https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

A synoptic clustering-based definition of South China Sea summer monsoon onset and application to seasonal prediction

Dzung Nguyen-Le[1]

[1]Department of Space and Earth Sciences, University of Science and Technology of Hanoi (USTH), Vietnam Academy of Science and Technology (VAST), Hanoi, Vietnam

Correspondence to: Dzung Nguyen-Le (nguyen-le.dung@usth.edu.vn); ORCID: 0000-0003-0214-3411

Abstract. Accurately predicting the South China Sea summer monsoon (SCSSM) onset date is of crucial importance for effective water resource management, agricultural planning, and disaster prevention across East Asia and the Western North Pacific. However, reliable predictions at seasonal timescales remain challenging. To address this, we propose a synoptic circulation–based approach that defines monsoon onset through persistent large-scale circulation regimes identified by clustering low-level atmospheric fields. Using ECMWF SEAS5 seasonal forecasts, we evaluate onset predictions derived from the proposed regime-based definition against those based on a conventional zonal wind criterion. The regime-based definition yields systematic improvements in deterministic correlations, potential predictability, and categorical and probabilistic skill metrics. Skill gains are evident during both a dependent training period and an independent forecast period, with enhanced performance persisting at long lead times of up to five months. The improved predictability reflects the multi-timescale controls on monsoon onset, wherein slowly varying boundary forcing modulates large-scale circulation and subseasonal variability triggers the onset transition. By emphasizing regime persistence and structural evolution, the proposed framework better isolates the predictable component of onset variability and enhances forecast robustness. These findings demonstrate that circulation-regime–based definitions offer a physically grounded and effective alternative for seasonal prediction of monsoon onset.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

1. Introduction

The onset of the South China Sea (SCS), also known as the East Sea in Vietnam, summer monsoon (SCSSM) marks a rapid reorganization of the Asian monsoon system, characterized by a transition from a dry, trade-wind-dominated spring regime to a convectively coupled summer circulation (Wang and LinHo, 2002; Ding et al., 2004; Wang et al., 2004; Ding et al., 2015). Climatologically occurring in mid-May, this transition is marked by a reversal of lower-tropospheric zonal winds over the SCS, enhanced low-level moisture convergence, and the eastward retreat of the western North Pacific (WNP) subtropical high (WNPSH), leading to the establishment of deep convection over the basin (Zhou and Chan, 2007; Wang et al., 2009). As a precursor to the broader East Asian summer monsoon, variability in the timing of SCSSM onset strongly regulates the initiation of the regional rainy season over Southeast Asia and southern China. Previous studies have further shown that this transition is sensitive to subseasonal disturbances and remote forcing, including intraseasonal oscillations, tropical–extratropical wave interactions, and large-scale Indo–Pacific circulation anomalies that modulate low-level westerlies and convective instability over the SCS (Ding and Chan, 2005; Geen, 2021).

Although the SCSSM onset occurs every year, its interannual variability exerts a strong control on regional climate through modulation of large-scale circulation, moisture transport, and convective activity. Variations in onset date (OD) alter the establishment and persistence of low-level westerlies over the SCS and the WNP, thereby redistributing summer rainfall across East and Southeast Asia and shaping regional drought and flood risks. An early SCSSM OD is often followed by suppressed summer rainfall over subtropical East Asia, including the middle and lower reaches of the Yangtze River basin, whereas a delayed OD tends to prolong the pre-monsoon rainfall regime and enhance flood potential (Jiang et al., 2018; He and Zhu, 2015). Over Southeast Asia, early OD favors enhanced convection and an increased likelihood of extreme rainfall events, while delayed OD is associated with reduced low-level moisture convergence and suppressed heavy rainfall in May (Hu et al., 2022a, b). The SCSSM OD also modulates tropical cyclone activity over the WNP by influencing

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

background vorticity, vertical wind shear, and monsoon trough development, with earlier OD linked to increased tropical cyclone genesis and a higher frequency of landfalling storms along the southeastern coast of China (Chen et al., 2017; Huangfu et al., 2017; Wang and Chen, 2018). Together, these mechanisms identify SCSSM OD variability as a key driver of seasonal-scale climate variability with direct implications for agriculture, water resources, and disaster risk management.

Motivated by these impacts, substantial efforts have been devoted to predicting the SCSSM OD using dynamical seasonal forecast systems. Despite the chaotic nature of atmosphere, previous studies demonstrate that useful predictive skill for interannual variability in SCSSM onset exists at seasonal lead times of approximately one to three months (e.g., Zhu and Li, 2017; Martin et al., 2019; Chevuturi et al., 2019, 2021; Attada et al., 2022). This seasonal predictability is derived primarily from slowly evolving boundary conditions, most notably the El Niño–Southern Oscillation (ENSO), which modulates large-scale circulation over the WNP through the Indo-Pacific Ocean Capacitor effect (Xie et al., 2016). As a result, operational forecast systems such as the UK Met Office GloSea5 and the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasting system 5 (SEAS5) can skillfully discriminate between early and late onset years up to three months in advance, particularly when onset is defined using robust circulation-based metrics, such as the 850-hPa zonal wind (U850) index of Wang et al. (2004), rather than local precipitation alone (Bombardi et al., 2017, 2020; Chevuturi et al., 2019; Martin et al., 2019).

However, extending this predictive capability to longer lead times of four to five months remains a formidable challenge. State-of-the-art dynamical models encounter a pronounced spring predictability barrier, during which the influence of ENSO conditions from the preceding winter weakens and stochastic atmospheric variability associated with the seasonal transition becomes dominant, often rendering forecasts initialized in winter (e.g., preceding December or January) statistically insignificant for predicting a May onset (Martin et al., 2019). In addition to intrinsic predictability limits, systematic model deficiencies persist. Coupled systems such as ECMWF SEAS5 frequently exhibit a cold sea surface temperature (SST) bias over the SCS and struggle to

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

realistically simulate the northward propagation of intraseasonal oscillations, leading to a systematic late-onset bias and degraded forecast skill at extended lead times (Chevuturi et al., 2021; Bui-Minh et al., 2024). Together, these limitations suggest that existing linear indices and conventional dynamical frameworks may not adequately capture the nonlinear precursors required to bridge the predictability gap between seasonal (~three-month) and long-range (~five-month) forecasts.

In parallel with advances in dynamical seasonal prediction, recent studies have proposed alternative circulation-based approaches to defining monsoon onset using clustering of synoptic-scale atmospheric patterns (e.g., Borah et al., 2013; Dai et al., 2021; Bui-Minh et al., 2024). These approaches typically employ self-organizing maps (SOM; Kohonen, 2001) to project high-dimensional atmospheric fields onto a finite set of representative circulation patterns, which are subsequently grouped using clustering algorithms such as K-means. The resulting clusters objectively distinguish between pre-monsoon and monsoon circulation regimes, allowing onset to be identified as the sustained transition from dry, pre-monsoon patterns to convectively active monsoon-associated patterns. By construction, this framework captures the full spatial structure and temporal evolution of the large-scale circulation, rather than relying on a single variable or threshold-based index. Bui-Minh et al. (2024) shows that synoptic-pattern clustering provides a more physically consistent and temporally coherent identification of monsoon onset than traditional index-based definitions. Importantly, such pattern-based definitions are more closely aligned with the dynamical structures resolved by numerical models, suggesting potential advantages for seasonal prediction.

Despite this promise, the extent to which synoptic-pattern-based definitions can improve the predictability of SCSSM OD in operational seasonal forecast systems, particularly at lead times exceeding three months, has not yet been systematically assessed. This study addresses this gap by developing a synoptic clustering–based definition of SCSSM onset using the combination of SOM and clustering techniques and evaluating its performance within a seasonal prediction framework. Specifically, the proposed onset definition will be applied to retrospective forecasts from the ECMWF SEAS5 system to assess whether it enhances prediction skill for SCSSM onset date at

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

extended lead times. By explicitly linking monsoon onset to large-scale circulation regimes rather than single-variable indices, this work aims to provide a more robust and dynamically meaningful framework for understanding and predicting the SCSSM onset. The remainder of this paper is organized as follows. Section 2 describes the data and methodology. Section 3 presents the synoptic clustering–based definition of the SCSSM OD. Section 4 evaluates its performance in seasonal prediction. Finally, discussions and conclusions are given in Section 5.

2. Datasets and methodology

2.1 Datasets

2.2.1 Observational data

The fifth generation of the ECMWF Reanalysis dataset (ERA5; Hersbach et al., 2020) for the period 1979 to the present, with a horizontal resolution of 0.25° × 0.25°, is used to analyze atmospheric circulation. The analysis focuses on 850-hPa zonal wind (U850), meridional wind (V850), and geopotential height (Z850). Pentad (5-day) means are computed from hourly ERA5 data. Monthly SST data are taken from the UK Met Office Hadley Centre Global Sea Ice and Sea Surface Temperature dataset (HadISST; Rayner et al., 2003), with a resolution of 1° × 1°, available from 1870 to the present. Anomalies are calculated by subtracting the 1981–2010 climatology from detrended data over the corresponding period of record.

2.2.2 Model data

The ECMWF SEAS5 (Johnson et al., 2019) coupled forecasting system consists of the Integrated Forecast System (IFS) atmospheric model, the HTESSEL land surface model, and the Nucleus for European Modelling of the Ocean (NEMO). SEAS5 provides seasonal forecasts at an O320 horizontal resolution (approximately 36 km) with 91 vertical levels in the atmosphere, and an ORCA 0.25° grid (approximately 27 km) with 75 vertical levels in the ocean. The SEAS5 hindcasts and forecasts are initialized on the first day of each month and integrated for seven months, using a

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

21-member ensemble for hindcasts over 1981–2016 and a 51-member ensemble for real-time forecasts from 2017 onward.

2.2 Methodology

2.2.1 Clustering

The methodology applied in this study follows Nguyen-Le et al. (2017, 2019) and is designed to classify synoptic-scale atmospheric circulation patterns over the SCS and surrounding regions before and after the SCSSM onset using SOM. From a synoptic perspective, SCSSM onset is characterized by the development of the low-level monsoon trough and cross-equatorial flow near 105°E, the eastward retreat of the WNPSH, and the establishment of a coherent monsoonal meridional circulation (Liu et al., 2016; Huangfu et al., 2017; Hu et al., 2018). Based on these dynamical features, SOM input vectors are constructed from pentad-mean ERA5 reanalysis fields of Z850, U850, and V850, which are concatenated for each pentad to form a single high-dimensional input vector for SOM training. A total of 864 input vectors is generated, corresponding to 18 pentads per year (from pentad 19, 1–5 April, to pentad 36, 25–29 June) over the 38-year period 1979–2016. The spatial domain spans 5°S–25°N and 95°E–135°E, yielding 161 longitude points and 81 latitude points. Consequently, each raw input vector contains $3 \times 161 \times 81 = 39123$ elements. Because these elements are not statistically independent owing to strong spatial coherence and inter-variable correlations, direct use of the full input vectors is neither efficient nor optimal. To reduce dimensionality and remove redundant information, principal component analysis (PCA) is applied jointly to the three original reanalysis fields. Prior to PCA, each variable is normalized to ensure comparable variance contributions, as both PCA and SOMs are sensitive to variable scaling. Only empirical orthogonal functions explaining 99% of the total variance are retained. The corresponding principal components (PCs) are then used as input to the SOM in place of the original fields. This procedure reduces the dimensionality of each input vector to $d = 144$, corresponding to a reduction

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

factor of approximately 270, thereby improving statistical independence among input elements and substantially reducing computational cost.

The SOM training outcome depends on both the lattice size and several training parameters, including the learning rate, neighborhood radius, and training length. A trial-and-error approach is therefore employed to identify an optimal SOM configuration. Multiple lattice sizes (4 × 4, 5 × 5, 6 × 6, 7 × 7, and 8 × 8) are tested and evaluated based on node occupancy, quantization error (QE), and topographic error (TE). Specifically, an excessive number of samples assigned to individual nodes (e.g., >10) indicates insufficient pattern separation and suggests the need for a larger lattice, whereas the presence of empty nodes implies over-partitioning and an excessively large lattice. Based on these criteria, a SOM with a 6 × 6 hexagonal lattice (36 nodes) is selected, using a learning rate of 0.2 and a neighborhood radius of three. To ensure numerical stability and convergence, the SOM is trained for two million iterations. Given the relatively limited sample size, bootstrap resampling is incorporated into the training procedure, whereby samples are randomly drawn with replacement from the original dataset. A total of 1000 SOM realizations are generated using this configuration. The final "master" SOM is selected as the realization exhibiting the lowest QE and TE (Fig. S1) and a relatively smooth Sammon mapping (Sammon, 1969), indicating good preservation of the topological relationships among patterns (Fig. S2).

Because the SOM is designed to preserve the topological structure of the original high-dimensional input space in a two-dimensional lattice, nodes that are adjacent in the SOM represent more similar circulation patterns than those that are farther apart (Vesanto and Alhoniemi, 2000). This similarity can be quantified using the unified distance matrix (U-matrix), which measures the average distance between the codebook vector of a given node and those of its nearest neighbors (Ultsch and Siemon, 1990). To enhance the objectivity and robustness of pattern classification, a two-step clustering strategy is adopted. First, the SOM organizes the atmospheric states according to their topological similarity. Second, the resulting SOM nodes are further grouped into a smaller number of circulation regimes based on the U-matrix structure using K-means clustering. A known

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

limitation of K-means is the need to predefine the number of clusters, K. To address this issue, K is determined objectively by examining the spatial distribution of U-matrix values. Potential clusters are identified around local minima of the U-matrix, while cluster boundaries correspond to nodes with relatively large U-matrix values separating adjacent minima. Fig. S3 shows the U-matrix of the master SOM, with values normalized to the unit interval. Based on the number, spatial separation, and surrounding high-distance boundaries of the local minima, the SOM lattice is partitioned into six distinct clusters (K = 6), representing six characteristic synoptic-scale circulation regimes during the April–June period. Four clusters correspond to pronounced local minima located in the top-left, bottom-left, bottom-middle, and bottom-right regions of the lattice. In addition, elevated U-matrix values in the top-middle and top-right portions of the SOM indicate circulation patterns that are markedly different from those elsewhere, supporting their identification as two additional, independent regimes. Further methodological details regarding the combined use of SOMs and U-matrix-guided clustering can be found in Nishiyama et al. (2007) and Nguyen-Le et al. (2017, 2019).

2.2.1 Prediction

In the prediction phase, the circulation clustering results are applied to prognostic atmospheric fields from SEAS5. Pentad-mean forecasts of Z850, U850, and V850 fields from pentad 19 (1–5 April) to pentad 36 (25–29 June) for the period 2017–2024 are extracted over the same spatial domain used in the SOM training. Forecasts initialized from December of the preceding year through April, corresponding to lead times of approximately five to one months, are analyzed to assess lead-time dependence. The forecast fields are normalized using the means and standard deviations derived from the ERA5 training period and projected onto the same PCA space to construct a forecast vector $\mathbf{p} = [p_1, p_2, \ldots, p_d]$, where $d = 144$. The Euclidean distance between $\mathbf{p}$ and the centroid of each synoptic-scale circulation cluster (C1–C6) is then computed, and each forecast pentad is assigned to the cluster with the minimum distance. The SCSSM onset date is subsequently identified using the synoptic clustering-based definition introduced in Section 3.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

2.3 Measures of prediction skill

Following Chevuturi et al. (2019, 2021), the prediction skill of the SCSSM OD is evaluated separately for the dependent and independent periods by comparing onset dates diagnosed from SEAS5 hindcasts (1981–2016) and forecasts (2017–2024) with those derived from ERA5 reanalysis. OD are calculated for each individual ensemble member as well as for the ensemble mean, allowing assessment of both deterministic and probabilistic forecast skill. Skill is further evaluated for different initialization months to examine lead-time dependence.

2.3.1 Deterministic prediction skill

The ability of SEAS5 to reproduce the observed interannual variability of the SCSSM OD is quantified using the Pearson correlation coefficient (r) between predicted and observed onset dates. The statistical significance of the correlation is assessed using a two-tailed Student's t test, with $p < 0.05$ indicating significance.

Because correlation alone does not fully characterize predictability, potential predictability is further evaluated using the ratio of predictable component (RPC) (Eade et al., 2014), defined as

$$RPC = \frac{r}{\sqrt{\sigma_{sig}^2/\sigma_{tot}^2}} \quad (1)$$

where r is the correlation between the ensemble-mean prediction and observations, σ_{sig}^2 is the variance of the ensemble mean representing the predictable signal, and σ_{tot}^2 is the total variance across all ensemble members. An RPC value close to unity indicates that the forecast system effectively extracts the available predictable signal, whereas values substantially below unity suggest underconfident or overly dispersed ensemble forecasts.

2.3.2 Categorical skill of ensemble-mean forecasts

To evaluate the skill of SEAS5 in predicting the relative timing of SCSSM onset, OD are classified as early, normal, or late based on the forecast–observation difference. A forecast is

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

classified as early if the predicted OD is more than one pentad earlier than observed (< −1 pentad), normal if the difference lies within ±1 pentad, and late if the predicted OD is more than one pentad later than observed (> 1 pentad). This categorical framework allows assessment of forecast skill in capturing relative onset timing, even when the exact OD is not reproduced.

For deterministic forecasts based on ensemble-mean OD, categorical prediction skill is quantified using Accuracy (ACC) and the Heidke Skill Score (HSS) (WCRP, 2015). ACC measures the fraction of correct categorical forecasts relative to the total number of forecasts and is defined as

$$ACC = \frac{1}{N}\sum_{i=1}^{C} n(F_i, O_i) \quad (2)$$

where $C(F_i, O_i)$ equals 1 if the forecast category F_i matches the observed category O_i (early–normal–late), and 0 otherwise, and N is the total number of forecasts. ACC ranges from 0 (no skill) to 1 (perfect accuracy).

The HSS evaluates forecast accuracy relative to random chance and is defined as

$$HSS = \frac{\frac{1}{N}\sum_{i=1}^{C} n(F_i,O_i) - \frac{1}{N^2}\sum_{i=1}^{C} n(F_i)n(O_i)}{1-\frac{1}{N^2}\sum_{i=1}^{C} n(F_i)n(O_i)} \quad (3)$$

where $C(F_i)$ and $n(O_i)$ denote the marginal frequencies of forecast and observed categories, respectively. HSS ranges from negative values (worse than random) to 1 (perfect skill), with 0 indicating no skill relative to climatological chance.

2.3.3 Probabilistic skill of ensemble forecasts

Probabilistic prediction skill is evaluated using the full ensemble information. For each forecast, probabilities for the three onset categories (early/normal/late) are defined as the fraction of ensemble members predicting each category. These probabilistic forecasts are verified against observations using the Brier Skill Score (BSS) and the Ranked Probability Skill Score (RPSS), which assess forecast reliability, resolution, and discrimination relative to a climatological reference forecast. The discrete BSS (dBSS) for each category is defined as

$$dBSS = 1 - \frac{BS_c}{BS_{clim}+D} \quad (4)$$

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

where BS_c is the Brier Score for a given category,

$$BS_c = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2 \quad (5)$$

where F_i denoting the forecast probability, O_i the observed category indicator (1 for the observed category and 0 otherwise), and N the number of forecasts. The reference score BS_{clim} is computed assuming equal climatological probabilities (1/3). The dBSS ranges from negative values (worse than climatology) to 1 (perfect skill), with 0 indicating no skill.

The discrete RPSS (dRPSS) evaluates cumulative probability errors across ordered onset categories (early–normal–late) and is defined as

$$dRPSS = 1 - \frac{RPS}{RPS_{clim}+D} \quad (6)$$

where the Ranked Probability Score (RPS) is given by

$$RPS = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{(C-1)}\left(\sum_{j=1}^{C}(F_j - O_j)^2\right)\right] \quad (7)$$

with $C = 3$ denoting the number of ordered categories. The reference score RPS_{clim} is calculated using climatological probabilities of 1/3 for each category. The dRPSS similarly ranges from negative values to 1, with 0 indicating no skill relative to climatology, and summarizes forecast errors arising from systematic biases while accounting for discrimination and resolution.

To account for finite ensemble size, a discrete correction term D is applied to both dBSS and dRPSS following Weigel et al. (2007),

$$D = \frac{1}{M}\sum_{i=1}^{K}\left[p_i\left(1 - p_i - 2 \times \sum_{j=i+1}^{K} p_j\right)\right] \quad (8)$$

where M is the ensemble size and p_k is the climatological forecast probability for category k. The correction term is averaged across all years to account for variations in ensemble size and becomes negligible for large ensembles. This correction is particularly important for small ensembles, where limited probability spread can otherwise lead to artificially inflated skill estimates.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

3. Synoptic clustering–based definition of the South China Sea summer monsoon onset

3.1 Definition of SCSSM onset based on synoptic circulation regimes

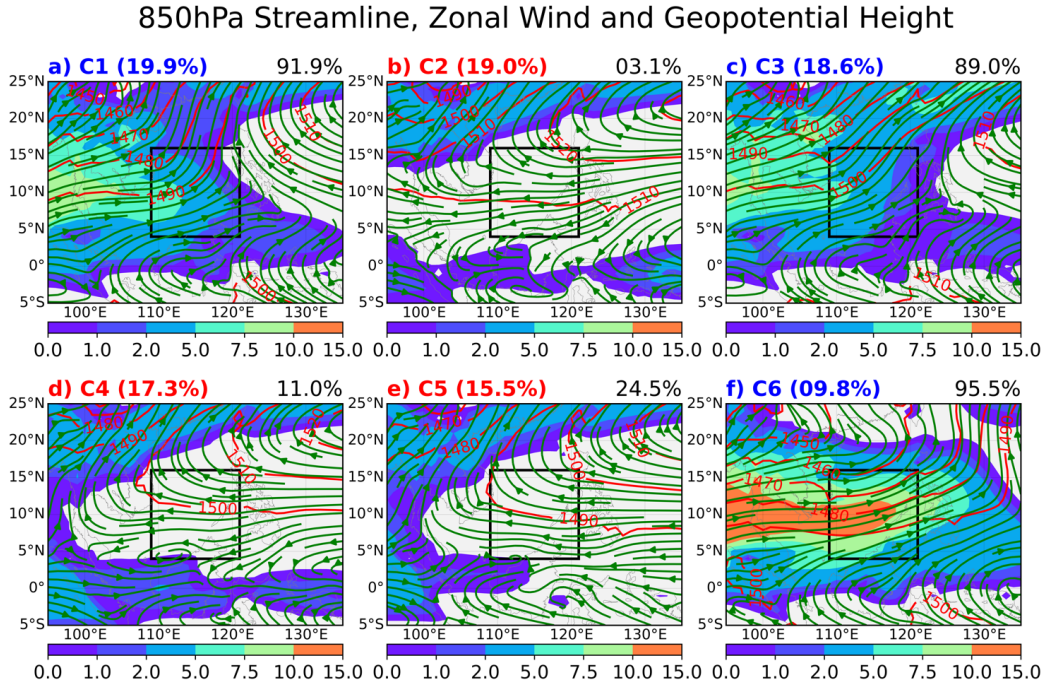


Figure 1: Synoptic circulation regimes associated with the SCSSM identified from SOM clustering. Shown are composites of 850-hPa winds (vectors), zonal wind (shades) and geopotential height (contours) for six clusters (C1–C6). The black box denotes the SCS region used for onset definition based on the U850 index of Wang et al. (2004). Cluster labels and their relative frequencies of occurrence (%) are shown in each panel, while numbers in the upper-right corner indicate the percentage of occurrences after the SCSSM OD defined by Wang et al. (2004). Clusters (a) C1, (c) C3, and (f) C6 correspond to monsoon regimes, whereas (b) C2, (d) C4, and (e) C5 represent pre-monsoon regimes.

Figure 1 presents the six synoptic circulation regimes (C1–C6) identified from the combination of SOM and K-means clustering, illustrated by composites of 850-hPa winds and geopotential height. Together, these clusters represent the dominant low-level circulation regimes spanning the seasonal transition from pre-monsoon to monsoon conditions. Specifically, clusters C2, C4, and C5 are representative of pre-monsoon regimes. These patterns are characterized by easterly or northeasterly low-level flow over the SCS region, associated with the westward extension or

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

persistence of the WNPSH. Geopotential height contours are largely zonal and tightly packed north of the SCS, indicating weak low-level convergence and unfavorable conditions for sustained deep convection over the basin. Consistent with this interpretation, these clusters occur almost exclusively before the SCSSM onset defined by the U850 index of Wang et al. (2004), as reflected by their low post-onset occurrence frequencies (3.1% for C2, 11.0% for C4, and 24.0% for C5).

In contrast, clusters C1, C3, and C6 correspond to monsoon regimes, characterized by the establishment of low-level westerlies over the SCS and enhanced cyclonic curvature of the flow. These regimes are associated with an eastward retreat and weakening of the WNPSH, strengthened geopotential height gradients across the SCS, and enhanced low-level convergence conducive to deep convection. Among the monsoon regimes, C6 represents the most mature and canonical monsoon state, characterized by strong and spatially coherent southwesterly flow extending across the SCS into the WNP and a pronounced eastward retreat of the WNPSH. This regime occurs almost exclusively after the onset defined by Wang et al. (2004), with a post-onset occurrence frequency of 95.5%, indicating a fully established monsoon circulation. C3 also corresponds to a mature monsoon regime, exhibiting stronger low-level westerlies over the SCS and a further eastward withdrawal of the WNPSH compared to C1, reflecting a more developed monsoon structure. While C1 represents a relatively weaker or early-stage monsoon circulation, the high post-onset occurrence frequencies of C3 (89.0%) and C1 (91.9%) confirm that all three regimes are robustly associated with post-onset monsoon conditions and collectively capture variability in monsoon strength and maturity.

The clear dynamical separation between pre-monsoon and monsoon circulation patterns, together with their contrasting post-onset occurrence frequencies, demonstrates that the combination of SOM and K-means clustering effectively captures the large-scale circulation transition underlying SCSSM onset. Importantly, the presence of multiple monsoon regimes indicates that SCSSM onset is not associated with a single circulation pattern, but rather with a regime transition toward persistent low-level westerlies and enhanced convergence marked by the eastward retreat of the WNPSH over

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

the SCS. Based on this regime framework, the OD of the SCSSM can be defined objectively using synoptic circulation transitions rather than a single-variable threshold.

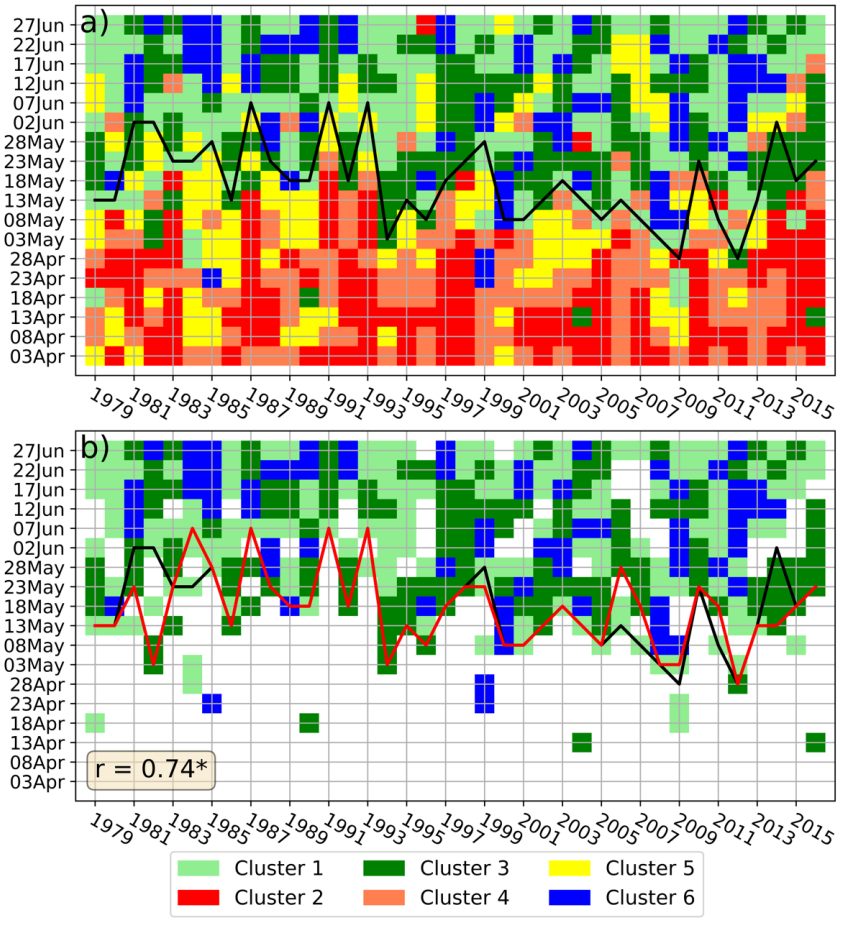
The OD of the SCSSM is defined as the first pentad from pentad 24 (26–30 April onward) that satisfies the following three criteria: ***a) Immediate regime transition:*** *The synoptic circulation in the onset pentad and the subsequent pentad must both belong to the monsoon regimes (C1, C3, or C6);* ***b) Persistence criterion:*** *Within the subsequent four pentads (including the onset pentad), the synoptic circulation must belong to the monsoon regimes in at least three pentads; and* ***c) Mature monsoon condition:*** *Within the same four-pentad window (including the onset pentad), cluster C3 or C6 must occur at least once.*

The first criterion ensures that the identified onset corresponds to a genuine transition into monsoon-type circulation rather than a transient synoptic fluctuation, thereby excluding "bogus" onsets associated with brief westerly intrusions (Wang et al., 2004). The second criterion enforces circulation persistence, filtering out short-lived or weak monsoon-like disturbances commonly observed during the pre-monsoon period. The third criterion further requires the emergence of C3 and C6, which represents a more developed monsoon circulation state, thereby ensuring that the onset is associated with a sufficiently mature and dynamically robust monsoon configuration. Together, these criteria explicitly link SCSSM onset to a sustained regime transition characterized by persistent low-level westerlies over the SCS and the retreat of the WNPSH, providing a dynamically grounded alternative to traditional single-index onset definitions.

Figure 2 compares the SCSSM onset dates identified using the synoptic clustering–based definition (red line) with those derived from the U850 index of Wang et al. (2004) (black line). The two onset definitions show strong agreement, with a statistically significant correlation of 0.74 ($p < 0.01$), indicating that the new definition captures the dominant interannual variability of SCSSM onset timing identified by the conventional circulation-based index. This overall consistency demonstrates that the synoptic clustering–based framework remains aligned with established onset metrics while being explicitly grounded in large-scale circulation regimes. Despite this overall

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

coherence, differences occur in individual years. The clustering-based onset is sometimes identified later than the Wang et al. (2004) onset when brief or weak westerlies are not followed by sustained monsoon circulation, reflecting the imposed persistence and maturity criteria. In contrast, earlier onset identification occurs in years with an early and coherent transition to monsoon-type circulation preceding the U850 threshold. These differences highlight the sensitivity of index-based definitions to short-lived zonal wind reversals and underscore the advantage of a regime-based framework that explicitly accounts for circulation persistence. Overall, the synoptic clustering–based definition preserves the large-scale timing of SCSSM onset while providing a more dynamically constrained and physically interpretable onset identification, supporting its application to seasonal prediction.



https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

Figure 2: Interannual evolution of synoptic circulation regimes and SCSSM OD. Colored shading indicates the dominant synoptic circulation regime (C1–C6) for each pentad from early April to late June for individual years during 1979–2016. (a) OD identified using U850 index of Wang et al. (2004) (red line). (b) Comparison between OD identified by the synoptic clustering–based definition (red line) and the U850-based definition of Wang et al. (2004) (black line). The correlation between the two onset time series is statistically significant ($r = 0.74$, $p < 0.01$). Only monsoon regimes (C1, C2, and C6) are shown in (b).

3.2 Climatology and interannual variability of the SCSSM onset

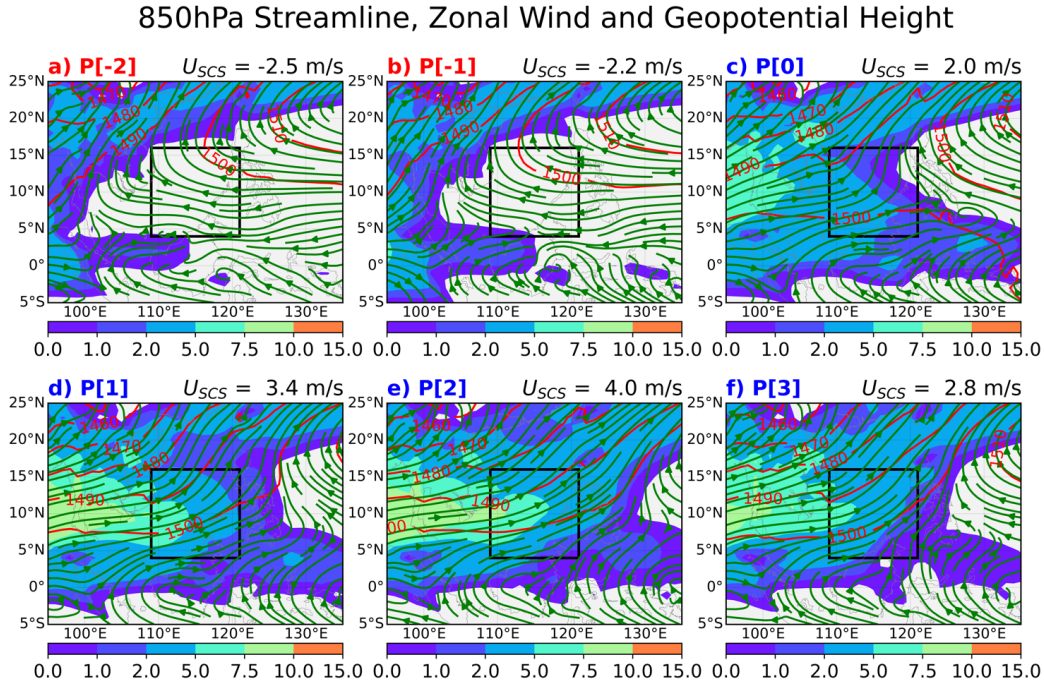


Figure 3: Climatological evolution of 850-hPa winds (vectors), zonal wind (shades) and geopotential height (contours) from two pentads before to three pentads after the SCSSM onset, based on the synoptic clustering–based definition. Panels show composites for P[−2] and P[−1] (pre-onset), P[0] (onset), and P[1]–P[3] (post-onset). The black box denotes the SCS region used to compute the area-averaged zonal wind (U_{SCS}), whose values (m s^{-1}) are indicated in the upper-right corner of each panel.

Figure 3 illustrates the climatological evolution of low-level circulation from two pentads prior to SCSSM onset to three pentads after onset, based on the synoptic clustering–based definition. The composites show 850-hPa winds and geopotential height, with the black box indicating the South China Sea (SCS) region used to calculate area-averaged zonal wind (U_{SCS}). During the pre-onset

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

phase P[−2] and P[−1] (Figs. 3a–b), the SCS is dominated by easterly to northeasterly low-level flow, reflected by negative U_{SCS} values (−2.5 and −2.2 m s^{-1}, respectively). This circulation is associated with a westward-extending WNPSH, as indicated by the orientation and packing of geopotential height contours north of the SCS. Low-level convergence over the basin is weak, consistent with suppressed convective conditions characteristic of the pre-monsoon regime.

At the onset pentad P[0] (Fig. 3c), an abrupt circulation transition occurs. Low-level winds over the SCS reverse to westerly, with U_{SCS} becoming positive (2.0 m s^{-1}), marking the establishment of monsoon flow. This transition is accompanied by an eastward retreat of the WNPSH and enhanced meridional geopotential height gradients across the SCS, indicating strengthened low-level convergence favorable for deep convection.

In the post-onset period P[1]–P[3] (Figs. 3d–f), westerly flow intensifies and becomes more spatially coherent across the SCS, with U_{SCS} increasing to 3.4–4.0 m s^{-1} before slightly weakening by P[3]. The circulation exhibits a well-developed monsoon structure, characterized by persistent southwesterlies extending from the equatorial Indian Ocean into the SCS and WNP. The sustained positive U_{SCS} and stable geopotential height configuration demonstrate that the onset identified by the clustering-based definition corresponds to a robust and persistent transition into the summer monsoon regime rather than a transient wind reversal.

The climatological evolution shown in Figure 3 is consistent with the combined influence of slowly varying boundary forcing and subseasonal atmospheric variability on SCSSM onset. On seasonal timescales, the transition from pre-onset easterlies to sustained post-onset westerlies reflects the gradual weakening and eastward retreat of the WNPSH, a process that is strongly modulated by ENSO through the Indo-Pacific Ocean Capacitor mechanism (Xie et al., 2009, 2016). This ENSO-related background modulation is further illustrated in Figure 4, which shows composite anomalies of SST and 850-hPa winds for early and late SCSSM onset years during the preceding winter (December–January–February) and spring (March–April–May).

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

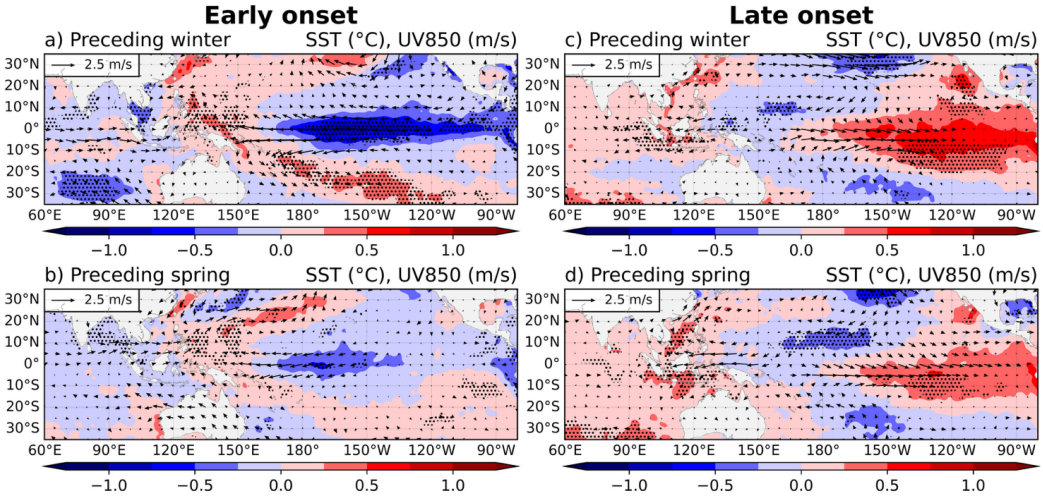


Figure 4: Composite anomalies of SST (shading, °C) and 850-hPa winds (vectors, m/s) associated with early and late SCSSM onset years. (a, b) Early-onset composites for the preceding winter and spring, respectively. (c, d) Late-onset composites for the preceding winter (December–January–February) and spring (March–April–May). Stippling denotes regions where SST anomalies are statistically significant at the 95% confidence level. Wind vectors indicate anomalous low-level circulation, with the reference vector shown in the upper-left corner of each panel.

For early onset years (Figs. 4a–b), the preceding winter is characterized by cold SST anomalies over the central–eastern equatorial Pacific, warm anomalies in the WNP and cold anomalies in the Indian Ocean, consistent with La Niña–like conditions. Associated low-level wind anomalies exhibit weakened easterlies over the western Pacific and enhanced westerly tendencies extending toward the SCS. During the preceding spring, these anomalies persist and intensify, promoting a background circulation favorable for the early establishment of low-level westerlies over the SCS and facilitating an earlier monsoon transition.

In contrast, late onset years (Figs. 4c–d) are associated with warm and cold SST anomalies over the central–eastern equatorial Pacific and the WNP, respectively, during the preceding winter, indicative of El Niño conditions, together with warming in the tropical Indian Ocean. The corresponding low-level wind anomalies show strengthened easterlies over the western Pacific, reinforcing the WNPSH and suppressing westerly development over the SCS. These conditions

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

persist into spring, delaying the seasonal weakening of the WNPSH and inhibiting the establishment of monsoon-type circulation over the SCS, thereby favoring a later onset.

Superimposed on this ENSO-modulated seasonal preconditioning, the abrupt reversal of low-level winds at the onset pentad and the rapid strengthening of westerlies in the subsequent pentads (Fig. 3) highlight the critical role of subseasonal disturbances in triggering SCSSM onset once the large-scale background state becomes favorable. Intraseasonal oscillations, particularly those associated with the Madden–Julian Oscillation, have been shown to modulate low-level westerlies, convection, and moisture convergence over the SCS, providing the immediate dynamical trigger for monsoon onset (Wang et al., 2009). The coexistence of gradual ENSO-related background evolution and abrupt synoptic-scale transitions underscores that SCSSM onset emerges from the interaction between low-frequency boundary forcing and high-frequency atmospheric variability. It suggests that the synoptic clustering–based onset definition naturally captures this multi-timescale behavior by linking onset to both a persistent circulation regime transition and the large-scale background state. This makes the proposed definition suitable for applications to seasonal prediction.

4. Application to seasonal prediction of SCSSM onset

This section evaluates the applicability of the synoptic clustering–based SCSSM onset definition for seasonal prediction using SEAS5 forecasts. Prediction skill is assessed separately for a dependent training period (1981–2016) and an independent forecast period (2017–2024) to examine both in-sample performance and out-of-sample robustness. Results obtained using the clustering-based definition of the present study (NL26) are systematically compared with those based on the conventional U850 index of Wang et al. (2004; W04).

4.1 Prediction skill during the training (hindcast) period (1981–2016)

First, performance of the ECMWF SEAS5 seasonal prediction system in forecasting SCSSM OD using both the W04 and NL26 definitions is evaluated during the dependent training period (1981–2016). Forecasts initialized from December (~five-month lead time) to April (one-month lead

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

time) are examined to assess lead-time dependence, and both deterministic and probabilistic skill metrics are considered.

4.1.1 Deterministic prediction skill

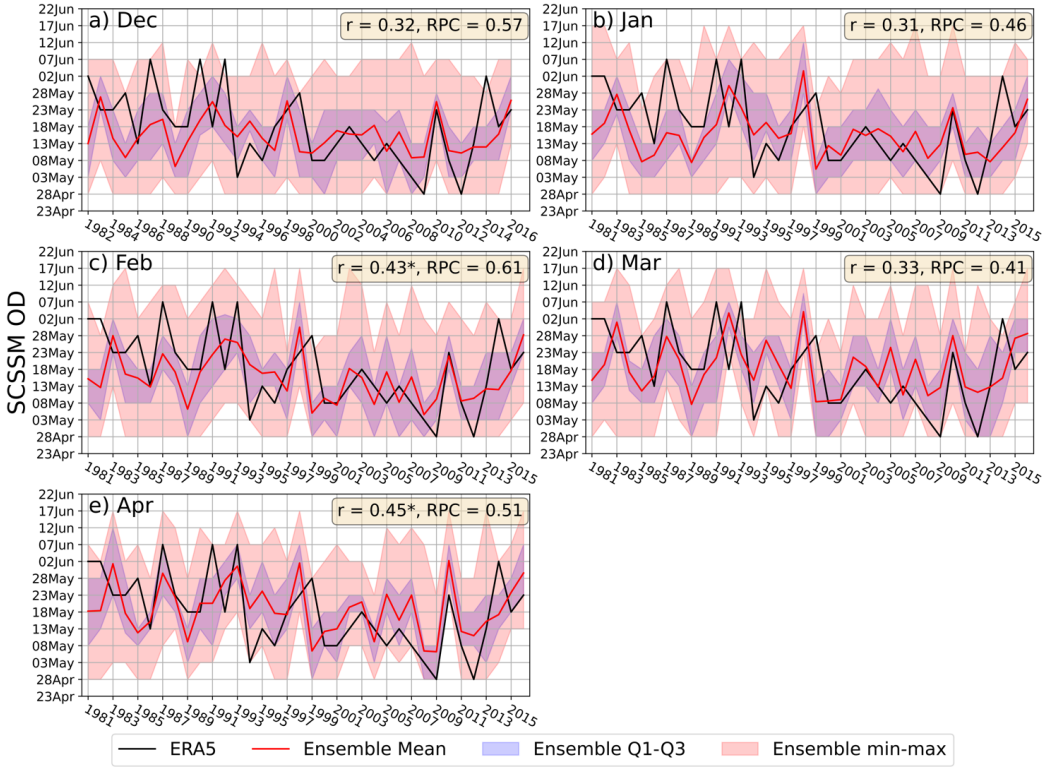
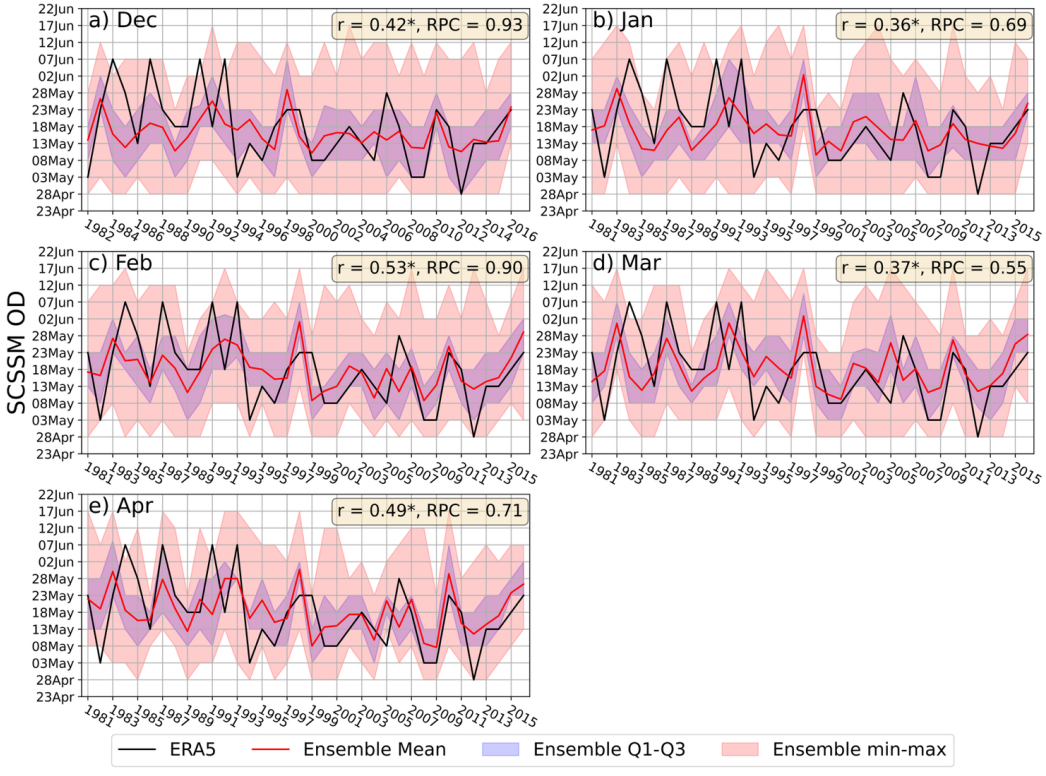


Figure 5: Hindcast performance of the ECMWF SEAS5 system in predicting SCSSM onset during the dependent training period (1981–2016) using the U850-based definition of Wang et al. (2004). Time series show observed onset dates from ERA5 (black) and ensemble-mean predictions (red) for forecasts initialized from December to April. Blue shading denotes the interquartile range (Q1–Q3), while red shading indicates the full ensemble spread. The Pearson correlation coefficient (r) and ratio of predictable component (RPC) are shown for each initialization month. The * denotes r is statistically significant at the 95% confidence level.

Figure 5 shows the deterministic prediction skill of SEAS5 when SCSSM onset is defined using the W04. Correlation coefficients (r) between ensemble-mean predictions and observed onset dates range from 0.31 to 0.45, with modest improvements toward shorter lead times , and statistically

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

447 significant correlations obtained only for forecasts initialized in February and April. The ratio of
448 predictable component (RPC) generally remains below 0.65, indicating that a substantial fraction of
449 ensemble spread is not associated with the predictable signal. This behavior suggests that OD defined
450 purely by a local zonal-wind threshold are sensitive to short-lived wind reversals and synoptic noise,
451 which limits deterministic predictability even during the dependent period.



452

453 **Figure 6:** Same as Fig. 5, but for SCSSM OD defined using the synoptic clustering–based definition (NL26).

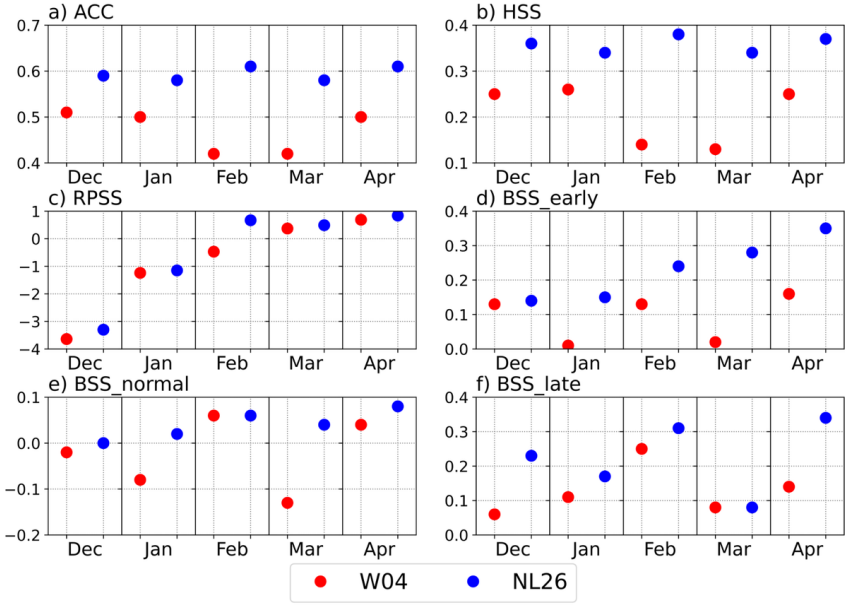
454 In contrast, Figure 6 demonstrates a clear improvement in deterministic forecast skill when the
455 NL26 definition is applied. Correlation coefficients increase to 0.36–0.53 for all initialization
456 months, with statistically significant correlations throughout and maximum skill for February
457 initialization (r = 0.53) and April initialization (r=0.49). Importantly, skill improvements are already
458 evident at long lead times: for December initialization, the correlation increases from 0.32 (W04) to

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

459 0.42 (NL26), indicating enhanced sensitivity to slowly varying boundary forcing when onset is
460 defined in terms of circulation regimes rather than a single-point wind index.

461 Even more striking is the improvement in potential predictability. RPC values increase
462 substantially under the NL26 definition, reaching 0.90–0.93 for December and February
463 initializations and remaining above 0.55 for all start months. These near-unity RPC values indicate
464 that the ensemble mean captures most of the predictable signal and that ensemble dispersion is more
465 consistent with signal amplitude. This result suggests that the clustering-based definition effectively
466 filters out unpredictable synoptic-scale fluctuations and aligns the onset metric more closely with
467 large-scale circulation features that are inherently predictable on seasonal timescales.

468 ***4.1.2 Categorical and probabilistic prediction skill***



469

470 **Figure 7:** Categorical and probabilistic prediction skill of SEAS5 for SCSSM OD during the dependent
471 training period (1981–2016), comparing the U850-based definition of Wang et al. (2004; W04) and the
472 synoptic clustering–based definition (NL26). Shown are (a) Accuracy (ACC), (b) Heidke Skill Score (HSS),
473 (c) Ranked Probability Skill Score (RPSS), and Brier Skill Scores (BSS) for (d) early, (e) normal, and (f) late
474 onset categories, for forecasts initialized from December to April.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

Figure 7 further compares forecast skill between the W04 and NL26 definitions using categorical and probabilistic metrics, including Accuracy (ACC), Heidke Skill Score (HSS), Ranked Probability Skill Score (RPSS), and Brier Skill Scores (BSS) for early, normal, and late onset categories. For deterministic categorical forecasts (Figs. 7a–b), NL26 consistently outperforms W04 across all initialization months. ACC increases from approximately 0.42–0.51 under W04 to 0.58–0.61 under NL26, while HSS improves from values near 0.13–0.26 to 0.34–0.38. These improvements indicate both higher hit rates and substantially greater skill relative to random chance, confirming that the regime-based onset classification yields more robust categorical predictions.

Differences are even more pronounced for probabilistic skill metrics. Under the W04 definition, RPSS values are negative or near zero for December–January initializations, indicating little to no skill relative to climatology (Fig. 7c). In contrast, RPSS becomes positive for all initialization months under NL26, with values approaching unity for February to April initializations, reflecting a marked improvement in the reliability and discrimination of ensemble-based probabilistic forecasts when onset is defined in terms of persistent circulation regimes. Consistent improvements are also evident in category-specific BSS (Figs. 7d–f). For early and late onset categories, NL26 yields systematically higher BSS values across all lead times, indicating enhanced probabilistic resolution for extreme onset years. Skill for the normal category remains modest for both definitions, reflecting the intrinsic difficulty of predicting near-average onset timing; nevertheless, NL26 generally maintains neutral to positive BSS, whereas W04 often exhibits negative skill.

Overall, results from the dependent training period demonstrate that defining SCSSM onset using the NL26 enhances deterministic, categorical, and probabilistic prediction skill relative to the WL04. By explicitly incorporating circulation persistence and regime transitions, the clustering-based definition aligns onset timing more closely with large-scale, predictable circulation features represented in SEAS5, leading to higher correlations, improved ensemble reliability, and more skillful probabilistic forecasts. While these results establish the potential of the proposed definition under in-sample conditions, it is essential to assess whether the skill improvements persist under true

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

out-of-sample prediction. This is examined in the following subsection using independent SEAS5 forecasts for the period 2017–2024.

4.2 Prediction skill during the independent forecast period (2017–2024)

This subsection evaluates the out-of-sample performance of the synoptic clustering–based SCSSM onset definition using independent SEAS5 forecasts for 2017–2024, which were not included in the training of the SOM circulation regimes.

4.2.1 Deterministic prediction skill

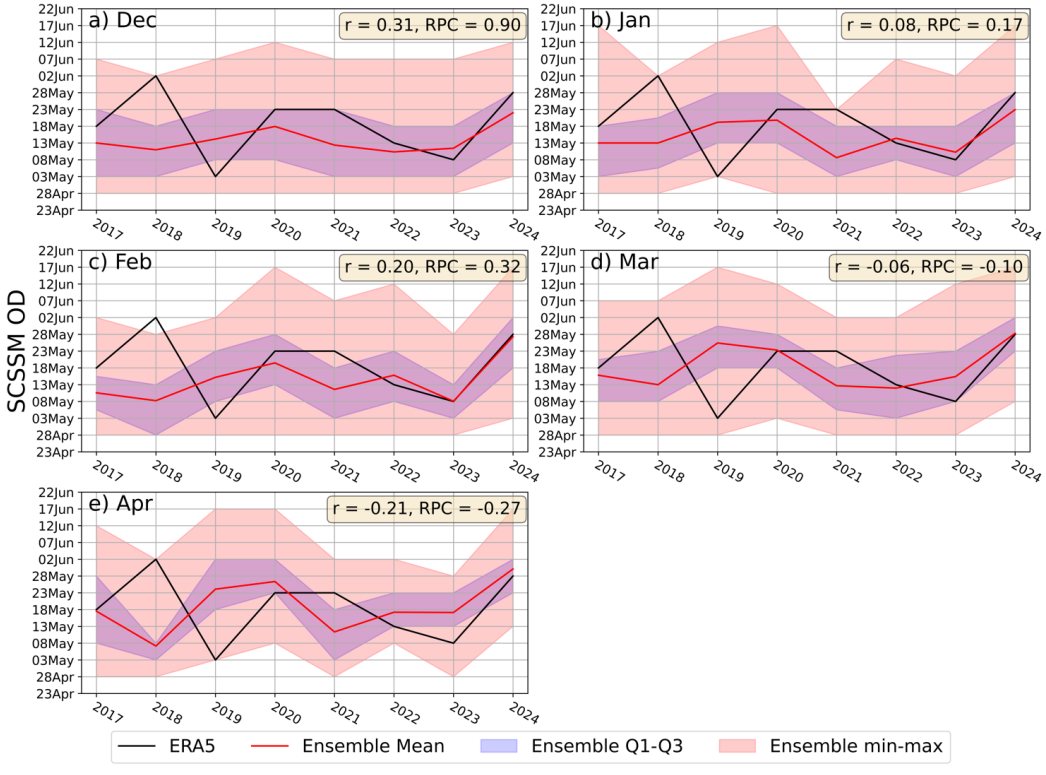


Figure 8: Forecast performance of the ECMWF SEAS5 system in predicting SCSSM onset during the independent forecast period (2017–2024) using the U850-based definition of Wang et al. (2004). Time series show ERA5 onset dates (black) and ensemble-mean forecasts (red), with blue shading denotes the interquartile range (Q1–Q3) and red shading indicates the full ensemble spread. The Pearson correlation coefficient (r) and ratio of predictable component (RPC) are shown for each initialization month. The * denotes r is statistically significant at the 95% confidence level.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

As expected for an independent forecast period with a limited sample size, overall deterministic prediction skill is reduced compared to the dependent period (Fig. 8). When SCSSM OD is defined using W04, correlations between SEAS5 ensemble-mean and ERA5 are generally weak and often not statistically significant, particularly for forecasts initialized from January onward. In several cases, both the correlation coefficient and the RPC are close to zero or negative, indicating little usable predictability. Meanwhile, the NL26 yields consistently higher r and markedly improved RPC values across most initialization months (Fig. 9). Forecasts initialized from January to April show the largest gains, with r exceeding 0.5 in several months and RPC values approaching or exceeding unity, indicating that a substantial fraction of the predictable signal is captured by the ensemble mean. These improvements suggest that the clustering-based definition better aligns forecast variability with predictable large-scale circulation features, even under out-of-sample conditions.

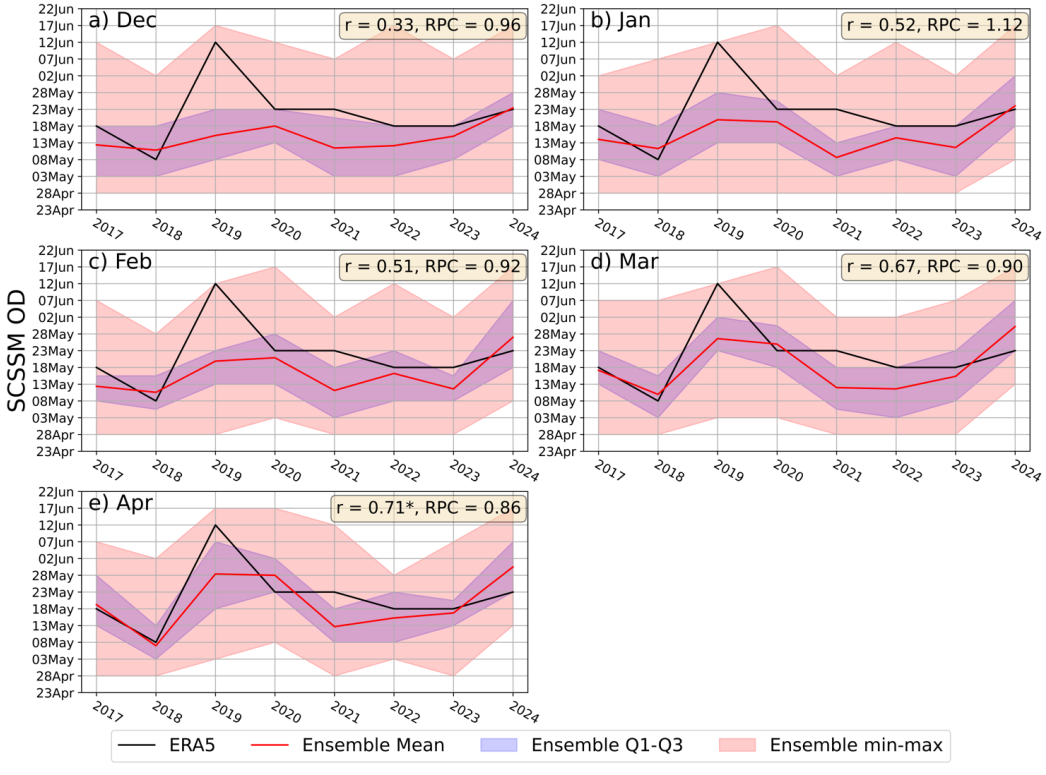


Figure 9: Same as Fig. 8, but for SCSSM OD defined using the synoptic clustering–based definition (NL26).

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

4.2.2 Categorical and probabilistic prediction skill

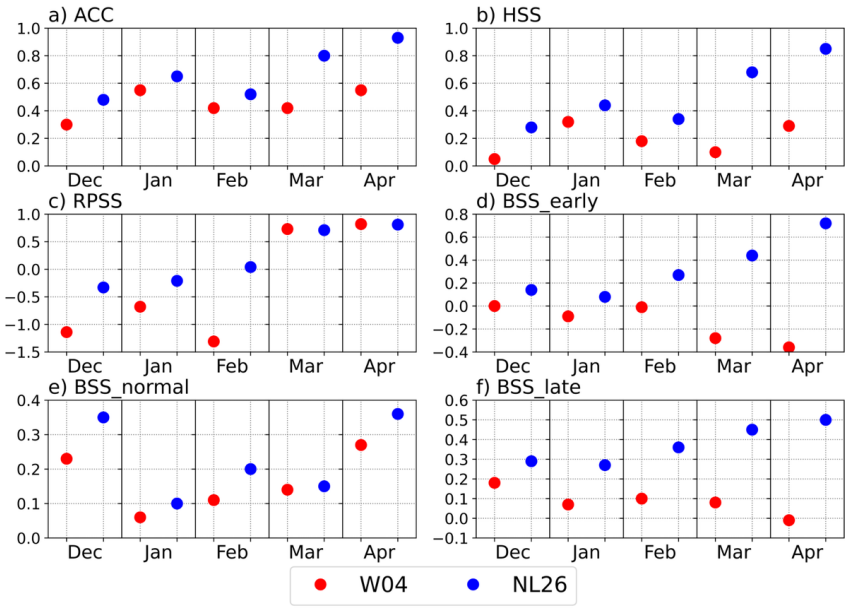


Figure 10: Categorical and probabilistic prediction skill of SEAS5 for SCSSM onset during the independent forecast period (2017–2024), comparing the U850-based definition of Wang et al. (2004; W04) and the synoptic clustering–based definition (NL26). Shown are (a) Accuracy (ACC), (b) Heidke Skill Score (HSS), (c) Ranked Probability Skill Score (RPSS), and Brier Skill Scores (BSS) for (d) early, (e) normal, and (f) late onset categories, for forecasts initialized from December to April.

Categorical skill metrics further highlight the advantages of the synoptic clustering–based definition during the independent forecast period (Fig. 10). For deterministic ensemble-mean forecasts, both ACC and HSS obtained with NL26 consistently exceed those derived from W04 across all initialization months, with particularly pronounced improvements for forecasts initialized in March and April. HSS values under W04 are generally small and occasionally near zero, indicating limited skill relative to climatology, whereas NL26 yields positive and substantially larger HSS values, demonstrating improved discrimination among early, normal, and late onset categories.

Differences are even more pronounced for probabilistic prediction skill. RPSS values based on W04 are frequently negative or close to zero, especially for early-season initializations, indicating

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

that probabilistic forecasts often perform no better than climatology. In contrast, RPSS values under NL26 are positive for most initialization months and increase systematically toward shorter lead times, reflecting enhanced reliability and resolution of ensemble-based probabilities. Category-specific BSS further confirm that the clustering-based definition improves probabilistic skill across all onset categories. The largest gains occur for early and late onset categories, which are most strongly linked to persistent circulation regime transitions and therefore benefit directly from the regime-based onset criteria. Skill for the "normal" category is also generally higher under NL26, although improvements are more modest, consistent with the broader spread and inherently lower predictability of near-climatological onset timing.

Overall, results from the independent forecast period indicate that the synoptic clustering–based definition of SCSSM onset consistently outperforms the conventional U850-based index across deterministic, categorical, and probabilistic metrics, despite the expected reduction in absolute skill under true out-of-sample conditions. The clustering-based definition yields higher correlations, improved ensemble reliability, and more skillful categorical and probabilistic forecasts, demonstrating its robustness and transferability beyond the training period.

5. Discussions and Conclusions

5.1 Discussions

This study demonstrates that defining SCSSM OD in terms of persistent synoptic circulation regimes substantially improves seasonal prediction skill compared with traditional index-based definitions. The superior performance of the NL26 approach can be understood by considering both the physical processes governing monsoon onset and the intrinsic limits of seasonal predictability. Conventional onset definitions, such as the WL04, rely on local threshold exceedance of zonal winds and are therefore highly sensitive to short-lived wind reversals associated with transient synoptic disturbances. Such events may satisfy index criteria without representing a genuine transition into the summer monsoon circulation, introducing noise into diagnosed OD and degrading forecast

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

verification. This sensitivity is reflected in the relatively modest deterministic correlations and low RPC obtained using the W04 definition, particularly at longer lead times. In contrast, the synoptic clustering–based definition explicitly links SCSSM onset to a regime transition in the large-scale circulation, characterized by the establishment and persistence of low-level westerlies over SCS and the eastward retreat of the WNPSH. By requiring both continuity and persistence of monsoon-type regimes, the definition filters out spurious or short-lived events and emphasizes circulation structures that evolve on longer timescales and are therefore more predictable. This physically grounded framing explains the systematic improvements in deterministic, categorical, and probabilistic skill metrics obtained with the clustering-based definition.

The results further underscore the multi-timescale nature of SCSSM onset. On interannual timescales, slowly varying boundary forcing, most notably ENSO-related SST anomalies, modulates the large-scale Indo–Pacific circulation and preconditions the timing of low-level westerly establishment over the SCS. This source of predictability is effectively captured by SEAS5 when onset is defined in terms of circulation regimes. Superimposed on this background evolution, subseasonal disturbances, including intraseasonal oscillations, provide the immediate dynamical trigger for onset once the large-scale environment becomes favorable. The clustering-based definition naturally accommodates this interaction by identifying onset as a sustained regime shift rather than a single instantaneous threshold crossing.

5.2 Conclusions

This study introduces a synoptic clustering–based definition of SCSSM onset that conceptualizes onset as a persistent transition between large-scale circulation regimes, rather than a local or instantaneous wind threshold. By identifying onset through sustained occupation of monsoon-type circulation patterns derived from self-organizing maps and clustering of low-level fields, the proposed definition reduces sensitivity to short-lived synoptic fluctuations and more faithfully represents the dynamical evolution of the monsoon system. Application to ECMWF SEAS5 forecasts

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

demonstrates that the clustering-based definition yields systematic and robust improvements in seasonal prediction skill relative to the conventional U850-based index of Wang et al. (2004). Skill gains are evident across deterministic, potential predictability, categorical, and probabilistic metrics during the dependent training period (1981–2016) and persist during the independent forecast period (2017–2024), indicating that the approach generalizes well beyond the training sample.

The improved forecast performance is physically consistent with the dominant controls on SCSSM onset variability, which arise from the interaction between slowly varying boundary forcing and higher-frequency atmospheric variability. By emphasizing circulation persistence and structural maturity, the regime-based framework more effectively isolates the predictable component of onset variability and improves ensemble reliability, particularly for early and late onset events. These results indicate that predictability-oriented, circulation-regime–based definitions offer a physically meaningful and practically useful alternative to traditional index-based approaches for diagnosing and forecasting monsoon onset. Although this study focuses on the SCSSM, the methodology is readily applicable to other monsoon systems and climate transition processes in which circulation persistence plays a key role in predictability. Future work should also explore the integration of advanced pattern-recognition approaches, including deep learning, to further enhance regime identification and forecast skill.

Supplementary Information The manuscript contains supplementary figures.

Acknowledgments The author gratefully acknowledges Dr. Le Duc for valuable guidance and assistance with the implementation of the SOM and K-means clustering techniques.

Fundings This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability All datasets analyzed in this study are publicly available from the following sources: ERA5 (ECMWF; https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5);

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

HadISST (UK Met Office Hadley Centre; https://www.metoffice.gov.uk/hadobs/hadisst); SEAS5 (ECMWF; https://cds.climate.copernicus.eu/datasets).

Code availability The code used in this study is available from the corresponding author upon reasonable request.

Author contributions The author conceived the study, performed the analysis, and wrote the manuscript.

Declarations

Competing interests The author has no conflicts of interest to declare.

Ethical approval Ethical approval is not applicable to this article.

References

Attada R, Ehsan MA, Pillai PA (2022) Evaluation of Potential Predictability of Indian Summer Monsoon Rainfall in ECMWF's Fifth-Generation Seasonal Forecast System (SEAS5). Pure Appl Geophys 179:4639–4655. https://doi.org/10.1007/s00024-022-03184-9

Bombardi, R. J., Pegion, K. V., Kinter, J. L., Cash, B. A., and Adams, J. M.: Sub-seasonal Predictability of the Onset and Demise of the Rainy Season over Monsoonal Regions, Front. Earth Sci., 5, https://doi.org/10.3389/feart.2017.00014, 2017.

Bombardi, R. J., Moron, V., and Goodnight, J. S.: Detection, variability, and predictability of monsoon onset and withdrawal dates: A review, Intl Journal of Climatology, 40, 641–667, https://doi.org/10.1002/joc.6264, 2020.

Borah, N., Sahai, A. K., Chattopadhyay, R., Joseph, S., Abhilash, S., and Goswami, B. N.: A self-organizing map–based ensemble forecast system for extended range prediction of active/break cycles

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

of Indian summer monsoon, JGR Atmospheres, 118, 9022–9034, https://doi.org/10.1002/jgrd.50688, 2013.

Bui-Minh, T., Doan, Q.-V., Nguyen, K.-C., Phan-Van, T., Trinh-Tuan, L., Cong, T., and Trinh-Minh, N.: Determining the onset of summer rainfall over Vietnam using self-organizing maps, Clim Dyn, 62, 9189–9206, https://doi.org/10.1007/s00382-024-07385-x, 2024.

Chen, T.-C., Tsay, J.-D., Matsumoto, J., and Alpert, J.: Impact of the Summer Monsoon Westerlies on the South China Sea Tropical Cyclone Genesis in May, Weather and Forecasting, 32, 925–947, https://doi.org/10.1175/WAF-D-16-0189.1, 2017.

Chevuturi, A., Turner, A. G., Woolnough, S. J., Martin, G. M., and MacLachlan, C.: Indian summer monsoon onset forecast skill in the UK Met Office initialized coupled seasonal forecasting system (GloSea5-GC2), Clim Dyn, 52, 6599–6617, https://doi.org/10.1007/s00382-018-4536-1, 2019.

Chevuturi, A., Turner, A. G., Johnson, S., Weisheimer, A., Shonk, J. K. P., Stockdale, T. N., and Senan, R.: Forecast skill of the Indian monsoon and its onset in the ECMWF seasonal forecasting system 5 (SEAS5), Clim Dyn, 56, 2941–2957, https://doi.org/10.1007/s00382-020-05624-5, 2021.

Dai, L., Cheng, T. F., and Lu, M.: Define East Asian Monsoon Annual Cycle via a Self-Organizing Map-Based Approach, Geophysical Research Letters, 48, e2020GL089542, https://doi.org/10.1029/2020GL089542, 2021.

Ding, Y.: SEASONAL MARCH OF THE EAST-ASIAN SUMMER MONSOON, in: World Scientific Series on Asia-Pacific Weather and Climate, vol. 02, WORLD SCIENTIFIC, 3–53, https://doi.org/10.1142/9789812701411_0001, 2004.

Ding, Y., Li, C., and Liu, Y.: Overview of the South China sea monsoon experiment, Adv. Atmos. Sci., 21, 343–360, https://doi.org/10.1007/BF02915563, 2004.

Ding, Y. and Chan, J. C. L.: The East Asian summer monsoon: an overview, Meteorol. Atmos. Phys., 89, 117–142, https://doi.org/10.1007/s00703-005-0125-z, 2005.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

663 Ding, Y., Liu, Y., Song, Y., and Zhang, J.: From MONEX to the global monsoon: A review of
664 monsoon system research, Adv. Atmos. Sci., 32, 10–31, https://doi.org/10.1007/s00376-014-0008-7,
665 2015.

666 Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., and Robinson, N.: Do
667 seasonal-to-decadal climate predictions underestimate the predictability of the real world?,
668 Geophysical Research Letters, 41, 5620–5628, https://doi.org/10.1002/2014GL061146, 2014.

669 Geen, R.: Forecasting South China Sea Monsoon Onset Using Insight From Theory, Geophysical
670 Research Letters, 48, e2020GL091444, https://doi.org/10.1029/2020GL091444, 2021.

671 He, J. and Zhu, Z.: The relation of South China Sea monsoon onset with the subsequent rainfall over
672 the subtropical East Asia, Intl Journal of Climatology, 35, 4547–4556,
673 https://doi.org/10.1002/joc.4305, 2015.

674 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J.,
675 Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G.,
676 Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis,
677 M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan,
678 R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., De
679 Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis,
680 Quart J Royal Meteoro Soc, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

681 Hu, P., Chen, W., Huang, R., and Nath, D.: On the weakening relationship between the South China
682 Sea summer monsoon onset and cross-equatorial flow after the late 1990s, Intl Journal of
683 Climatology, 38, 3202–3208, https://doi.org/10.1002/joc.5472, 2018.

684 Hu, P., Chen, W., Chen, S., Liu, Y., Wang, L., and Huang, R.: The Leading Mode and Factors for
685 Coherent Variations among the Subsystems of Tropical Asian Summer Monsoon Onset, Journal of
686 Climate, 35, 1597–1612, https://doi.org/10.1175/JCLI-D-21-0101.1, 2022a.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

687 Hu, P., Chen, W., Li, Z., Chen, S., Wang, L., and Liu, Y.: Close Linkage of the South China Sea
688 Summer Monsoon Onset and Extreme Rainfall in May over Southeast Asia: Role of the Synoptic-
689 Scale Systems, Journal of Climate, 35, 4347–4362, https://doi.org/10.1175/JCLI-D-21-0740.1,
690 2022b.

691 Huangfu, J., Huang, R., and Chen, W.: Relationship between the South China Sea summer monsoon
692 onset and tropical cyclone genesis over the western North Pacific, Intl Journal of Climatology, 37,
693 5206–5210, https://doi.org/10.1002/joc.5141, 2017.

694 Jiang, X., Wang, Z., and Li, Z.: Signature of the South China Sea summer monsoon onset on spring-
695 to-summer transition of rainfall in the middle and lower reaches of the Yangtze River basin, Clim
696 Dyn, 51, 3785–3796, https://doi.org/10.1007/s00382-018-4110-x, 2018.

697 Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L.,
698 Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H.,
699 and Monge-Sanz, B. M.: SEAS5: the new ECMWF seasonal forecast system, Geoscientific Model
700 Development, 12, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019, 2019.

701 Li, X., Wang, L., Zhong, S., and Liu, L.: Comparative analysis of indices in capturing the onset and
702 withdrawal of the South Asian Summer Monsoon, Environ. Res. Commun., 6, 031007,
703 https://doi.org/10.1088/2515-7620/ad352b, 2024.

704 Liu, B., Zhu, C., Yuan, Y., and Xu, K.: Two Types of Interannual Variability of South China Sea
705 Summer Monsoon Onset Related to the SST Anomalies before and after 1993/94, Journal of
706 Climate, 29, 6957–6971, https://doi.org/10.1175/JCLI-D-16-0065.1, 2016.

707 Martin, G. M., Chevuturi, A., Comer, R. E., Dunstone, N. J., Scaife, A. A., and Zhang, D.:
708 Predictability of South China Sea Summer Monsoon Onset, Adv. Atmos. Sci., 36, 253–260,
709 https://doi.org/10.1007/s00376-018-8100-z, 2019.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

Nguyen-Le, D. and Yamada, T. J.: Using Weather Pattern Recognition to Classify and Predict Summertime Heavy Rainfall Occurrence over the Upper Nan River Basin, Northwestern Thailand, Weather and Forecasting, 34, 345–360, https://doi.org/10.1175/WAF-D-18-0122.1, 2019.

Nguyen-Le, D., Yamada, T. J., and Tran-Anh, D.: Classification and forecast of heavy rainfall in northern Kyushu during Baiu season using weather pattern recognition, Atmospheric Science Letters, 18, 324–329, https://doi.org/10.1002/asl.759, 2017.

Nishiyama, K., Endo, S., Jinno, K., Bertacchi Uvo, C., Olsson, J., and Berndtsson, R.: Identification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a Self-Organizing Map, Atmospheric Research, 83, 185–200, https://doi.org/10.1016/j.atmosres.2005.10.015, 2007.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res., 108, 2002JD002670, https://doi.org/10.1029/2002JD002670, 2003.

Sammon, J. W.: A Nonlinear Mapping for Data Structure Analysis, IEEE Transactions on Computers, C–18, 401–409, https://doi.org/10.1109/T-C.1969.222678, 1969.

Ultsch, A. and Siemon, H. P: Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. in Proceedings of the International Neural Network Conference (INNC-90), Paris, France, July 9–13, 1990 1. Dordrecht, Netherlands (eds. Widrow, B. & Angeniol, B.) 1, 305--308 (Kluwer Academic Press, 1990).

Vesanto, J. and Alhoniemi, E.: Clustering of the self-organizing map, IEEE Transactions on Neural Networks, 11, 586–600, https://doi.org/10.1109/72.846731, 2000.

Wang, B., Huang, F., Wu, Z., Yang, J., Fu, X., and Kikuchi, K.: Multi-scale climate variability of the South China Sea monsoon: A review, Dynamics of Atmospheres and Oceans, 47, 15–37, https://doi.org/10.1016/j.dynatmoce.2008.09.004, 2009.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

734 Wang, B. and LinHo: Rainy Season of the Asian–Pacific Summer Monsoon*, J. Climate, 15, 386–
735 398, https://doi.org/10.1175/1520-0442(2002)015%3C0386:RSOTAP%3E2.0.CO;2, 2002.

736 Wang, B., LinHo, Zhang, Y., and Lu, M.-M.: Definition of South China Sea Monsoon Onset and
737 Commencement of the East Asia Summer Monsoon*, Journal of Climate, 17, 699–710,
738 https://doi.org/10.1175/2932.1, 2004.

739 Wang, B., Ding, Q., Fu, X., Kang, I., Jin, K., Shukla, J., and Doblas-Reyes, F.: Fundamental
740 challenge in simulation and prediction of summer monsoon rainfall, Geophysical Research Letters,
741 32, 2005GL022734, https://doi.org/10.1029/2005GL022734, 2005.

742 Wang, L. and Chen, G.: Relationship between South China Sea summer monsoon onset and
743 landfalling tropical cyclone frequency in China, Intl Journal of Climatology, 38, 3209–3214,
744 https://doi.org/10.1002/joc.5485, 2018.

745 World Climate Research Programme (WCRP): Forecast verification—issues, methods, and FAQ,
746 http://www.cawcr.gov.au/projects/verification (last access: 25 January 2026), 2015.

747 Weigel, A. P., Liniger, M. A., and Appenzeller, C.: The Discrete Brier and Ranked Probability Skill
748 Scores, Monthly Weather Review, 135, 118–124, https://doi.org/10.1175/MWR3280.1, 2007.

749 Xie, S.-P., Hu, K., Hafner, J., Tokinaga, H., Du, Y., Huang, G., and Sampe, T.: Indian Ocean
750 Capacitor Effect on Indo–Western Pacific Climate during the Summer following El Niño, Journal of
751 Climate, 22, 730–747, https://doi.org/10.1175/2008JCLI2544.1, 2009.

752 Xie, S.-P., Kosaka, Y., Du, Y., Hu, K., Chowdary, J. S., and Huang, G.: Indo-western Pacific ocean
753 capacitor and coherent climate anomalies in post-ENSO summer: A review, Adv. Atmos. Sci., 33,
754 411–432, https://doi.org/10.1007/s00376-015-5192-6, 2016.

755 Zhou, W. and Chan, J. C. L.: ENSO and the South China Sea summer monsoon onset, Intl Journal of
756 Climatology, 27, 157–167, https://doi.org/10.1002/joc.1380, 2007.

https://doi.org/10.5194/egusphere-2026-358
Preprint. Discussion started: 25 February 2026

757 Zhu, Z. and Li, T.: Empirical prediction of the onset dates of South China Sea summer monsoon,

758 Clim Dyn, 48, 1633–1645, https://doi.org/10.1007/s00382-016-3164-x, 2017.