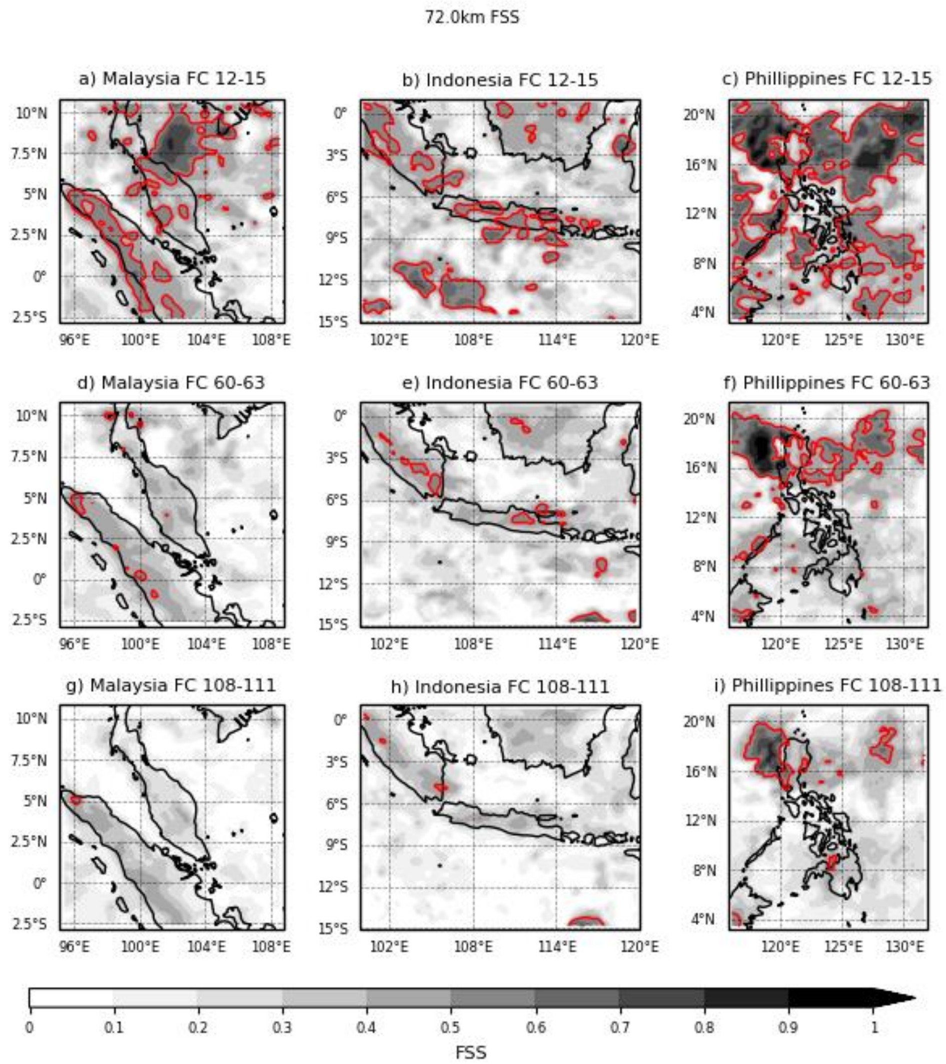

Figure 7: Localised FSS (LFSS) for the 72 km neighbourhood scale 3-hourly accumulated precipitation exceeding 95th percentile aggregated over all 4.5 km forecasts in Oct 2018-Mar 2019 for the Malaysia model domain (left panels), the Indonesia model domain (centre panels), and Philippines model domain (right panels). LFSS is calculated for lead hours 12-15, 60-63 and 108-111. The red line shows the LFSS=0.5 contour.

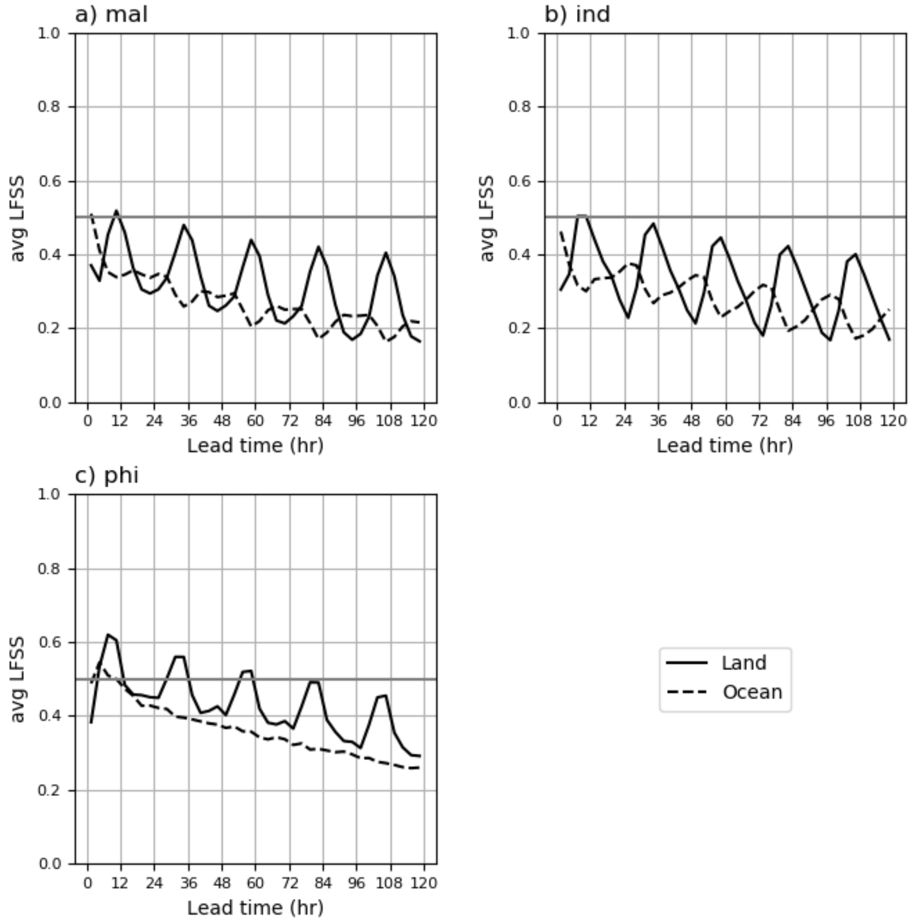

Figure 8: eFSS calculated using land and ocean points for the 72 km neighbourhood scale 3-hourly accumulated precipitation exceeding 95th percentile aggregated over all 4.5 km forecasts in Oct 2018-Mar 2019 for a) the Malaysia model domain, b) the Indonesia model domain, and c) Philippines model domain. Solid lines show eFSS over land and dashed lines show eFSS over ocean.

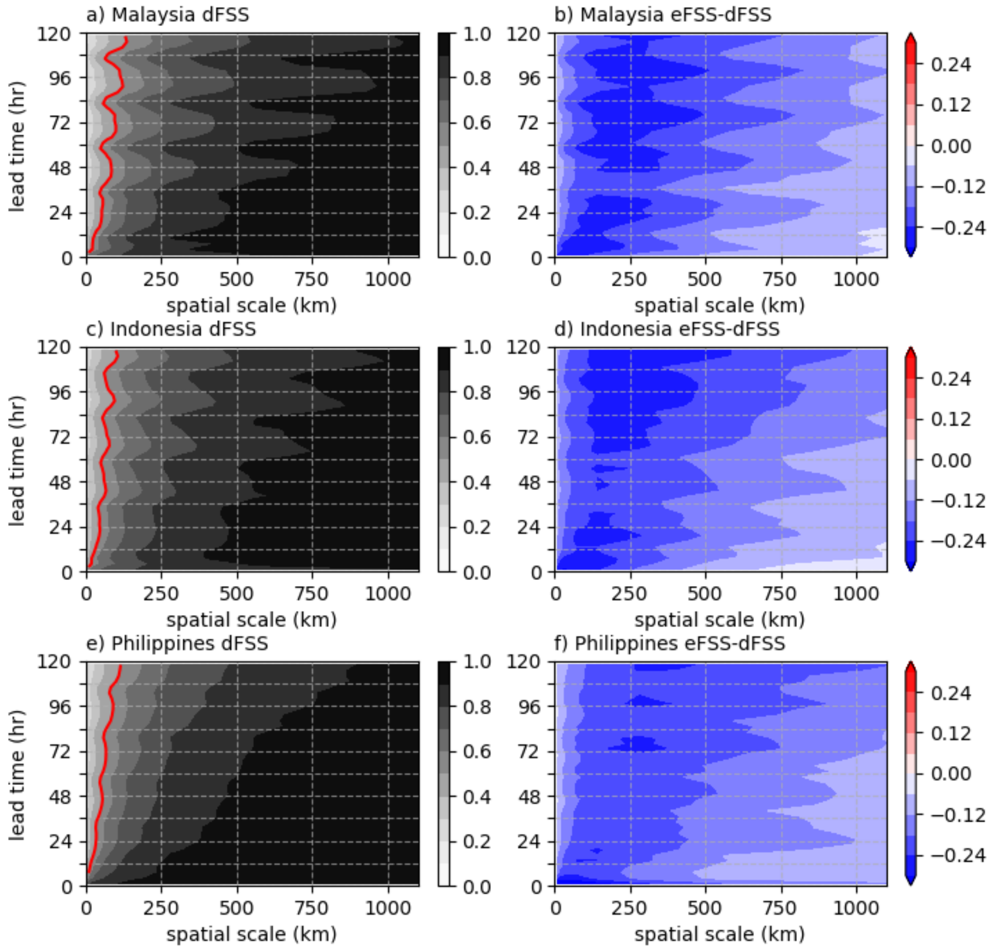

Figure 9: Dispersion Fraction Skill Score (dFSS) of 3-hourly accumulated precipitation exceeding 95th percentile aggregated over all 4.5 km forecasts in Oct 2018-Mar 2019 as function of spatial scale (x-axis) and lead time (y-axis) for a) Malaysia domain (Note - local time is UTC+8), c) Indonesia domain and e) Philippines domain. The red line shows the dFSS=0.5 contour for 3-hourly forecasts. The difference between eFSS and dFSS for the same forecasts as in a), c) and e) are shown in b), d) and f) such that a negative value indicates the ensemble is under-spread and a positive value indicates the ensemble is over-spread.

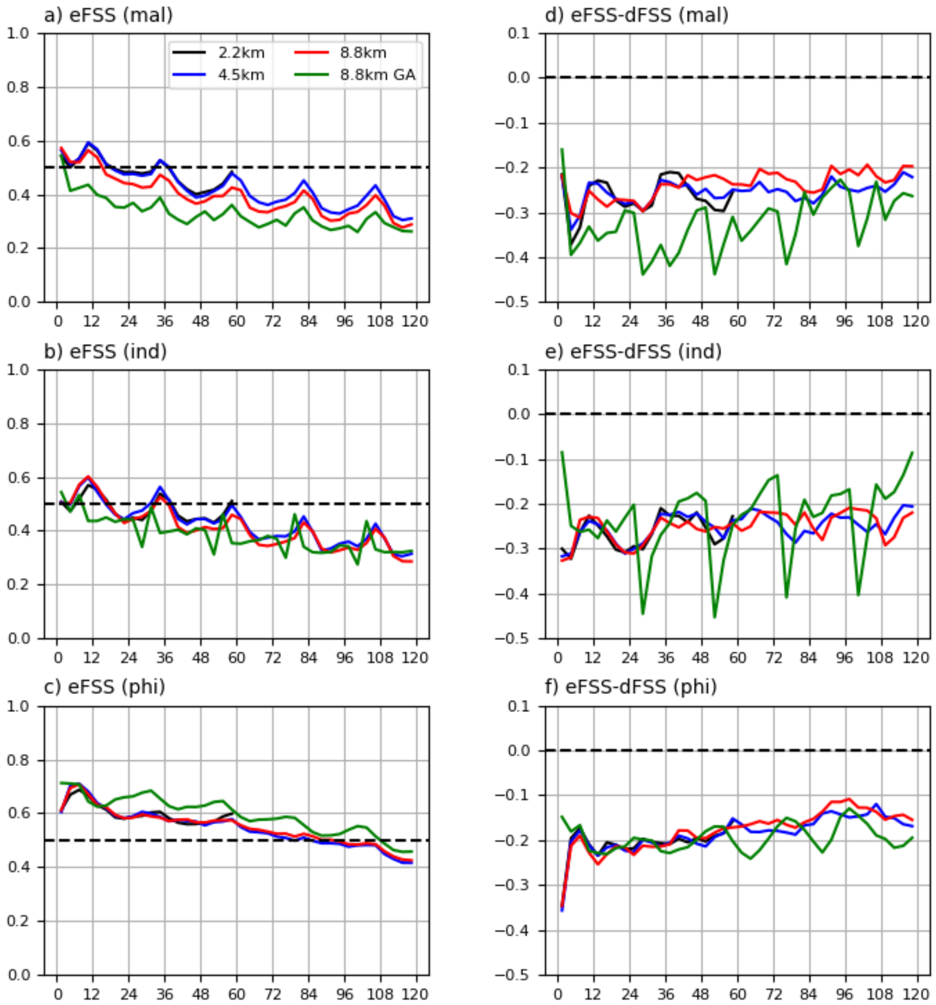

Figure 10: Fraction Skill Score (a-c) and the difference with the dispersion Fraction skill score (d-f) at a spatial scale of 144 km of 3-hourly accumulated precipitation exceeding 95th percentile aggregated over all forecasts in Oct 2018-Mar 2019 as a function of lead time (x-axis) for a) & d) Malaysia model domain, b) & e) Indonesia model domain, and c) & f) Philippines model domain. The color of the lines indicates the type of forecast; 2.2 km CP (black), 4.5 km CP (blue), 8.8 km CP (red), and 8.8 km GA (green). Horizontal dashed line in a)-c) indicates a skilful forecast (eFSS>0.5). In d)-f) negative values indicated an under-spread ensemble, positive

values indicate an over-spread ensemble. Analysis in d)-f) is performed over a subregion of the stated 2.2 km region (see figure discussion).