Weather patterns in Southeast Asia: enhancing high-impact weather sub-seasonal forecast skill

Article

Published Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1002/qj.4378

RESEARCH ARTICLE

Weather patterns in Southeast Asia: Enhancing high-impact weather subseasonal forecast skill

Paula L. M. Gonzalez[1,3] | **Emma Howard**[1,3] | **Samantha Ferrett**[1,2,3] | **Thomas H. A. Frame**[3] | **Oscar Martínez-Alvarado**[1,3] | **John Methven**[3] | **Steven J. Woolnough**[1,3]

[1]National Centre for Atmospheric Science, University of Reading, Reading, UK

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

[3]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Correspondence
Paula L. M. Gonzalez, NCAS/Department of Meteorology, University of Reading, Meteorology Building, Whiteknights Road, Earley Gate, Reading, RG6 6ET, UK.
Email: p.gonzalez@reading.ac.uk

Abstract

While skilful forecasts of heavy rainfall are highly desirable for weather warnings and mitigating impacts, forecasting such events is notoriously difficult, even with the most advanced numerical weather prediction models, due to the strong dependence on convective-scale processes. The large-scale circulation, on the other hand, is typically more predictable. Weather patterns (WPs) are a set of circulation types obtained statistically that can be used to characterize regional weather and harness the predictability of the large-scale circulation. In this work we produce pattern-conditioned probabilistic rainfall forecasts by projecting the horizontal winds from the Met Office GloSea5 prediction system on to WPs and then using the observed relationship between each WP and rainfall estimated by satellite. The WPs are derived following a novel two-tier clustering technique: the WPs in the first tier represent planetary-scale variability, such as El Niño–Southern Oscillation (ENSO), while the WPs in the second tier capture synoptic-scale variability. We investigate WP predictability as well as the improvement in skill of subseasonal rainfall forecasts gained by this technique. GloSea5 predicts the WP occurrence with skill extending beyond lead day 10. The pattern-conditioned rainfall forecasts were evaluated against climatological forecasts and model-simulated rainfall hindcasts. We show that the pattern-conditioned forecasts are skilful and outperform the model-simulated rainfall hindcasts for lead times extending to days 10–20, depending on the specific exceedance criteria and region. Spatial aggregation leads to increased levels of skill, but not to a significant extension of the skilful prediction horizon. These results constitute a fundamental step for the development of subseasonal prediction systems for Southeast Asia.

KEYWORDS

cluster analysis, extended-range forecasts, extreme rainfall, high-impact weather, predictability, S2S prediction, weather patterns

1 | INTRODUCTION

Reliable high-impact weather (HIW) forecasts facilitate planning, preparation, and timely deployment of emergency services (Vitart *et al.*, 2017; White *et al.*, 2017; Moron *et al.*, 2018). The ability to predict the risk of HIW events at forecast horizons beyond the first forecast week is critical for the reduction of their societal impact. Over Southeast Asia, the most impactful HIW events, in terms of economic and social insecurity, are floods and landslides (Gupta, 2010; Eckstein *et al.*, 2019). Such extreme events are associated with heavy precipitation due to tropical cyclones and convective weather systems. Precipitation, however, is very poorly captured by global numerical weather prediction systems at lead times beyond 1–2 days in the Tropics (Vogel *et al.*, 2020), although greater skill at lead times of 1–3 days can be obtained with convection-permitting ensemble forecasts (Ferrett *et al.*, 2021). In the case of Southeast Asia, the intricate land–sea contrast due to the many islands and the complexity of the mountainous terrain leads to very small spatial and temporal features in precipitation that most general circulation models cannot capture (Love *et al.*, 2011; Birch *et al.*, 2016; Ferrett *et al.*, 2020). By contrast, there are variables related to the large-scale atmospheric circulation that can be predicted more reliably than rainfall, such as geopotential height or horizontal wind components above the mixed layer (Boer, 2003; Hohenegger and Schar, 2007; Zhang *et al.*, 2007; Selz and Craig, 2015; Ying and Zhang, 2017), and this has led to the development of post-processing methods such as model output statistics (MOS) to relate precipitation to the more predictable components (Glahn and Lowry, 1972).

Over the last two decades, clustering and other statistical methodologies have been used to identify weather patterns (WPs) and characterise these as the more predictable features of the large-scale circulation and associated surface impacts (Huth *et al.*, 2008; Philipp *et al.*, 2010; Ghil *et al.*, 2019). Furthermore, these methods have been shown to be a promising technique for the enhancement of skill in the subseasonal-to-seasonal (S2S) prediction range (Mariotti *et al.*, 2020; Mastrantonas *et al.*, 2022). A limitation of the WP approach is that each pattern has a fixed spatial structure, while many of the relevant weather phenomena may be propagating features (e.g., equatorial waves, tropical cyclones). However, many features recur in similar geographical locations, associated with orography, land–sea contrast, and the more stationary elements of the larger-scale flow. In recent years, several authors have explored the subseasonal predictability of WPs over different regions (Vigaud *et al.*, 2018; Wang *et al.*, 2019; Cortesi *et al.*, 2021). Even though there have been studies applying WPs over Southeast Asia (Moron *et al.*, 2015;

Hassim and Timbal, 2019), to our knowledge their use for the prediction of precipitation into the subseasonal range has not been explored.

In a companion article, Howard *et al.* (2021). (referred to as H21 hereafter) introduced a novel technique for the identification of WPs over Southeast Asia based on clustering lower tropospheric horizontal wind data. Two approaches were explored and compared: a "tiered" approach using a planetary-scale tropical domain (20% of the Earth's surface) to define the first tier of clusters, followed by a second tier conditional on the occurrence of the first, clustering only in the Southeast Asia region (25% of the tier-1 area). The second "flat" approach used a single clustering step on the smaller regional domain only, imposing the same number of clusters as the second tier of the tiered approach. Their results show that the WPs obtained through both methodologies capture the variability and circulation patterns associated with the seasonal cycle over the region, and are therefore very likely to represent that aspect of precipitation variability. H21 also show that there is a strong connection between the tiered and flat WPs and extreme precipitation over the region, which is of greater relevance from an impact-based perspective. Furthermore, through the use of a perfect pattern forecast approach they demonstrated that the pattern-conditioned extreme rainfall predictions have the potential to outperform the climatology and also Madden–Julian Oscillation (MJO)-conditioned forecasts.

Here, the objective is to design and evaluate pattern-conditioned forecasts of extreme rainfall exceedance and exploit their potential to extend the range of skilful HIW predictions across Southeast Asia. In particular, the study addresses the following research questions.

- Can we derive pattern-conditioned predictions of extreme precipitation that are skilful on subseasonal-to-seasonal (S2S) timescales relative to climatological forecasts?
- Can they outperform the rainfall predictions simulated by global models and at which scales?
- Does one of the clustering methodologies ("tiered" or "flat") lead to more skilful pattern-conditioned predictions than the other?

The rest of the article is organised as follows: Section 2 describes the data and methodology followed to derive pattern-conditioned forecasts. Section 3 focuses on the ability of the UK Met Office GloSea5 extended-range prediction system to capture the regimes and their climatological features, as well as the system's skill for predicting pattern assignment. Section 4 introduces an implementation of the pattern-conditioned forecasts of

precipitation exceedance and an initial assessment of their skill compared with climatological forecast benchmarks as well as with the simulated precipitation. This section also explores the impact on skill of the spatial aggregation of the predictions. A high-impact weather case study illustrating the implementation of the methods is presented in Section 5. Finally, a discussion and conclusions regarding the results are included in Section 6.

2 | DATA AND METHODOLOGY

2.1 | Data

Weather patterns were identified through the application of k-means clustering to daily-mean 850-hPa zonal and meridional wind fields from the ERA5 reanalysis set (Hersbach *et al.,* 2020) over the period 1979–2018. The pattern assignment used to represent "truth" in evaluation of the predictions was based on the 0000 UTC ERA5 wind fields, corresponding to the available valid time of the forecast fields described below.

Daily precipitation estimates were obtained from the National Aeronautics and Space Administration (NASA) Global Precipitation Measurement (GPM) dataset (Huffman *et al.,* 2015) over the period June 2000–December 2018. The GPM Dual-frequency Precipitation Radar (DPR) and Global Microwave Imager (GMI) Combined Precipitation L3 daily version 06 product was aggregated and interpolated to a $1.5° \times 1.5°$ grid to match the forecast dataset. This dataset was used to calculate climatological probabilities of precipitation exceedance, as the truth for assessing forecast skill, and to condition them given the occurrence of the different weather patterns.

A UK Met Office GloSea5 hindcast set (GS5: MacLachlan *et al.,* 2015.) corresponding to operational forecast starts in the period between July 2017 and June 2018 was obtained from the World Meteorological Organization (WMO) S2S dataset (Vitart *et al.,* 2017) through the European Centre for Medium-Range Weather Forecasts (ECMWF) archive at the common 1.5° horizontal resolution.[1] This set comprises four start times per month (on days 1, 9, 17, and 25 of each month), 23 years of hindcasts for each start time (1993–2015), and seven ensemble members. Fields corresponding to 850-hPa zonal and meridional winds at 0000 UTC and daily total precipitation accumulation (24 hours up to each valid time) were considered. This hindcast set includes forecasts for days 0–60, but only the first 36 forecast steps (35 days) were used in this study.

2.2 | Weather-pattern definitions

A full description of the methodology to define WPs can be found in H21, but a brief description is included here for completeness. In H21, two approaches were explored and compared. The first, called the two-tiered approach, made use of a planetary-scale domain (35°S–35°N, 60°E–180°E) to define the first tier of clusters (eight "tier-1 regimes"). These clusters were able to isolate large-scale climatological features such as the monsoon circulation and its seasonal cycle and the influence of El Niño–Southern Oscillation (ENSO). A second cluster analysis was performed over a smaller Southeast Asia region (15°S–25°N, 90°E–140°E) using only data from the days assigned to one tier-1 regime to identify between five and eight "tier 2" weather patterns. The result is 51 "tier-2" weather patterns, each linked to only one "tier-1" regime. These will be referred to as the "tiered set" of WPs. The smaller-scale tier-2 WPs capture the synoptic variability affecting the region and the connection with tropical phenomena such as the MJO, Boreal Summer Intraseasonal Oscillation (BSISO), cold surges, tropical cyclones, and equatorial waves. The second "flat" approach considered a single clustering step on the smaller regional domain only, imposing the same number of 51 clusters as in the tiered approach, for ease of comparison.

2.3 | Weather-pattern assignment

No new pattern definitions were implemented for the GS5 hindcasts. Instead, for each hindcast date and ensemble member, an assignment of the atmospheric state to a pattern in the flat, tier-1 and tier-2 categories is done by finding the minimum Euclidean distance between the 850-hPa 0000 UTC zonal and meridional simulated fields and the patterns or "centroids" obtained from the clustering analysis of ERA5 data, as calculated in H21 (Howard *et al.,* 2021). The full expression of the Euclidean distance to be minimised is

$$ED(r) = \frac{1}{N_x \times N_y} \sum_{x,y} \left[(u_f(x,y) - u_c(r,x,y))^2 \right. $$
$$\left. + (v_f(x,y) - v_c(r,x,y))^2 \right], \quad (1)$$

where u_f and v_f represent the simulated zonal and meridional wind fields, and $u_c(r)$ and $v_c(r)$ represent the zonal and meridional components of the ERA5 pattern centroid r. For each forecast, the pattern r is assigned as the one that minimises $ED(r)$. This assignment was completed for forecast leads 0–35 days. It is important to point out that no bias correction was performed on the GS5 forecasts prior to the WP assignment.

2.4 | Pattern-conditioned forecasts of precipitation threshold exceedance

Based on the hypothesis that an ensemble global prediction system is able to forecast the synoptic and larger-scale circulation described by the WPs better than precipitation, a skilful prediction of extreme rainfall risk might be obtained from combining the predictions of pattern membership (in this case, from the GS5 hindcast set) with the pattern-conditioned climatological probabilities of precipitation exceedance derived from observations (in this case, the combination of ERA5 patterns and GPM precipitation). The diagram presented in Figure 1 describes the methodology, which was applied to both the "tiered" and "flat" WP sets, and to different criteria for extreme precipitation.

Results here present an initial implementation of the pattern-conditioned hindcasts of precipitation exceedance. Two exceedance criteria were considered, namely a fixed threshold on the grid-cell precipitation ($25 \, \text{mm·day}^{-1}$) and a percentile definition (90th percentile, P90 hereafter). The calculation of P90, which is consistent with the one used in H21, considers a climatology built using a running 60-day window around each day of the year and for each grid cell, regardless of whether it was rainy or not. This climatology presents a seasonal cycle. While the conclusions we present are qualitatively valid for both criteria, generally we found lower levels of skill with the $25 \, \text{mm·day}^{-1}$ threshold than using P90. This can be explained, at least partially, by considering that, even though the comparison between the two exceedance criteria changes with the region and season, P90 is more frequently exceeded than $25 \, \text{mm·day}^{-1}$ over Southeast Asia (see Figure S2 in the Supporting Information), i.e. the $25 \, \text{mm·day}^{-1}$ threshold represents a more extreme case than P90. In this work we mainly present results for the $25 \, \text{mm·day}^{-1}$ threshold, while the P90 results are included in the Supplementary Information. In the case of the model-simulated precipitation probability of extreme rainfall, the exceedance was evaluated with respect to both GPM precipitation and the hindcast-based lead-dependent climatology.

3 | WEATHER PATTERNS IN THE GS5 HINDCASTS

The analysis of the WPs obtained from GS5, including average frequencies, seasonality, and persistence, reveals that the GS5 system has a reasonable representation of both the tiered and flat patterns, albeit with some biases and drifts. In terms of the typical persistence, Table 1 presents some key statistics for GS5 patterns and how they compare with pattern occurrence in ERA5. The most significant difference is observed in the persistence of tier-1 regimes, which tend to last longer in GS5 than in ERA5. This is observed in spite of two facts: (1) persistence is limited by the maximum GS5 hindcast length (60 days) and (2) regime events might have started before the beginning of the forecast, so that shorter durations than in ERA5 would be expected. In terms of the shorter-lived WPs, GS5 slightly overestimates the durations of tier-2 WPs while the opposite is true for the flat WPs, though these differences are not significant. To investigate the causes of the increased duration of tier-1 events in the GS5 hindcasts, we can disaggregate the statistics further by the tier-1 regime (Table 2). Quite noticeably, the overestimation seems to be explained mainly by regime 5 (R5), the eastward extension of the Asian summer monsoon, for which persistence is significantly overestimated. For most other regimes, however, the hindcasts reproduce events that are on average shorter than in ERA5, though the differences are statistically

Pattern-conditioned forecasts of extreme precipitation over SE Asia

Given a pattern definition (flat or tiered) and exceedance threshold (e.g., 25mm/day):
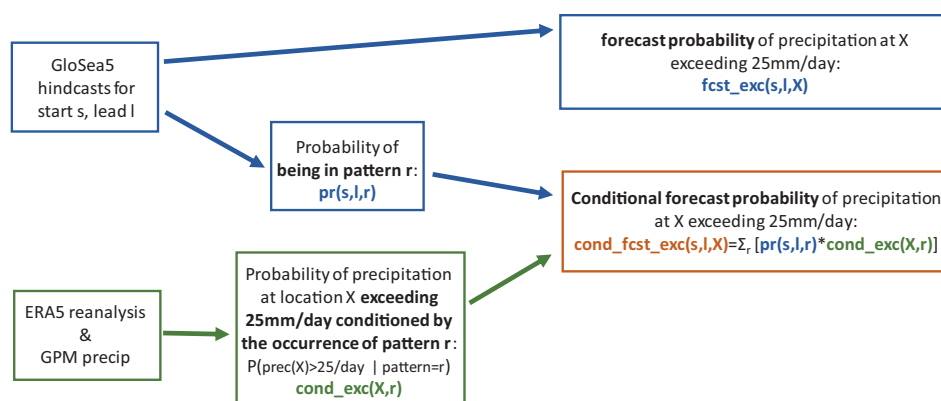
FIGURE 1 Schematic describing the pathways to generate a direct precipitation simulation and pattern-conditioned forecast of extreme precipitation risk. Boxes indicated in blue indicate information derived from the GS5 forecasts. Boxes in green indicate variables derived from observations and reanalysis. Finally, the regime-conditioned forecasts are described in the orange box [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Statistics for the typical duration of weather pattern events in days. (a) ERA5 0000 UTC patterns for 1979–2016. Values between parentheses reflect results from considering only events shorter than 61 days. (b) GS5 hindcast patterns considering the period 1993–2015, 7 ensemble members, and lead times up to day 60. P90 and P99 denote the 90th and 99th percentiles. *Red and **blue indicate hindcast values higher and lower than those in ERA5, respectively. Bold italic values indicate that the means are statistically different with respect to a two-tailed t-test with $p < 0.001$ [Colour table can be viewed in online version at wileyonlinelibrary.com]

(a)	Mean	Median	P90	P99
Tier 1	6.8 (6.6)	3.0	18.0	45.0 (42.7)
Tier 2	2.4	2.0	5.0	11.0
Flat	2.4	2.0	5.0	10.0
(b)	Mean	Median	P90	P99
Tier 1	*8.7**	*4.0**	*25.0**	*59.0**
Tier 2	*2.5**	2.0	5.0	13.0
Flat	*2.3***	2.0	*4.0***	10.0

TABLE 2 Statistics of the average duration of tier-1 regimes individually in days. (a) ERA5 0000 UTC regimes for 1979–2016. Values between parentheses reflect results from considering only events shorter than 61 days (five events). (b) UK Met Office GloSea5 hindcast regimes considering the period 1993–2015, seven ensemble members, and lead times up to day 60. P90 and P99 denote the 90th and 99th percentiles. *Red and **blue indicate hindcast values higher and lower than those in ERA5, respectively. Bold italic values indicate that the means are statistically different with respect to a two-tailed t-test with $p < 0.001$ [Colour table can be viewed in online version at wileyonlinelibrary.com]

(a)	Mean	Median	P90	P99
R1	5.7	3.0	14.7	36.1
R2	5.4 (5.1)	2.5 (2.0)	12.0	41.7 (40.2)
R3	8.3 (8.1)	3.0	24 (23.4)	50.3 (49.3)
R4	10.6 (9.7)	5.0 (4.0)	26.5 (25.0)	60.3 (49.6)
R5	7.6	4.0	20.0	32.5
R6	6.0	4.0	13.6	28.9
R7	7.4	4.0	21.0	37.9
R8	4.4	2.0	12.0	22.4
(b)	Mean	Median	P90	P99
R1	5.0***	3.0	*12.0***	*30.0***
R2	4.5***	3.0*	*10.0***	*28.0***
R3	7.3***	3.0	*21.0***	*43.0***
R4	*5.3***	3.0***	*12.0***	*29.0***
R5	*17.0**	*8.0**	*49.0**	*61.0**
R6	*4.1***	*2.0***	*9.0***	29.0*
R7	6.4***	3.0***	*16.0***	*31.0***
R8	3.9***	2.0	*10.0***	23.1*

significant mainly for R4 and R6. Another significant difference is that tier-1 R3 and R4 are actually the ones that tend to be more persistent in ERA5, rather than R5.

The WP frequency and seasonality need to be considered as a function of hindcast lead time. Figure 2 presents the annual mean frequencies of the tiered (Figure 2a) and flat patterns (Figure 2b). In each case, the values obtained from the ERA5 pattern assignment (at 0000 UTC) are included as the leftmost column. In both WP sets, the proportion of days assigned to each tiered and flat WP at early lead times matches ERA5 very closely. However, as lead time increases some model drifts become evident. The most significant one is a drift to a higher frequency in R5 occurrence (purple tones), which grows mainly at the expense of R4 occurrence (red tones). Both R4 and R5 represent the Asian summer monsoon flow. From results in H21, this suggests a bias of GS5 towards intense boreal summer monsoon precipitation conditions with an eastward extension of the westerly winds into the South China Sea. This model drift could explain the overestimation of the tier-1 R5 persistence discussed earlier. A more subtle drift is also observed for regime R1 (blue tones), the frequency of which increases at the expense of R2 (orange tones) and R3 (green tones). This suggests that GS5 has a bias towards an extension to the east of the austral monsoon precipitation over Indonesia (see H21, Figure 3). The idea that these sets of regimes are linked in their frequencies (R1–R2/R3 and R4–R5) is based on the statistics of regime transitions and their strong seasonality as described in H21.

A similar behaviour is also seen in the flat WPs, as a gradual increase in the WPs numbered around 30 and a decrease in the ones numbered between 23 and 27. This is consistent with the strong co-occurrence between many of the flat and tiered WPs, as shown in H21. A trend in the frequency of the group of flat WPs associated with the austral monsoon is not noticeable. However, an increase in the frequency of the flat WP4 is clear. This pattern is associated with austral summer monsoon conditions with intense westerly winds over Indonesia (see H21, Figure S2).

H21 showed that the planetary-scale tier-1 regimes capture the different phases of the seasonal cycle of the region, but tier-2 WP frequencies are also seasonally conditioned, as are those of the flat WPs. Figure 3 describes the seasonality of the tiered WPs in GS5 for a set of different lead times and compares it with ERA5 (Figure 3a). Since the frequencies corresponding to GS5 are more sparse due to the system having four starts a month, they
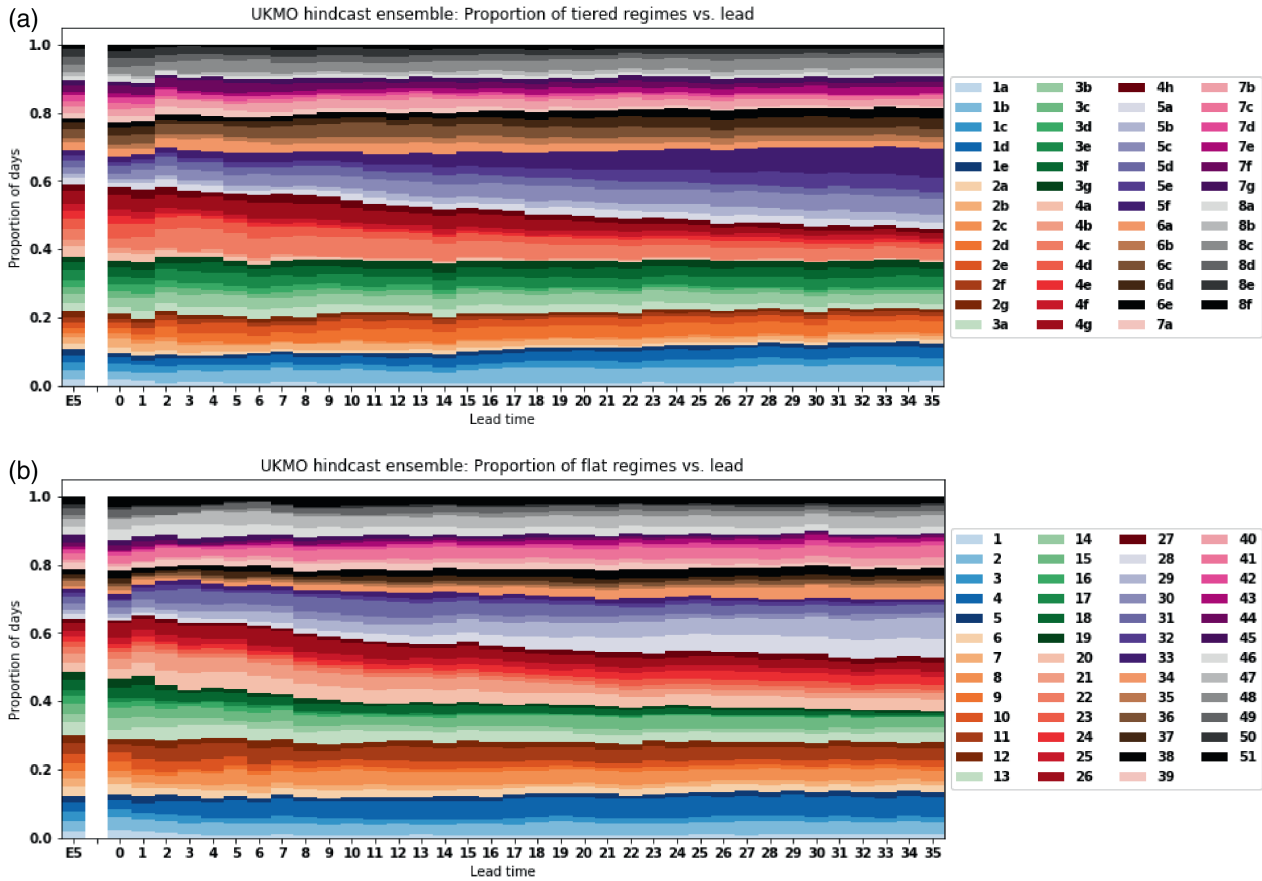
FIGURE 2 Annual normalised frequency of tiered and flat patterns as a function of lead time. The leftmost column presents the frequencies for ERA5 patterns (assigned using data at 0000 UTC). (a) Tiered patterns; (b) flat patterns [Colour figure can be viewed at wileyonlinelibrary.com]

have been aggregated by month of the year for clarity. The match between GS5 and ERA5 is very good for lead time 0 and good for day 7 (Figure 3b,c), but from day 14 (Figure 3d) a shift in the frequencies from R4 to R5 in the months between July and September is clear. Another tendency towards more R1 and fewer R2 patterns can be seen in March and April. These biases are consistent with the drifts observed in the annual mean frequencies (Figure 2). Consistently, the flat WPs show drifts in WPs 0–10 at the beginning of the calendar year, and in WPs 20–30 from May to September (Figure S1). This is not surprising, given the high-degree of co-occurrence between the flat and tiered WPs (H21). Nonetheless, this analysis shows that it is easier to interpret the model biases in terms of shifts in the seasonal properties of the monsoon, using the tiered WPs.

3.1 | Skill of weather-pattern assignment

To assess the skill of GS5 to predict the observed patterns, we consider the multicategory Brier score (BS: Brier, 1950.)

and its associated skill score (BSS). These can be defined as

$$BS = \frac{1}{F} \sum_t \sum_k \{p(k,t) - o(k,t)\}^2, \tag{2}$$

$$BSS = 1 - \frac{BS}{BS_{ref}}, \tag{3}$$

where p represents the forecast probability of occurrence of WP k at time t and o is a binary indicator of the observed occurrence of this pattern (yes/no). To build the skill score, the BS of a given forecast is compared with a reference forecast, ref, most commonly the climatological probability of occurrence of pattern k on that day of the year.

To address the bias induced by the limited size of the ensemble considered here, a correction was introduced to the BSS as suggested by Weigel *et al.* (2007). Since the correction is applied to the climatological reference, the bias-adjusted discrete BSS can be described as

$$DBSS = 1 - \frac{BS}{DBS_{clim}}, \tag{4}$$
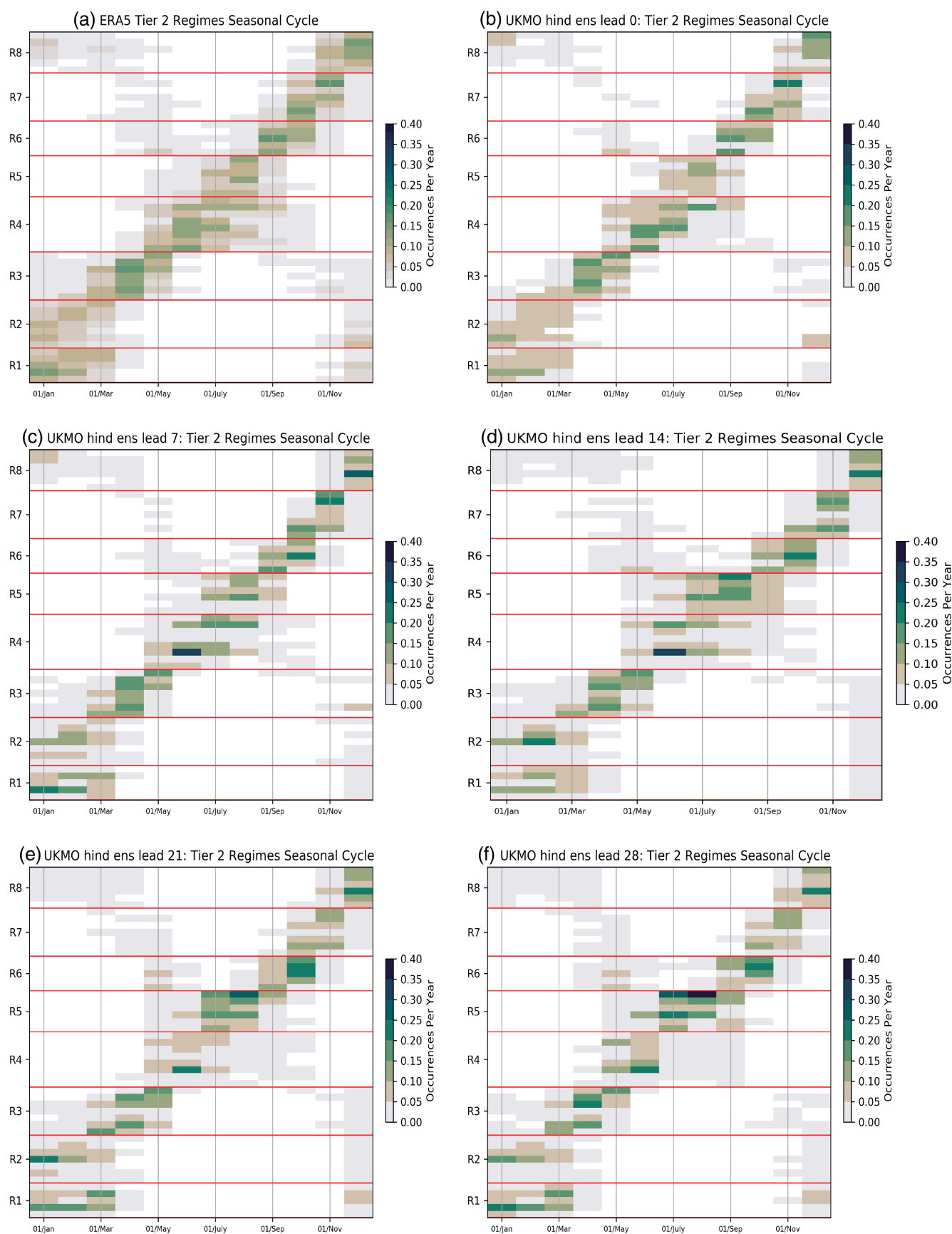
FIGURE 3   Frequency of each tiered regime as a function of the month. Units are occurrences per year. (a) ERA-Interim 0000 UTC;
(b) Lead 0; (c) Lead 7; (d) Lead 14; (e) Lead 21; (f) Lead 28 [Colour figure can be viewed at wileyonlinelibrary.com]
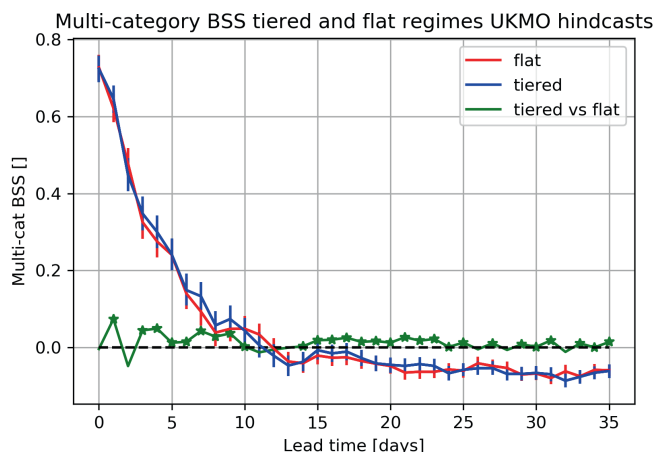
FIGURE 4 Discrete Brier skill score (DBSS) of the tiered and flat pattern assignments (blue and red, respectively) and of the method comparison (green) as a function of lead time. Vertical bars indicate 5–95% confidence intervals obtained from a 1,500-sample bootstrapping analysis. The stars in the green line indicate the lead times for which the tiered BSS is significantly greater than the flat-regime BSS. This means that it is higher than the 95th percentile of a 1,500-sample random BSS obtained from reshuffling of the forecast and observed probabilities [Colour figure can be viewed at wileyonlinelibrary.com]

where the corrected climatological discrete Brier score, DBS_{clim}, can be expanded as

$$\text{DBS}_{\text{clim}} = \frac{1}{F}\sum_t \sum_k \left[\{p_{\text{clim}}(k,t) - \text{obs}(k,t)\}^2 + D(k,t) \right];$$

$$D(k,t) = \frac{1}{M}p_{\text{clim}}(k,t)\{1 - p_{\text{clim}}(k,t)\}, \qquad (5)$$

where F represents the total number of forecasts, p_{clim} is the climatological probability of occurrence of pattern k (derived from ERA5) and also depends on the day of the year, $\text{obs}(k,t)$ is a binary indicator of the observed occurrence of pattern k at date t, and M represents the size of the hindcast ensemble. Time t refers to the valid time of the forecast, which is determined jointly by the forecast initial time and its lead time. The uncertainty of the BSS values was estimated through a 1,500-sample bootstrapping analysis, following the guidelines in Davidson and MacKinnon (2000).

The results are summarised in Figure 4. The BSS values show that GS5 is significantly more skilful than the climatology to identify WP assignment for lead times up to day 10 in the case of the tiered patterns and day 11 in the flat case. However, the objective comparison between the methods, where the flat assignment was considered as a reference for the BSS calculation (green line), reveals that the skill levels are very similar for the assignment of both pattern sets, with the tiered method slightly outperforming
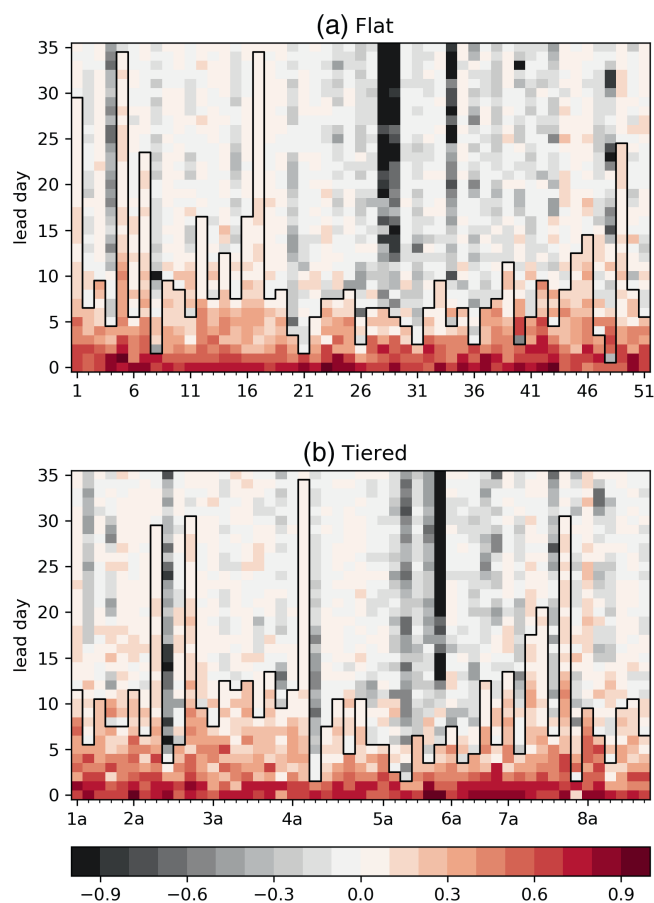
FIGURE 5 Discrete BSS for the assignment of individual patterns in (a) flat and (b) tiered sets. In each panel, the *x*-axis indicates the pattern and the *y*-axis indicates the lead time in days. The colour scale accounts for the BSS values and the thick black line is included to point out the sign change [Colour figure can be viewed at wileyonlinelibrary.com]

for most lead times. This difference in skill is statistically significant for most lead times according to a bootstrapping analysis of the tiered versus flat BSS (green line).

This skill assessment can be investigated further by analysing the performance to assign individual WPs. In this case, a binary discrete BSS score was applied considering two categories: pattern k versus all other patterns (Weigel *et al.,* 2007). These results are presented in Figure 5 and show that, with the exception of a few flat and tiered WPs, in most cases a BSS of zero is seen around lead day 10, in agreement with the multicategory assessment. The very negative BSS values seen in flat WPs in the 26–31 range and tier-2 WPs corresponding to tier-1 R5 can be related directly to the observed drift in WP modelled frequencies (see Figure 2). Other low-skill examples such as flat WP4 (Figure 5a) and tiered WP2d (Figure 5b) can be linked to biases in their modelled frequencies. The cases where the BSS changed sign at longer lead times were explored further as well, despite the fact that the skill levels remain
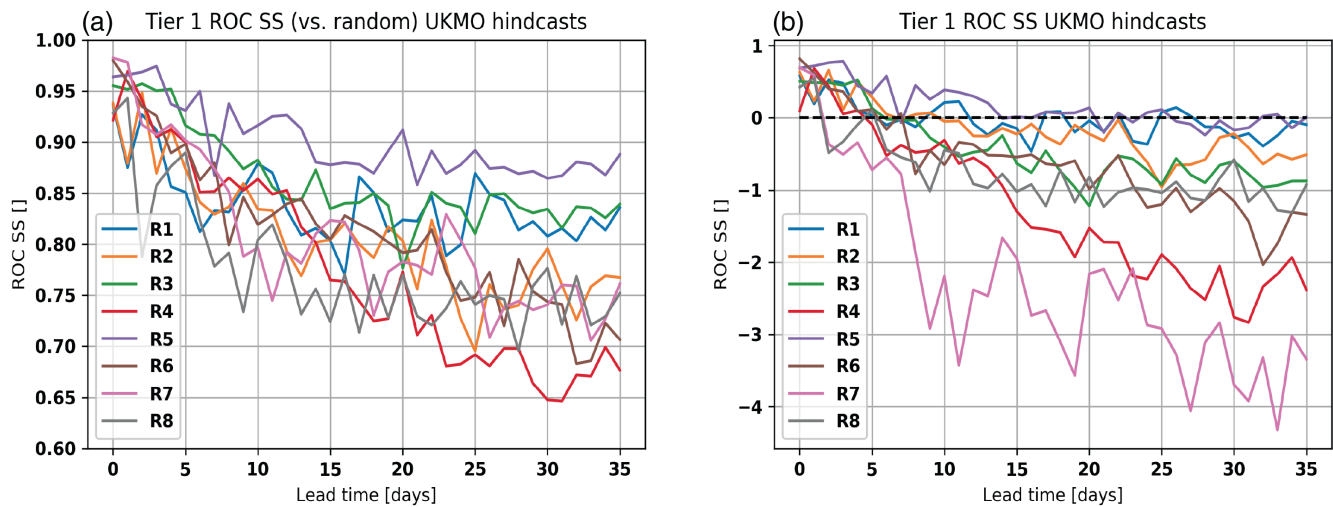
FIGURE 6   ROC skill score of the individual tier-1 regime assignments as a function of lead time. In (a) the skill score was calculated using a random assignment as a reference; in (b) the assignment according to the climatological probabilities of occurrence for each day of the year was considered as a reference [Colour figure can be viewed at wileyonlinelibrary.com]

quite low. Considering the results from H21, a connection was explored between the longer predictive skill for such patterns and their link with large-scale variability such as ENSO and the MJO, which might explain these particular patterns being more persistent. However, no evident connection was observed from this patternwise assessment.

We now assess the skill of GS5 to capture each of the tier-1 regimes, given that these contain information about large-scale quasistationary processes, which might result in predictive skill extending beyond the synoptic scale (H21). Figure 6 presents the relative operating characteristics (ROC, Mason, 1982.) skill score. The ROC score is derived from the area under the curve defined by the hit rates and false alarm rates varying across the different probabilities of occurrence of each event or, specifically in this case, of each tier-1 regime. A typical skill reference for the design of a ROC skill score is the use of a random assignment to normalize the area under the curve (as described in Wilks, (2019), section 9.4.6, equation 9.49). In this case, this corresponds to an equal probability of occurrence of each regime (Figure 6a). The fact that the individual ROC curves present significant jumps with lead time is a reflection of the limited sample size, and should therefore be considered a qualitative assessment of the single-regime assessment skill. These values are very high and suggest positive skill for the whole forecast period compared with a random assignment. This reflects the strong seasonal dependence present in the patterns but absent from the random pattern distribution. Therefore, a more significant reference for the assignment is the use of the climatological probabilities of occurrence of the patterns as a function of the day of the year derived from ERA5, as in the BSS analysis above (Figure 6b). These values show that most tier-1

regimes cannot be predicted skilfully beyond lead day 7, with the exception of R1, R2, and R5, which show positive skill-score values for lead times into the second hindcast week. Tier-1 R1 and R2 have been linked by H21 to the activity of ENSO, and it is therefore expected that they carry longer predictability. Tier-1 R5 has been linked to an eastward extension of the boreal summer monsoon, and it was shown above that the GS5 system has a drift towards high probabilities of occurrence of this pattern (Figure 2a). This might suggest that this extended prediction skill is not physical and reflects the model drift.The difference between the panels reveals that sole knowledge of the day of the year includes significant predictive information about the likelihood of occurrence of certain WPs.

4 | PATTERN-CONDITIONED PREDICTIONS OF HEAVY RAINFALL

The ultimate goal in the identification of WPs for Southeast Asia within this work was to demonstrate their potential use for the extended-range prediction of HIW, and in particular extreme rainfall. H21 showed that there is a strong connection between the tiered and flat WP sets considered here and extreme precipitation over the region. Furthermore, the use of a perfect WP forecast approach by H21 suggested that pattern-conditioned extreme rainfall predictions have the potential to outperform climatology and also MJO-conditioned forecasts over most months. Results presented in Section 2.3 indicate that GS5 is able to produce skilful probabilistic predictions of pattern membership for both methods and lead times extending into the second hindcast week. In this context, the next step

was to design and evaluate pattern-conditioned forecasts of extreme rainfall exceedance based on the GS5 hindcasts. This section presents an initial evaluation of such methods that serves as a demonstration of their potential use for the prediction of extreme precipitation over Southeast Asia.

4.1 | Skill of the predictions

To evaluate the skill of the pattern-conditioned rainfall exceedance hindcasts we used the BS (Brier, 1950) and its associated skill score (BSS), which can be described as

$$\mathrm{BS}_R(X) = \frac{1}{N_f} \sum_t [O(X, t) - \widetilde{PE}(X, t)]^2,$$

$$\mathrm{BSS}_R(X) = 1 - \frac{\mathrm{BS}_R}{\mathrm{BS}_{ref}}, \qquad (6)$$

where R is the regime set, $\widetilde{PE}$ is the predicted probability of exceedance for location X and time t, O is the binary indicator of observed exceedance, and N_f is the total number of predictions. The ($\sim$) operator indicates that the probability has been corrected to eliminate the specific time t from the relevant climatologies, to result in a cross-validated BSS that avoids skill inflation due to considering the verifying events in the climatological reference (a full expression can be found in equation 3 of H21). As in Equation 5, the time t refers to the valid time of the forecast. The correction to the BSS to account for the limited sample size was not implemented here, because it is unnecessary when the forecasting systems are benchmarked against each other rather than the climatology (Weigel *et al.,* 2007).

As a first instance, we evaluated the skill of the hindcasts at the daily scale and for the full domain. The different panels in Figure 7 present the BSS for 25 mm·day^{-1} exceedance corresponding to different hindcast lead times. Within each panel, the top row of subpanels presents the skill of the different methods relative to a forecast based on the GPM climatology that is a function of the day of the year:

(i) simulated precipitation,
(ii) flat pattern-conditioned predictions,
(iii) tiered pattern-conditioned predictions.

The bottom row of subpanels compares the skill between methods, for

(iv) flat versus simulated precipitation,
(v) tiered versus simulated precipitation,
(vi) tiered versus flat pattern-conditioned predictions.

In all the maps, positive BSS values (blue colours) indicate that the corresponding prediction is more skilful than the GPM-based climatological reference in a cross-validated sense. Additionally, points where this BSS is statistically significant are identified with dots. This means that the value is outside the 5–95% confidence interval of a random BSS, obtained from a 1,500-sample reshuffling of the forecast and observed probabilities. At lead day 0 (representing precipitation accumulated over the first 24 hours), the model-derived simulated precipitation hindcasts (Figure 7a(i)) show large areas of positive skill. Nonetheless, most areas of positive skill (blue) are not over land, with the largest region being connected with tropical cyclone (TC) activity to the east of the Philippines (H21). Moving to the pattern-conditioned predictions, the flat (Figure 7a(ii)) and tiered (Figure 7a(iii)) sets result in very similar skill patterns, with widespread positive values, though smaller in magnitude than for the hindcast simulated precipitation. Despite the small values, these methods lead to significant skill over more points over land (see Borneo, Sumatra, and the Philippines). The evolution with lead time reveals that the simulated precipitation exceedance hindcasts lose skill quickly, with minimal regions of positive skill by lead 7. While the positive BSS values observed for the pattern-conditioned hindcasts are very small, they are still positive and statistically significant over a significant portion of the land points at lead 7 (Figure 7b) and a reduced region at lead day 21 (Figure 7c), showing extended skill with respect to the simulated precipitation exceedance. This advantage is made clearer by the comparisons presented in subpanels (iv)–(vi), in which BSS calculations for each method are benchmarked against the others. These figures show that, at the initial lead time, simulated precipitation exceedance outperforms the regime-based methods mainly over some limited ocean areas (red colours in Figure 7a(iv) and Figure 7a(v)). These areas, however, have mostly disappeared by lead day 7 (Figure 7b), indicating that the pattern-conditioned hindcasts outperform the simulated precipitation exceedance significantly over most of the domain for lead times equal to or longer than 7 days. It is important to note that the fact that the BSS of this comparison is positive does not imply that the pattern-conditioned hindcasts remain skilful at longer lead times. The comparison between regime-based methods (subpanel vi) shows that, for longer lead times, the tiered pattern-conditioned hindcasts outperform the flat-based method over larger portions of the domain. The BSS values are only marginally positive, but statistically different from zero. The low BSS values are at least partially explained by the lack of sharpness in the pattern forecasts (extreme events are rarely forecast at high probabilities). It is also important to remark that it has been shown that negative BSS values that consider the
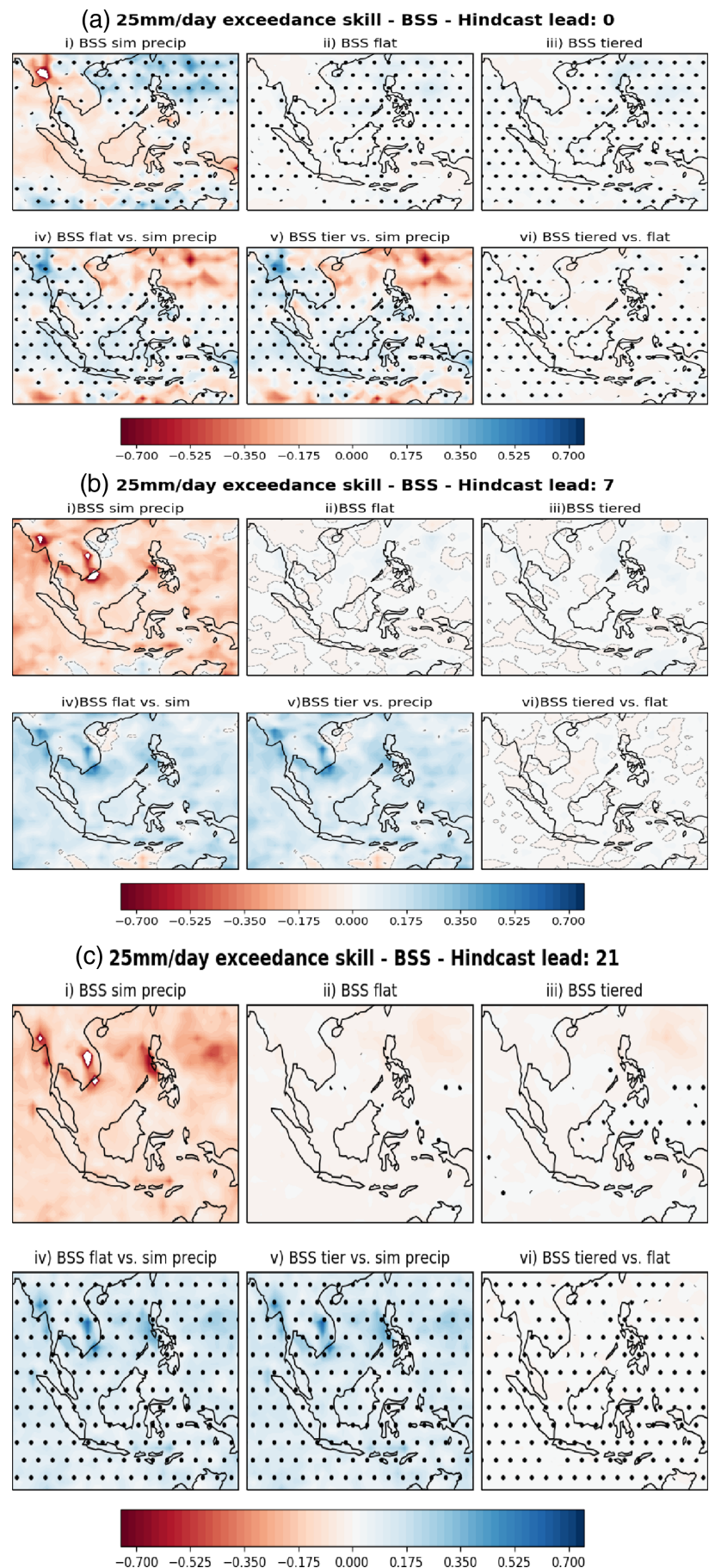
FIGURE 7 Daily BSS skill assessment of the hindcasts for 25 mm·day^{-1} exceedance. (a) Lead 0; (b) Lead 7; (c) Lead 21. In each plot, the first row of panels present the skill assessment of (i) simulated precipitation, (ii) flat pattern-conditioned precipitation, and (iii) tiered pattern-conditioned hindcast, all against the GPM seasonally varying reference. In the second row, the skill of different methods is compared for (iv) flat versus simulated precipitation, (v) tiered versus simulated precipitation, and (vi) tiered versus flat. The black dots indicate grid points where the BSS is statistically higher than 0, as described in Section 4.1 [Colour figure can be viewed at wileyonlinelibrary.com]
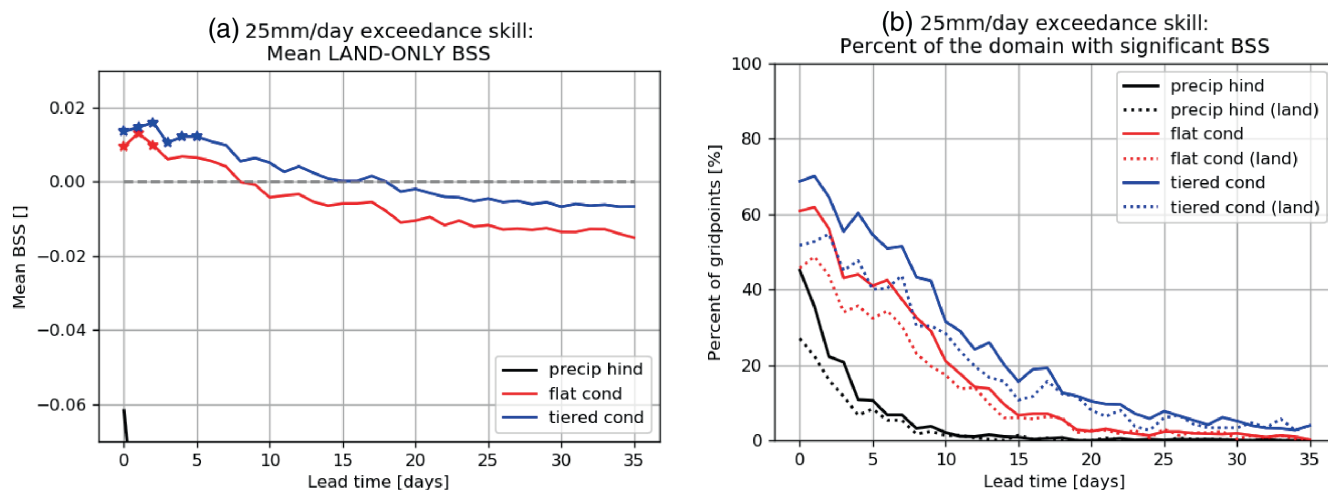
FIGURE 8 Summary BSS metrics for 25 mm·day^{-1} exceedance: (a) average land-only BSS over the complete SE Asia domain as a function of lead time; (b) percentage of the grid points where BSS is statistically significant as a function of lead time. In the latter, continuous lines correspond to all points and the dotted lines to land-only grid points. Statistical significance is assigned where the BSS is higher than 0 as described in Section 4.1, and significant lead times are indicated by a star symbol for each forecast type in (a) [Colour figure can be viewed at wileyonlinelibrary.com]

climatology as a reference can be masking the fact that the forecasts contain useful information (Mason, 2004).

The results for the P90 exceedance are analogous, albeit they reveal slightly higher levels of skill in the case of P90 exceedance than with a fixed threshold of 25 mm·day^{-1} (Figure S3).

A summary comparison between methods is presented in Figure 8, which shows the average land-only BSS over the full Southeast Asia domain (Figure 8a) and the percentage of grid points where the BSS is statistically significant (Figure 8b) for the three sets of predictions: simulated precipitation (black), flat-conditioned (red), and tiered-conditioned (blue). For every forecast lead, the pattern-conditioned methods outperform the simulated precipitation significantly, according to these metrics. The average land-only BSS for the simulated precipitation is negative for every lead time. In the case of the pattern-conditioned hindcasts, the average BSS remains positive for lead days into the second and third forecast weeks, depending on the case, but only the initial lead times (3–5 days) are statistically significant. For both exceedance cases, the tiered set shows average BSS values that are larger than in the flat case for lead times longer than a few days. This means that the average BSS for the tiered set shows positive skill for longer lead times (around 15 days for 25 mm·day^{-1} and 23 days for P90 in Figure S4), while the significant values are limited to the first week. As discussed above, BSS can be very unstable for the assessment of skill to predict rare events, and even negative values can be hiding useful information in the forecasts (Mason, 2004). Moving beyond the domain-mean BSS, Figure 8b also presents the percentage of the domain

with statistically significant BSS. In agreement with the findings in Figures 7 and S3, the pattern-conditioned hindcasts show significant skill over larger regions, whether we consider land-only points or the full domain. It is particularly noteworthy that the simulated precipitation has significant skill over less than 50% of the full domain at lead time 0 and decreases rapidly. For the land-only case (dotted black lines in b), the values are even smaller at 28%. This panel also shows that the forecasts conditioned by the tiered patterns are statistically significant over a larger area than the flat-conditioned case (5–10% larger).

4.2 | Countrywide skill assessment

The spatial variability of skill was also explored by considering the regional average of BSS over land points in six different boxes illustrated in Figure 9. The outer box indicates the full domain used to define the tier-2 and flat clusters, and the inner five boxes depict some regions of interest. Averages of the 25 mm·day^{-1} exceedance BSS over those six boxes are presented in Figure 10. Note that this is a measure of average skill, and not the skill of the forecast averaged over the regions. Simulated precipitation, however, shows little to no skill for every case. Only limited positive BSSs are observed over the Philippines and Vietnam for short lead times (Figure 10d,e). These results also show that there are significant differences between the levels of skill of the different subregions. For example, the pattern-conditioned hindcasts show positive skill over Borneo until around lead day 13 (and significant until day 11), but only until around lead day 4 over Indonesia and
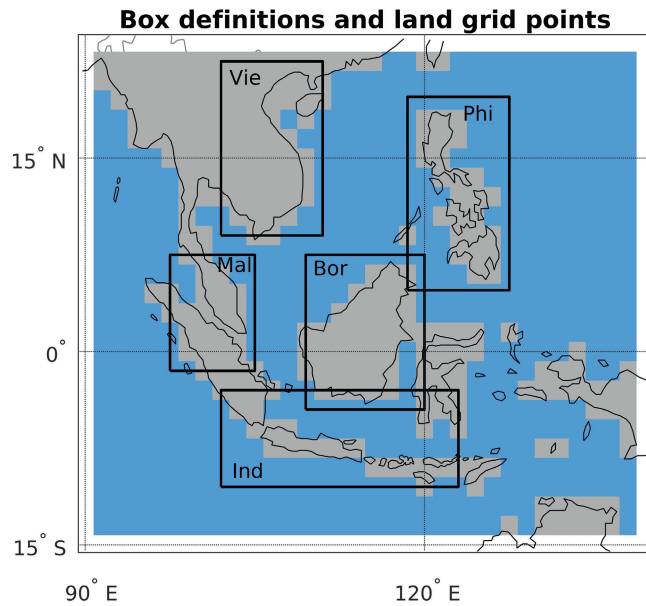
FIGURE 9 Map indicating the grid points and boxes considered in the regional skill assessments. Grey grid boxes indicate the land points. Black boxes describe the areas identified as Mal (mainly Peninsular Malaysia and part of N Sumatra), Vie (Vietnam and surrounding areas of Cambodia, Laos, Thailand), Phi (Philippines), Ind (Southern Indonesia), and Bor (Borneo). The GloSea5 land–sea mask at S2S resolution of 1.5° was obtained from https://apps.ecmwf.int/datasets/data/s2s [Colour figure can be viewed at wileyonlinelibrary.com]

lead day 6 over Vietnam and only significant until day 2 (Figure 10b,c,e). In some cases, there is a slight advantage of the tiered over the flat pattern-conditioned hindcasts for longer lead times, which implies that the BSS becomes negative more rapidly in the flat case. In general, the levels of skill are very similar and statistically significant differences between the pattern-conditioned methods are seen after lead day 5 over some of the regions (Malaysia, Borneo, Philippines).

The equivalent results for P90 exceedance are shown in Figure S5. A comparison between Figure 10 and Figure S5 reveals that there are slightly higher levels of skill of the pattern-conditioned hindcasts for the P90 exceedance case and that in general, the BSS values remain positive for longer lead times. This is in agreement with the comparison between Figures 7 and S3.

4.3 | Impact of spatial aggregation on prediction skill

Considering the intrinsic limitations of extended-range predictions (Vitart *et al.,* 2017; White *et al.,* 2017; Meehl *et al.,* 2021), in this subsection we conduct an evaluation of the GS5 skill beyond the grid-point level, by aggregating

over a group of nearby grid points. We do this with the aim of accounting for the potential inability of a subseasonal forecast system to predict the exact location of a HIW event. This evaluation is performed by considering the averaged grid-point based probabilities, which result in the expected spatial rate at which the precipitation criteria are exceeded within an N by N grid points box. This analysis is similar to the use of the fractions skill score (FSS: Roberts and Lean, 2008.) to evaluate the spatial scale at which the predictions are skilful.

The effect of the spatial scale on the average land-only BSS for $25\,mm \cdot day^{-1}$ exceedance is shown in Figure 11. Given that the S2S Database 1.5° resolution is quite coarse, only aggregations of up to 5×5 grid points were considered. The spatial aggregation results in higher BSS values for approximately the first 18 days of hindcast for both pattern-conditioned hindcasts (Figure 11a,b). The spatial aggregation increases skill levels and some more lead days become statistically significant, but it does not result in a significant extension of the skilful prediction horizon for the pattern-condition methods, especially in the tiered set. This can be interpreted, for the case of a dichotomous prediction, as having more hits and/or fewer misses and false alarms for the lead times at which skill was already positive. In the case of the simulated precipitation exceedance (Figure 11c), the increase in skill levels observed for about 12 lead days from the start of the forecast is much more significant, going from negative skill at grid-point level to BSS values more than double those in the pattern-conditioned cases for larger aggregations. In this specific case, it does result in an extension of predictability, but this is still restricted to the first forecast week. Analogous results were obtained for P90 exceedance S6, but with slightly longer skill horizons, consistent with P90 being—on average—a less extreme condition over the region.

5 | CASE STUDY

To illustrate the value of the methods presented before in the context of impact-based forecasting, we present here a case study focused on a HIW event that affected Jakarta and neighbouring regions at the beginning of 2020. This event caused severe flooding and landslides affecting the Greater Jakarta region, causing around 80 deaths and prompting the evacuation of over 100,000 people (JBA, 2020; Reliefweb, 2020). In addition to the severity of its impacts, we selected this event for being outside the GS5 hindcast period. This case study has previously been introduced by H21. The results presented here complement the qualitative analysis of the events presented in table 3 of H21, in which the value of a perfect pattern forecast was tested as a constraint to the probabilities of extreme rainfall
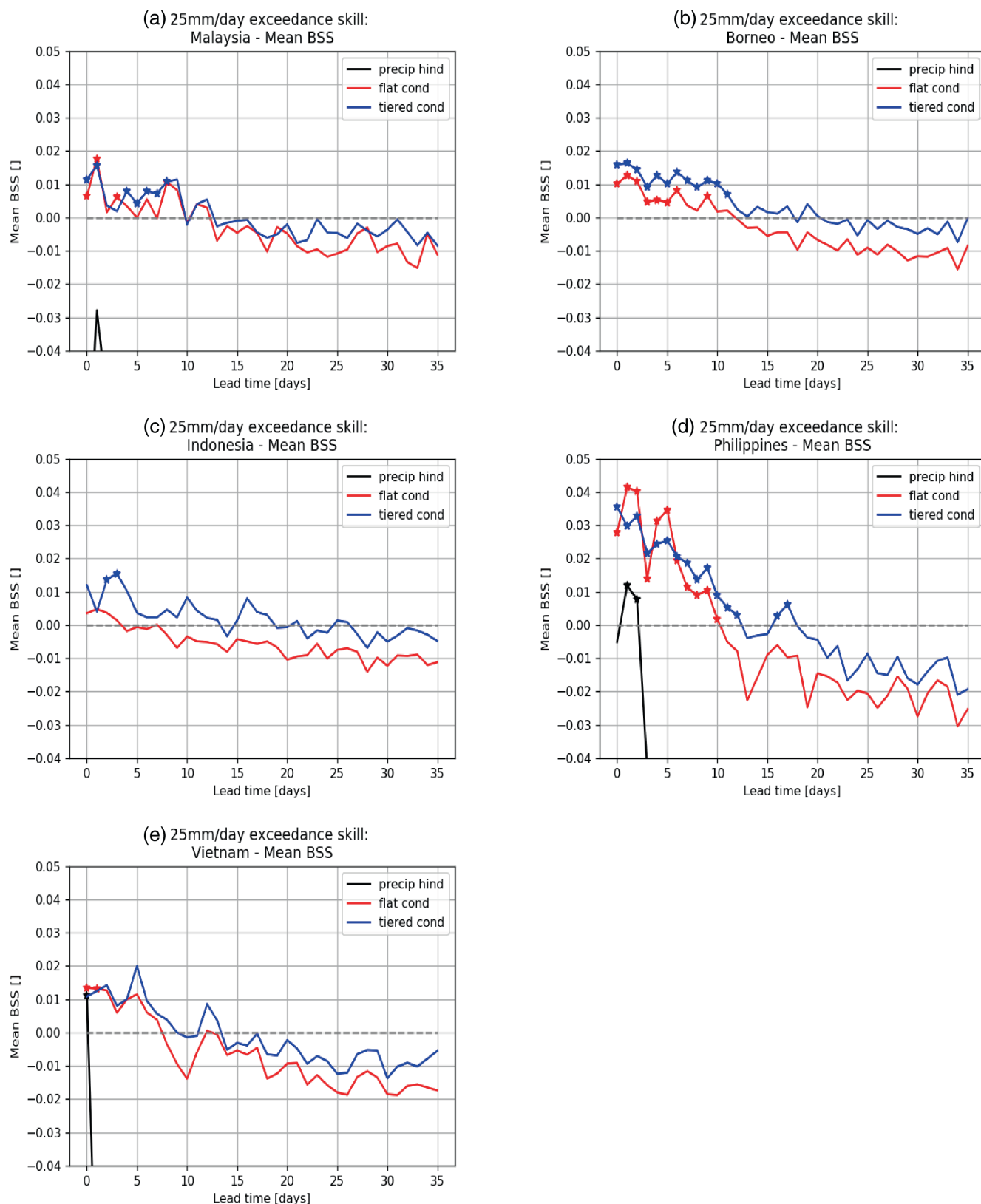
FIGURE 10 Average land-only BSS for 25 mm·day^{-1} precipitation exceedance over (a) Malaysia, (b) Borneo, (c) Indonesia, (d) Philippines, (e) Vietnam. Significant BSS values are indicated by a star symbol for each forecast type and assessed as described in Section 4.1 [Colour figure can be viewed at wileyonlinelibrary.com]

25mm/day exceedance skill: effect of spatial aggregation
Mean LAND-ONLY BSS
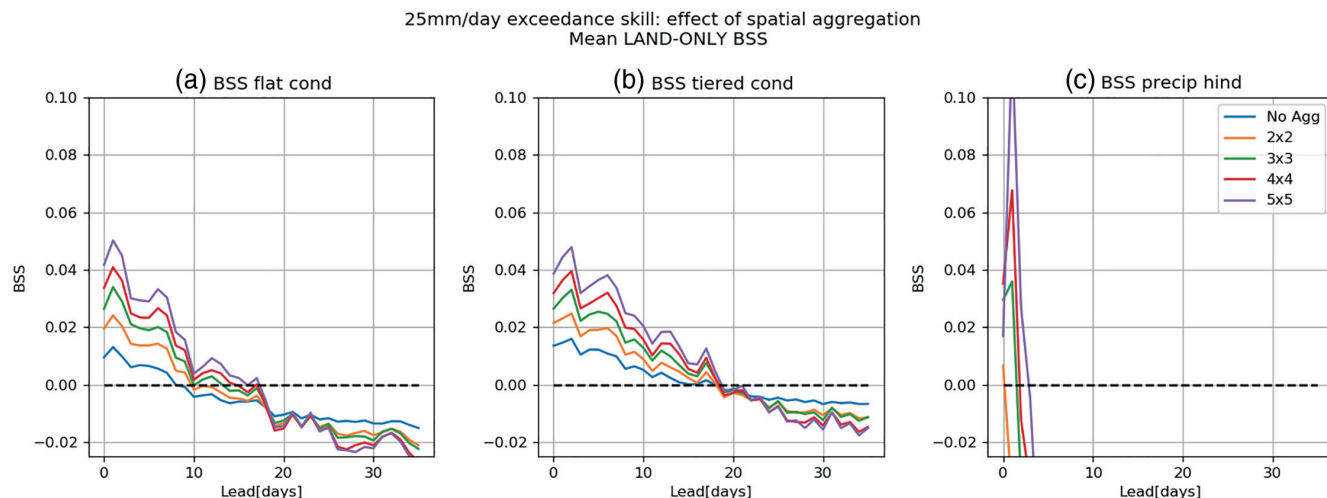
FIGURE 11 Effect of the spatial aggregation in mean land-only BSS for 25 mm·day^{-1} exceedance as a function of the grid-box size. No aggregation corresponds to the S2S 1.5° resolution and increasingly large square boxes were considered, with size expressed as multiples of a 1.5° grid box. (a) Flat conditioned hindcasts, (b) tiered conditioned hindcasts, (c) simulated precipitation. Significant BSS values up to lead day 20 are indicated by a star symbol for each box size and assessed as described in Section 4.1 [Colour figure can be viewed at wileyonlinelibrary.com]

over specific dates and regions. In this case we aim to test the full potential of the methods by including the forecast probabilities of WP membership.

We considered GS5 forecasts with valid dates in the period between December 30, 2019 and January 5, 2020. Those forecasts that are part of the S2S database are started every day and have four ensemble members (Vitart *et al.*, 2017). Four different start dates were chosen (December 9, 16, 23, and 30, 2019) which allowed the skill assessment of the predictions at lead times of 0–6 days (week 0), 7–13 days (week 1), 14–20 days (week 2), and 21–27 days (week 3). Forecast fields of 850-hPa wind vectors were used to derive probabilistic prediction for the flat and tiered WP memberships that were later converted into predicted probabilities of exceedance of the 25 mm·day^{-1} threshold (Section 2.4, Figure 1).

The precipitation features during that event as depicted by GPM are shown in Figure 12. The left panel indicates the grid boxes over which the threshold was exceeded and over how many days during the target week. The right panel presents the total precipitation accumulated over that same period. This shows that large portions of Indonesia, in particular West Borneo and Java, experienced extreme rainfall. In some locations, the exceedance criteria were met on three–four days over that seven-day period.

Forecast probabilities of exceedance derived from the flat and tiered WPs are presented in Figure 13. The probabilities presented in each of the four panels account for the expected frequency of exceedance of the threshold over the target period. The regions over which the forecast

Wekly 25mm/d Exceedance
Week start: 2019-12-30

Observed exceedance [days]   Precipitation [mm/week]
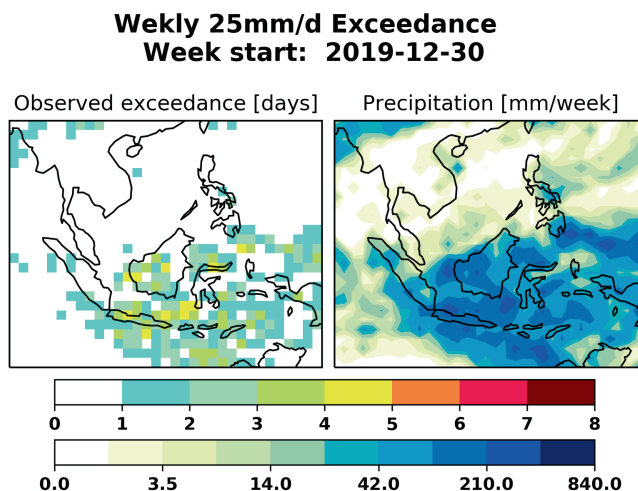
FIGURE 12 GPM 25 mm·day^{-1} precipitation for the week beginning on December 30, 2019. (a) Number of days in the week in which the threshold was exceeded. (b) Weekly accumulated precipitation [Colour figure can be viewed at wileyonlinelibrary.com]

probabilities are larger than the pattern-conditioned climatological exceedance are surrounded by a dashed line. Even though the probabilities presented here are relatively small, even for lead week 0, they are larger than the climatological odds over most of the region of interest for all the lead times considered. The flat and tiered methods provide similar results, though the probabilities derived from the flat patterns tend to be slightly higher. On closer inspection, one can note that for lead week 1 (Figure 13b) Java is not included in the areas over which the tiered
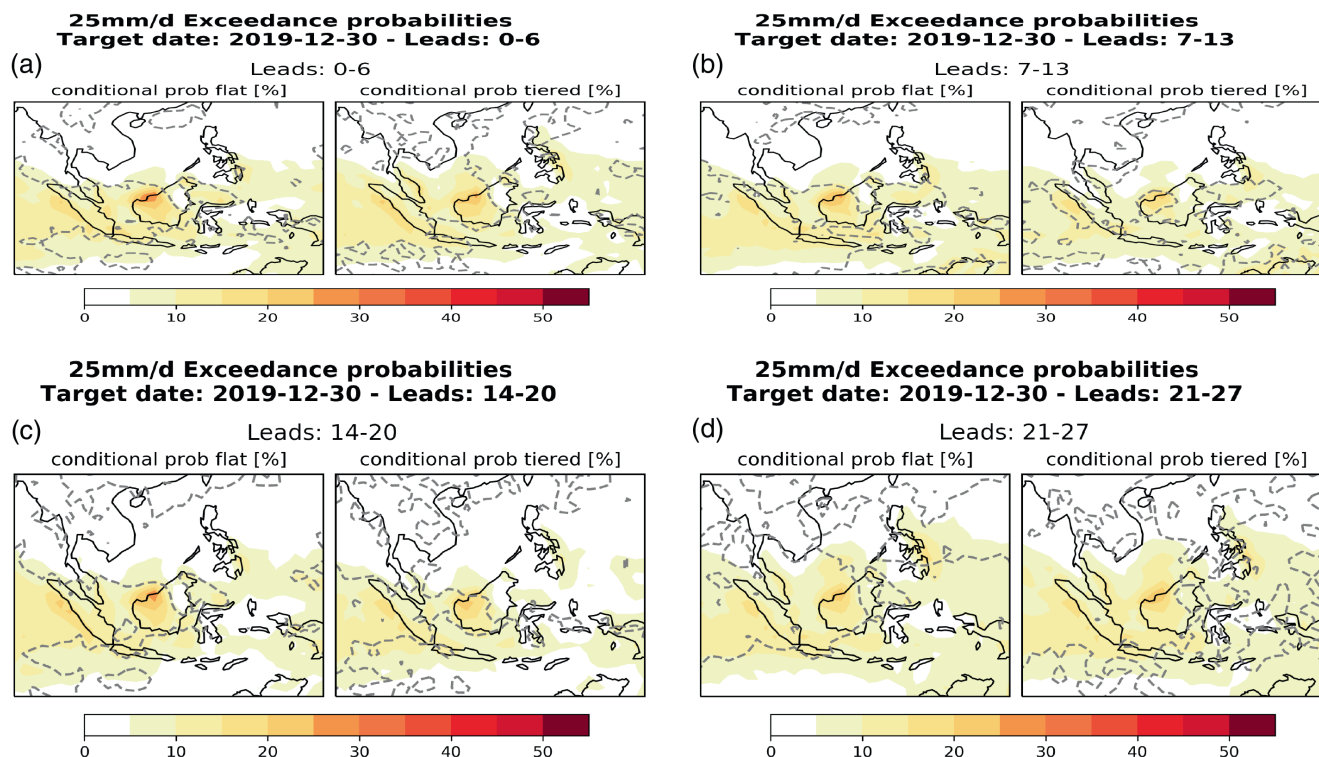
FIGURE 13 Weekly aggregated probabilities of exceedance of the 25 mm·day^{-1} threshold obtained from the flat and tiered pattern-conditioned hindcasts for the week starting on December 30, 2019. Obtained from lead days: (a) 0–6 (week 0); (b) 7–13 (week 1); (c) 14–20 (week 2); (d) 21–27 (week 3). Grey lines surround regions where the forecast probabilities exceed the pattern-conditioned climatological probabilities of exceedance [Colour figure can be viewed at wileyonlinelibrary.com]

pattern predicts above-normal probabilities. It is important to point out that the forecasts have four ensemble members, as opposed to seven members in the hindcasts, resulting in higher chances of duplication in the assigned patterns and therefore in a more limited ability to capture smaller-scale detail in the forecast risks.

A comparison between the probabilities of exceeding 25 mm·day^{-1} as directly simulated by the GS5 model and the GPM-derived climatological probability of exceedance shows that GS5 predicted above-normal probabilities of simulated precipitation exceeding 25 mm·day^{-1} in at least one day of the week over Java, but not over the rest of the focus region (Figure 14; note the difference in colour scale). Furthermore, the forecast probabilities derived from simulated precipitation decreased significantly with increasing lead time, and by lead week 2 the system does not provide any useful information over land. This is in agreement with what was more generally observed in Section 4.1 for the hindcast set.
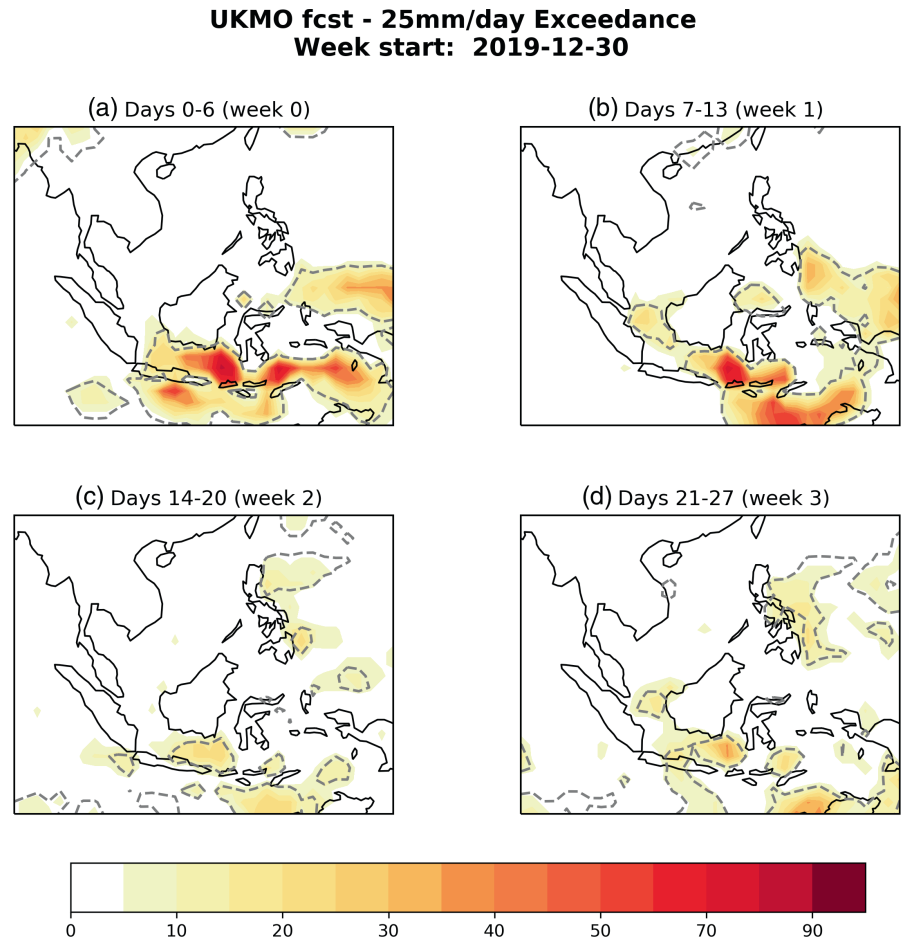
This simple analysis illustrates how the methodologies developed in this project could have been used to inform an early-warning system by providing a forecast of above-normal probabilities of exceedance. It is important to notice that this constitutes an example and that there are specific settings of the methods presented here

that would need to be tailored to the local needs and applications, such as the exceedance threshold or the relevant trigger probabilities for an alert. Furthermore, these are results from the analysis of a single case study and provide no information about the potential level of skills of a prototype early-warning system derived from the methodology presented. The true value of the implementation of these methods should be addressed by a robust verification considering a large set of forecast events.

6 | DISCUSSION AND CONCLUSIONS

This work has evaluated the use of weather patterns (WPs) defined through tiered and flat clustering methodologies to condition forecasts of extreme precipitation over Southeast Asia. For this purpose, the skill of the pattern-conditioned predictions derived from the UK Met Office GloSea5 (GS5) S2S hindcast was benchmarked against an observation-based seasonally varying climatology and the simulated precipitations obtained from the same prediction system. The analysis was performed for land points over the whole of Southeast Asia and over five

FIGURE 14 GloSea5 probabilities of daily simulated precipitation exceeding 25 mm·day^{-1} for the week beginning on December 30, 2019. Obtained from lead days: (a) 0–6 (week 0); (b) 7–13 (week 1); (c) 14–20 (week 2); (d) 21–27 (week 3) [Colour figure can be viewed at wileyonlinelibrary.com]

UKMO fcst - 25mm/day Exceedance
Week start: 2019-12-30

(a) Days 0-6 (week 0)
(b) Days 7-13 (week 1)
(c) Days 14-20 (week 2)
(d) Days 21-27 (week 3)

smaller regions: Malaysia, Borneo, Indonesia, the Philippines, and Vietnam.

GS5 showed skill for the assignment of both sets of WPs extending beyond lead day 10, with statistically significant larger skill in the case of the tiered WP set, attributed to the influence of the longer timescales associated with the tier-1 regimes: seasonal cycle, monsoons, and ENSO. These WP membership predictions were then convolved with GPM-based pattern-conditioned climatological probabilities of precipitation exceedance, resulting in the pattern-conditioned hindcasts for different extreme rainfall criteria. The performance of such predictions was evaluated against climatological benchmarks and the model-derived precipitation exceedance hindcasts. It was observed that, on average, the pattern-conditioned hindcasts were skilful and outperformed the model-based precipitation predictions for lead days extending to days 10–20, depending on the specific criteria and region. Both the flat and tiered pattern-conditioned hindcasts outperformed the model-derived precipitation exceedance predictions significantly over every region after a few initial lead days. The tiered pattern-conditioned hindcasts were shown to be more skilful than the flat counterpart for the land-only average skill and three out of the five regions

analysed for lead times longer than a few days. On average, these results show that GS5 has longer skill in the pattern-conditioned forecasts of precipitation exceedance than in the pattern assignment. This suggests that, despite forecasting a pattern that does not match the observed one exactly, the system might be assigning a similar pattern in terms of the relevant precipitation signatures. Potential improvements to the methodology presented could be derived from exploiting this feature, in particular given the small size of the GS5 hindcast ensemble. The comparison of the skill for predicting P90 and 25 mm·day^{-1} exceedance revealed higher levels of skill in the percentile exceedance, likely related to the fact that P90 is more frequently exceeded over the study region.

Acknowledging the limitations of extended-range initialised predictions and of the skill assessment of rare-event predictions, the skill evaluation was then repeated to consider spatial aggregation, allowing the system to have some errors in the specific location of a heavy precipitation event. The spatial aggregation was constructed by considering the spatial rate of heavy precipitation within boxes of increasing size. This analysis showed skill improvements in the pattern-conditioned hindcasts that extended up to lead day 20. In the case of

the model-derived precipitation exceedance hindcasts, the effect of spatial aggregation was beneficial only in the first 10 days of the hindcast, but the magnitude of the increases in skill was very significant and resulted in the emergence of skill over the first forecast week. The impact of the spatial aggregation on the pattern-conditioned hindcasts was an increase in skill but not a significant extension of the useful prediction horizon.

To illustrate the implementation of the methods in a particular situation, an out-of-sample extreme rainfall event was presented as a use case. This extreme rainfall event triggered flooding and landslides over Indonesia and caused widespread damage to property and life. We tested the application of the pattern-conditioned predictions of precipitation exceedance for this event using GS5 forecasts, and compared the resulting probabilities of exceedance with those derived from the precipitation forecasts. The analysis showed that the pattern-conditioned methods forecast above-normal probabilities of exceedance over the affected area as early as four weeks in advance, with the flat method resulting in marginally better predictions. The simulated precipitation forecast only resulted in above-normal probabilities for lead week 0 and over a more restricted area, showing no useful signal over land for longer lead times.

Overall, the results presented here demonstrate the use of weather patterns over Southeast Asia for the development of precipitation exceedance predictions that outperform the simulated precipitation for lead times extending into weeks 2 and 3, depending on the specific criteria and region. On average, the flat and tiered pattern definitions resulted in similar levels of skill, though there might be added value in the use of the tiered patterns for the evaluation of predicable circulation patterns or the identification of "windows of opportunity" for enhanced prediction. Furthermore, the tiered patterns enable attribution of changes in forecast risk to more predictable phenomena such as monsoonal variations, MJO, and ENSO. Weather risk is then considered in the context of the larger-scale environmental picture. Overall, the results presented here and those in the companion work of H21 suggest that the tiered methodology has the added value of providing a "forensic" power to understand past events and the inability of a prediction system to capture them, as well as identifying windows of opportunity for the prediction of HIW in the future. These features are in line with the guidelines derived from recent review work on the topic of S2S prediction (Moron *et al.,* 2018; Mariotti *et al.,* 2020; Meehl *et al.,* 2021).

The levels of skill obtained throughout this study are modest. Considering the challenges of assessing skill in the prediction of rare events (Casati *et al.,* 2008; Mason, 2012; Sillmann *et al.,* 2017; Contzen *et al.,* 2022; Ebert and Milne, 2022), further exploration with additional metrics might be advisable when attempting to develop a method tailored for concrete applications. However, the results presented here do suggest avenues for the improvements of the methods. Considering a probabilistic WP assignment that benefits from the identification of WPs that are "close" in their structure or impacts should be further explored. It was also noted that the GS5 S2S prediction ensemble is very small, with seven members in the hindcast set and four in the forecasts, which introduces a significant limitation. Additionally, it has to be pointed out that no form of calibration was applied to GS5 outputs, and such post-processing methodologies can also derive skill improvements. Even though these results show promise, the true value of the pattern-conditioned predictions should be evaluated in terms of the design of regional impact-based criteria (exceedance thresholds, spatial and temporal aggregation) tailored to a specific use, and after a thorough verification on events outside the training hindcast set and with multiple skill metrics that target different aspects of skill (resolution, sharpness, etc.).

Even though it might be possible to achieve similar levels of skill through model calibration or other post-processing methodologies, and this is worth pursuing, we consider that the methodology presented here has added value. The extra steps of evaluating the dominant weather patterns are of forensic value and allow us to identify conditions for which GS5 has more or less predictive skill. This is of value in terms of both acting on the forecasts and also informing model development. This work has presented an initial proof-of-concept for the use of pattern-conditioned forecasts over Southeast Asia, and the suggested refinements will be the focus of future work.

AUTHOR CONTRIBUTIONS

DATA AVAILABILITY STATEMENT

This study presents post-processing of data that are publicly available from the S2S database hosted by the ECMWF at https://apps.ecmwf.int/datasets/data/s2s. Sample code to replicate the process can be provided by the authors upon request.

ORCID

Paula L. M. Gonzalez https://orcid.org/0000-0003-0154-0087

Emma Howard https://orcid.org/0000-0003-0108-1220

Samantha Ferrett https://orcid.org/0000-0003-4726-847X

Oscar Martínez-Alvarado https://orcid.org/0000-0002-5285-0379

REFERENCES

Birch, C., Webster, S., Peatman, S., Parker, D., Matthews, A., Li, Y. and Hassim, M. (2016) Scale interactions between the MJO and the western maritime continent. *Journal of Climate*, 29, 2471–2492.

Boer, G. (2003) Predictability as a function of scale. *Atmosphere-Ocean*, 41, 203–215.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3 https://journals.ametsoc.org/view/journals/mwre/78/1/1520-049%3_1950_078_0001_vofeit_2_0_co_2.xml.

Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E., Brown, B. and Mason, S. (2008) Forecast verification: current status and future directions. *Meteorological Applications: A Journal of Forecasting, Practical Applications, Training Techniques and Modelling*, 15, 3–18.

Contzen, J., Dickhaus, T. and Lohmann, G. (2022) Variability and extremes: statistical validation of the alfred wegener institute earth system model (AWI-ESM). *Geoscientific Model Development*, 15, 1803–1820 https://gmd.copernicus.org/articles/15/1803/2022/.

Cortesi, N., Torralba, V., Lledó, L., Manrique-Suñén, A., Gonzalez-Reviriego, N., Soret, A. and Doblas-Reyes, F.J. (2021) Yearly evolution of euro-Atlantic weather regimes and of their sub-seasonal predictability. *Climate Dynamics*, 56, 1–32.

Davidson, R. and MacKinnon, J.G. (2000) Bootstrap tests: how many bootstraps? *Econometric Reviews*, 19, 55–68. https://doi.org/10.1080/07474930008800459.

Ebert, P.A. and Milne, P. (2022) Methodological and conceptual challenges in rare and severe event forecast verification. *Natural Hazards and Earth System Sciences*, 22, 539–557.

Eckstein, D., Künzel, V., Schäfer, L. and Winges, M. (2019) Global climate risk index 2020. *Tech. Rep.*, German watch, Berlin www.germanwatch.org.

Ferrett, S., Frame, T.H.A., Methven, J., Holloway, C.E., Webster, S., Stein, T.H.M. and Cafaro, C. (2021) Evaluating convection-permitting ensemble forecasts of precipitation over southeast asia. *Weather and Forecasting*, 36, 1199–1217 https://journals.ametsoc.org/view/journals/wefo/36/4/WAF-D-20%-0216.1.xml.

Ferrett, S., Yang, G.-Y., Woolnough, S.J., Methven, J., Hodges, K. and Holloway, C.E. (2020) Linking extreme precipitation in Southeast Asia to equatorial waves. *Quarterly Journal of the Royal Meteorological Society*, 146, 665–684.

Ghil, M., Groth, A., Kondrashov, D. and Robertson, A.W. (2019) Extratropical sub-seasonal to seasonal oscillations and multiple regimes: the dynamical systems view. In: *Sub-Seasonal to Seasonal Prediction*. Elsevier, pp. 119–142. https://www.sciencedirect.com/science/article/pii/B9780128117149000061?via%3Dihub

Glahn, H.R. and Lowry, D.A. (1972) The use of model output statistics (MoS) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, 11, 1203–1211 https://journals.ametsoc.org/view/journals/apme/11/8/1520-045%0_1972_011_1203_tuomos_2_0_co_2.xml.

Gupta, S. (2010) Synthesis report on asean countries disaster risks assessment – asean disater risk manage-ment initiative. Technical Report., World Bank, UNISDR, ASEAN and GFDRR. https://www.unisdr.org/we/inform/publications/18872.

Hassim, M.E. and Timbal, B. (2019) Observed rainfall trends over Singapore and the maritime continent from the perspective of regional-scale weather regimes. *Journal of Applied Meteorology and Climatology*, 58, 365–384.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R. and Schepers, D. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049.

Hohenegger, C. and Schar, C. (2007) Atmospheric predictability at synoptic versus cloud-resolving scales. *B. Am. Meteorol. Soc.*, 88, 1783–1794.

Howard, E., Thomas, S., Frame, T.H., Gonzalez, P.L., Methven, J., Martínez-Alvarado, O. and Woolnough, S.J. (2021) Weather patterns in southeast asia: relationship with tropical variability

and heavy precipitation. *Quarterly Journal of the Meteorological Society* https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4227.

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., et al. (2015) Integrated multi-satellite retrievals for GPM (IMERG) technical documentation. *NASA/GSFC Code*, 612, 47.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J. and Tveito, O.E. (2008) Classifications of atmospheric circulation patterns: recent advances and applications. *Ann. NY Acad. Sci.*, 1146, 105–152.

JBA (2020) *A stark reminder of jakarta's increasing vulnerability to flood risk*. Available at: https://www.jbarisk.com/flood-services/event-response/jakarta-flooding/%. [Accessed 17th March 2021].

Love, B.S., Matthews, A.J. and Lister, G.M. (2011) The diurnal cycle of precipitation over the maritime continent in a high-resolution atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, 137, 934–947.

MacLachlan, C., Arribas, A., Peterson, K.A., Maidens, A., Fereday, D., Scaife, A.A., Gordon, M., Vellinga, M., Williams, A., Comer, R.E., Camp, J., Xavier, P. and Madec, G. (2015) Global seasonal forecast system version 5 (glosea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141, 1072–1084 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2396.

Mariotti, A., Baggett, C., Barnes, E.A., Becker, E., Butler, A., Collins, D.C., Dirmeyer, P.A., Ferranti, L., Johnson, N.C., Jones, J., Kirtman, B.P., Lang, A.L., Molod, A., Newman, M., Robertson, A.W., Schubert, S., Waliser, D.E. and Albers, J. (2020) Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, 101, E608–E625 https://journals.ametsoc.org/view/journals/bams/101/5/bams-d-%18-0326.1.xml.

Mason, I. (1982) A model for assessment of weather forecasts. *Australian Meteorological Magazine*, 30, 291–303.

Mason, S. (2012) Do statistical models trade resolution for reliability? In: *ECMWF Annual Seminar on Seasonal Prediction*, Reading, UK. https://www.ecmwf.int/node/14827.

Mason, S.J. (2004) On using "climatology" as a reference strategy in the brier and ranked probability skill scores. In: *Monthly Weather Review*, Vol. 132, pp. 1891–1895 https://journals.ametsoc.org/view/journals/mwre/132/7/1520-04%93_2004_132_1891_oucaar_2.0.co_2.xml.

Mastrantonas, N., Magnusson, L., Pappenberger, F. and Matschullat, J. (2022) What do large-scale patterns teach us about extreme precipitation over the mediterranean at medium-and extended-range forecasts? *Quarterly Journal of the Royal Meteorological Society*, 148, 875–890.

Meehl, G.A., Richter, J.H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., Donat, M.G., England, M.H., Fyfe, J.C., Han, W., Kim, H., Kirtman, B.P., Kushnir, Y., Lovenduski, N.S., Mann, M.E., Merryfield, W.J., Nieves, V., Pegion, K., Rosenbloom, N., Sanchez, S.C., Scaife, A.A., Smith, D., Subramanian, A.C., Sun, L., Thompson, D., Ummenhofer, C.C. and Xie, S.P. (2021) Initialized earth system prediction from subseasonal to decadal timescales. *Nature Reviews Earth and Environment*, 2, 340–357. https://doi.org/10.1038/s43017-021-00155-x.

Moron, V., Robertson, A.W., Qian, J.-H. and Ghil, M. (2015) Weather types across the maritime continent: from the diurnal cycle to interannual variations. *Frontiers in Environmental Science*, 2, 65.

Moron, V., Robertson, A.W. and Vitart, F. (2018) Editorial: sub-seasonal to seasonal predictability and prediction of monsoon climates. *Frontiers in Environmental Science*, 6, 83 https://www.frontiersin.org/article/10.3389/fenvs.2018.00083.

Philipp, A., Bartholy, J., Beck, C., Erpicum, M., Esteban, P., Fettweis, X., Huth, R., James, P., Jourdain, S., Kreienkamp, F., Krennert, T., Lykoudis, S., Michalides, S.C., Pianko-Kluczynska, K., Post, P., Álvarez, D.R., Schiemann, R., Spekat, A. and Tymvios, F.S. (2010) Cost733cat–a database of weather and circulation type classifications. *Physics and Chemistry of the Earth*, 35, 360–373.

Reliefweb, OCHA Services. (2020) *Situation Update No. 1 - Massive Floods in Greater Jakarta Area, Indonesia - January 5, 2020*. Available at: https://reliefweb.int/report/indonesia/situation-update-no-1-massive-fl%oods-greater-jakarta-area-indonesia-05-january-2020/. [Accessed 17th March 2021].

Roberts, N.M. and Lean, H.W. (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136, 78–97 https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr%2123.1.xml.

Selz, T. and Craig, G.C. (2015) Upscale error growth in a high-resolution simulation of a summertime weather event over Europe. *Monthly Weather Review*, 143, 813–827.

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L.V., Hegerl, G., Seneviratne, S.I., Vautard, R., Zhang, X. and Zwiers, F.W. (2017) Understanding, modeling and predicting weather and climate extremes: challenges and opportunities. *Weather and Climate Extremes*, 18, 65–74 https://www.sciencedirect.com/science/article/pii/S2212094717%300440.

Vigaud, N., Robertson, A.W. and Tippett, M.K. (2018) Predictability of recurrent weather regimes over north america during winter from submonthly reforecasts. *Monthly Weather Review*, 146, 2559–2577.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A.W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R. and Zhang, L. (2017) The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 98, 163–173 https://journals.ametsoc.org/view/journals/bams/98/1/bams-d-1%6-0017.1.xml.

Vogel, P., Knippertz, P., Fink, A.H., Schlueter, A. and Gneiting, T. (2020) Skill of global raw and postprocessed ensemble predictions of rainfall in the tropics. *Weather and Forecasting*, 35, 2367–2385.

Wang, Y., Jin, S., Sun, X. and Wang, F. (2019) Winter weather regimes in southeastern China and its intraseasonal variations. *Atmosphere*, 10, 271.

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2007) The discrete brier and ranked probability skill scores. *Monthly Weather Review*, 135, 118–124 https://journals.ametsoc.org/view/journals/mwre/135/1/mwr3280%1.xml.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A.P., Eggen, B., Graham, R., Kjellström, E.,

Becker, E., Pegion, K.V., Holbrook, N.J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T.J., Street, R., Jones, L., Remenyi, T.A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B. and Zebiak, S.E. (2017) Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications*, 24, 315–325 https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.165%4.

Wilks, D.S. (2019) *Statistical Methods in the Atmospheric Sciences*, 4th edition. London, UK: Elsevier.

Ying, Y. and Zhang, F. (2017) Practical and intrinsic predictability of multi-scale weather and convectively coupled equatorial waves during the active phase of the MJO. *Journal of the Atmospheric Sciences*, 74, 3771–3785.

Zhang, F., Bei, N., Rotunno, R., Snyder, C. and Epifanio, C.C. (2007) Mesoscale predictability of moist baroclinic waves: convection-permitting experiments and multistage error growth dynamics. *Journal of the Atmospheric Sciences*, 64, 3579–3594.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.