Article

A machine learning model that outperforms conventional global subseasonal forecast models

Lei Chen[1,2,10], Xiaohui Zhong [1,10], Hao Li [1,10] ✉, Jie Wu[3,10], Bo Lu [3,4] ✉, Deliang Chen [5], Shang-Ping Xie [6], Libo Wu [7,8,9], Qingchen Chao[3], Chensen Lin [1], Zixin Hu[1] & Yuan Qi[1,2] ✉

Skillful subseasonal forecasts are crucial for various sectors of society but pose a grand scientific challenge. Recently, machine learning-based weather fore-casting models outperform the most successful numerical weather predic-tions generated by the European Centre for Medium-Range Weather Forecasts (ECMWF), but have not yet surpassed conventional models at subseasonal timescales. This paper introduces FuXi Subseasonal-to-Seasonal (FuXi-S2S), a machine learning model that provides global daily mean forecasts up to 42 days, encompassing five upper-air atmospheric variables at 13 pressure levels and 11 surface variables. FuXi-S2S, trained on 72 years of daily statistics from ECMWF ERA5 reanalysis data, outperforms the ECMWF's state-of-the-art Subseasonal-to-Seasonal model in ensemble mean and ensemble forecasts for total precipitation and outgoing longwave radiation, notably enhancing global precipitation forecast. The improved performance of FuXi-S2S can be pri-marily attributed to its superior capability to capture forecast uncertainty and accurately predict the Madden-Julian Oscillation (MJO), extending the skillful MJO prediction from 30 days to 36 days. Moreover, FuXi-S2S not only captures realistic teleconnections associated with the MJO but also emerges as a valu-able tool for discovering precursor signals, offering researchers insights and potentially establishing a new paradigm in Earth system science research.

Subseasonal forecasting, which predicts weather patterns from 2 to 6 weeks in advance, bridges a critical gap between short-term weather forecasts, typically up to 15 days, and longer-term climate forecasts that extend to seasonal and longer timescales[1]. Forecasting at this intermediate subseasonal timescale is indispensable for a variety of applications, including agricultural planning, disaster preparedness, mitigating impacts of extreme events such as heatwaves, droughts,

floods, and cold spells, and water resource management[2-5]. Despite its significant socioeconomic benefits, subseasonal forecasting has historically not received sufficient attention compared to medium-range weather and climate predictions. This gap existed because accurate subseasonal forecasts were once considered nearly impos-sible. Subseasonal forecasts are particularly challenging as they rely on both atmospheric initial conditions, essential in short-term

[1]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China. [2]Shanghai Academy of Artificial Intelligence for Science, Shanghai, China. [3]China Meteorological Administration Key Laboratory for Climate Prediction Studies, National Climate Center, Beijing, China. [4]Xiong'an Institute of Meteorological Artificial Intelligence, Xiong'an, China. [5]University of Gothenburg, Gothenburg, Sweden. [6]Scripps Institution of Oceanography, University of California San Diego, San Diego, CA, USA. [7]School of Data Science, Fudan University, Shanghai, China. [8]Institute for Big Data, Fudan University, Shanghai, China. [9]MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai, China. [10]These authors contributed equally: Lei Chen, Xiaohui Zhong, Hao Li, Jie Wu. ✉e-mail: lihao_lh@fudan.edu.cn; bolu@cma.gov.cn; qiyuan@fudan.edu.cn

weather forecasts, and boundary conditions at the Earth's surface, key to seasonal and climate forecasts[6,7]. However, neither of these conditions provides sufficient predictability, leaving subseasonal forecasts in a so-called predictability desert. Despite these challenges, recent advances in both physical and statistical modeling have enabled the regular production of subseasonal forecasts globally. Nonetheless, there remains a ongoing, strong demand for their further development to support informed decision-making across various sectors.

Developing an ensemble prediction system (EPS) based on traditional physics-based numerical weather prediction (NWP) models is a widely acknowledged and effective method for enhancing subseasonal forecast accuracy[8,9]. Major forecasting centers have implemented such EPS for subseasonal forecasts[3,10–12]. However, these systems often exhibit considerable biases[13–17], particularly in predicting extreme events[18]. The two primary challenges in this field are ensuring an adequate ensemble size within computational constraints and designing ensemble perturbations that accurately reflect uncertainty in key atmospheric and oceanic variability[19]. Enlarging the ensemble size is beneficial for forecast performance[20–22], but the substantial computational costs typically limit ensemble sizes to between 4 and 51 members across 11 international forecasting centers[12]. Given these computational limitations, machine learning model emerges as a promising alternative for direct subseasonal forecasting[23]. Machine learning models have the advantages of significantly higher computational efficiency, facilitating the generation of a large number of ensemble members which are crucial for prediction skill and reliability[24]. Recent advancements in machine learning for medium-range weather forecasting[25–31] have demonstrated that machine learning models can outperform the high-resolution forecasts (HRES) generated by the European Centre for Medium-Range Weather Forecasts (ECMWF), widely considered as the most accurate global weather forecasts[32].

Machine learning models have achieved made significant strides in medium-range weather forecasting and seasonal forecasting[33], but their success in subseasonal forecasting has been less pronounced[8,34,35]. This shortfall primarily stems from the limited range of variables incorporated into the models, and more importantly, from the inadequate methods employed for ensemble generation. Conventional machine learning techniques for ensemble forecasting, such as introducing random perturbations into initial conditions and altering model structures, overlook the background flow and consequently lead to rapid reduction in ensemble spread. The inadequate representation of the complexities limits the performance of these prior machine learning-based subseasonal forecasting models, which do not yet rival that of traditional EPS based on NWP models. To overcome these challenges, we introduce the FuXi Subseasonal-to-Seasonal (FuXi-S2S) model, representing a significant advancement in machine learning for subseasonal forecasting. This model is designed to generate global daily mean forecasts for 42 days from initialization. Unlike previous models that incorporated a limited set of variables, it incorporates a comprehensive suite of variables, instead of a couple of variables in previous models: 5 upper-air atmospheric variables at 13 pressure levels and 11 surface variables. Furthermore, it features a innovative perturbation module specifically designed to generate flow-dependent perturbations for ensemble forecasting. This module leverages vast amounts of historical data to learn probability distributions, thereby introducing flow-dependent perturbations directly into the model's hidden features. Compared to conventional NWP ensemble forecasting methods, which often struggle with constructing initial condition perturbations due to the complexities of multivariate interactions and the need to maintain dynamic balance and ensemble spread in simulations[36], our approach of introducing perturbations directly into the model's latent space, presenting an effective alternative. This perturbation module significantly enhances the performance of the FuXi-S2S forecasts, as demonstrated

in Supplementary Fig. 1. More details about the FuXi-S2S model architecture are available in "Methods".

Remarkably, FuXi-S2S outperforms the ECMWF Subseasonal to Seasonal (S2S) ensemble, which is recognized as the most skillful S2S modeling system, in producing both the ensemble mean and probabilistic forecasts[5,37]. Its efficacy is particularly evident in extreme total precipitation (TP) forecasting, as exemplified by its accurate forecasts for the 2022 Pakistan floods. Such capability is closely related to FuXi-S2S's improved prediction of the Madden–Julian Oscillation (MJO)[38,39], a key driver of global climate patterns, extending the skillful MJO prediction from 30 days to 36 days. These results further confirm that the notable improvement in FuXi-S2S's performance can be primarily attributed to the innovative perturbation module for ensemble generation. Another promising result is the ability of the FuXi-S2S model to identify potential precursor signals to physical processes. Beyond mere accuracy, in many applications involving machine learning forecasts, it is imperative to understand and validate the decision-making mechanisms of these models. Such understanding not only leads to enhanced trust in the models' predictions but also increases the likelihood of implementing effective actions, particularly in mitigating the risks associated with extreme events. Therefore, interpreting machine learning models to align their reasoning with established knowledge becomes crucial. Recent developments in explainable machine learning (XML)[40–45] methods have facilitated this interpretation. This study delves into the 2022 Pakistan floods, investigating the FuXi-S2S model's predictions to identify key geographic regions that significantly impact its predictive accuracy. This is achieved through the generation and analysis of saliency maps[46], wherein the identified regions in close alignment with insights from previous studies[47]. Therefore, we argue that FuXi-S2S transcends traditional NWP models in terms of accuracy and speed, potentially unveiling previously unrecognized processes within Earth's system in subseasonal forecasting[48,49].

Results

This study conducts a thorough evaluation of the 51-member FuXi-S2S forecasts by analyzing testing data spanning from 2017 to 2021. It compares the performance of FuXi-S2S with that of the 11-member ECMWF S2S reforecasts from the model cycle C47r3 over the same period. The analysis primarily focuses on average forecasts for week 3 (days 15–21), week 4 (days 22–28), week 5 (days 29–35), and week 6 (days 36–42), weeks 3–4, and weeks 5–6. The evaluation employs a comprehensive set of metrics, including deterministic metrics for the ensemble mean, probabilistic metrics for all ensemble members, prediction skills specific for MJO forecasts, and tailored assessments for extreme events, notably the 2022 Pakistan floods. Furthermore, the study explores the underlying processes driving the FuXi-S2S model's predictions for the 2022 Pakistan floods. This is accomplished by generating and analyzing the saliency maps, which provide profound insights into the model's predictive processes.

Additional evaluations, including an analysis of energy spectra[50], are available in the Supplementary Material.

Deterministic metrics

This subsection compares the performance of ensemble mean forecasts from FuXi-S2S and ECMWF S2S based on deterministic metrics. Figure 1 presents the globally-averaged and latitude-weighted temporal anomaly correlation coefficient (TCC) for both FuXi-S2S and ECMWF S2S, considering four variables: TP, 2 m temperature (T2M), geopotential at 500 hPa (Z500), and outgoing longwave radiation (OLR), across forecast lead times of 3, 4, 5, 6, 3–4, and 5–6 weeks. Significance testing is conducted as described in "Evaluation method". When the FuXi-S2S forecasts do not show a statistically significant improvement over the ECMWF S2S reforecasts, these are indicated with a pale color scheme. It is evident that the ensemble mean
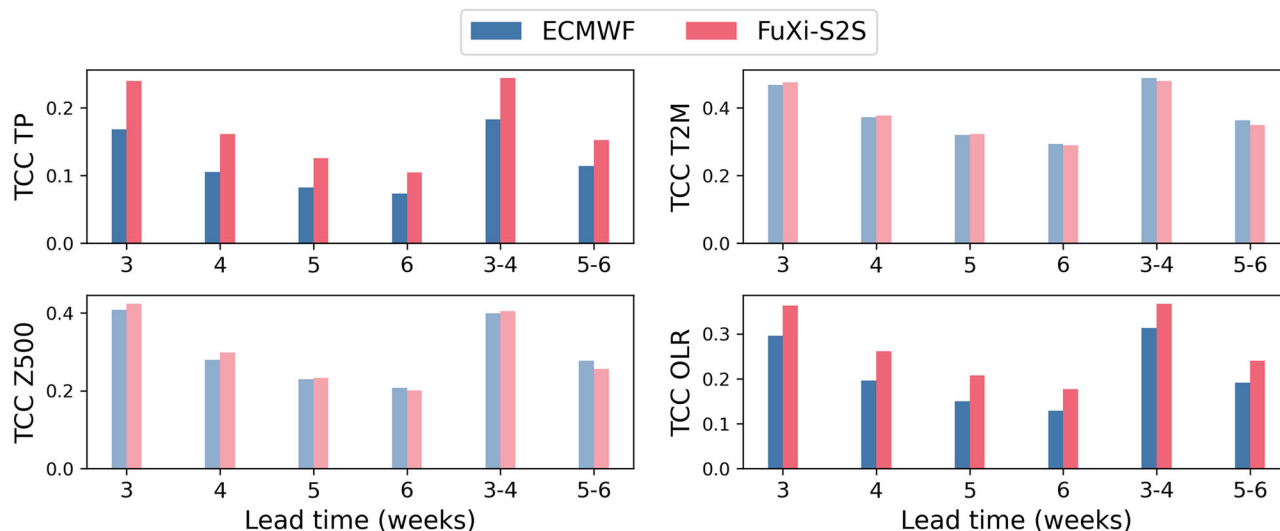
Fig. 1 | Comparison of globally-averaged and latitude-weighted temporal anomaly correlation coefficient (TCC) of the ensemble mean between ECMWF subseasonal-to-seasonal (S2S) reforecasts (in blue) and FuXi-S2S forecasts (in red) for total precipitation (TP), 2 m temperature (T2M), geopotential at 500 hPa (Z500), and outgoing longwave radiation (OLR). Rows 1 and 2 represent the performance across these variables, utilizing all testing data from the period spanning from 2017 to 2021. A bootstrapping approach, repeated 1000 times, is used for significance testing. When the FuXi-S2S forecasts fail to show a statistically significant improvement over the ECMWF S2S reforecasts at the 97.5% confidence level, a pale color scheme is used to denote these results.

forecasts from FuXi-S2S significantly outperform ECMWF S2S for TP and OLR, but not for T2M and Z500. The analysis is based on the averaged TCC computed from all testing data spanning the period from 2017 to 2021. The FuXi-S2S forecasts generally demonstrate higher TCC values than the ECMWF S2S reforecasts for TP and OLR at all lead times, while comparable TCC values for Z500 and T2M. Specifically, regarding Z500, the FuXi-S2S forecasts are superior to the ECMWF S2S reforecasts at lead times of 3, 4, 5, and 3–4 weeks, and have inferior performance at lead times of 6 and 5–6 weeks.

Supplementary Fig. 2 provides the spatial distributions of temporally-averaged TCC for both ECMWF S2S and FuXi-S2S, along with the differences in TCC between FuXi-S2S and ECMWF S2S for TP, T2M, Z500, and OLR forecasts at lead times of 3–4 and 5–6 weeks, respectively. The spatial distributions of TCC reveal considerably higher values over tropics, and greater values over oceans than over land. The TCC differences are described in red (positive values), blue (negative values), and white (zero values) patterns, suggesting whether FuXi-S2S's performance is superior, inferior, or equivalent to ECMWF S2S, respectively. Overall, FuXi-S2S demonstrates positive TCC differences for TP and OLR in most regions worldwide, consistent with the findings presented in Fig. 1. Moreover, FuXi-S2S also outperforms ECMWF in a majority of extra-tropical regions for both T2M and Z500, although its performance is generally less skillful in the tropical areas.

Probabilistic metrics

Deterministic metrics, evaluated using the ensemble mean, exhibit limited predictive skill, with TCC values below 0.5 for all subseasonal forecast lead times. Therefore, ensemble forecasts are essential for detecting predictable signals at subseasonal timescales.

The first two rows of Fig. 2 present the spatial distribution of the temporally-averaged ranked probability skill score (RPSS)[51,52] for ECMWF S2S and FuXi-S2S, as well as the RPSS differences between FuXi-S2S and ECMWF S2S for TP forecasts over 3–4 and 5–6 week lead times. This analysis utilizes RPSS data which are temporally averaged from 2017 to 2021. The red contour lines in the first and second columns highlight areas with positive RPSS values, which indicate more skillful prediction than climatology forecast can be obtained over these areas. Notably, FuXi-S2S predicts more areas with positive RPSS values than ECMWF S2S. The color coding in the right panels of Fig. 2

(red, blue, and white) indicates regions where FuXi-S2S performs better, worse, or equivalently compared to ECMWF S2S, respectively. The global distribution of RPSS suggests that both ECMWF S2S and FuXi-S2S primarily exhibit skill in tropical regions, whereas they lack skill in the extra-tropics compared to climatology. In contrast, RPSS demonstrates positive values (depicted in red color) in tropical regions, indicating enhanced predictive skills relative to climatology. Moreover, the RPSS values are notably higher over oceans compared to land areas. Predominantly, FuXi-S2S demonstrates nearly global positive RPSS differences for TP, except in some tropical regions where both models have quite high RPSS values. Compared to ECMWF S2S, whose skillful predictions are primarily confined to tropical ocean areas, FuXi-S2S demonstrates the capability of skillful predictions over more extra-tropical regions, such as East Asia, the North Pacific, and the Arctic.

The latitude-weighted RPSS for the same 4 variables as in Fig. 1 over forecast lead times of 3, 4, 5, 6, 3–4, and 5–6 weeks are given in Supplementary Fig. 6. FuXi-S2S shows higher RPSS values than ECMWF S2S across most regions for all the examined variables: TP, T2M, Z500, and OLR. This superiority is especially noticeable in extra-tropical averages. However, in the tropics, ECMWF S2S outperforms FuXi-S2S at lead times of 3 to 6 weeks for 1-week averages, whereas FuXi-S2S surpasses ECMWF S2S for 2-week averages. This discrepancy in performance likely arises from the fact that 1-week averages filter out variability with periods shorter than 2 weeks, while 2-week averages attenuate variability with periods shorter than 4 weeks. Thus, the skill differences between the 1-week and 2-week averages may reflect FuXi-S2S's enhanced ability in capturing lower-frequency variability. Furthermore, a previous study[37] suggests that dynamical S2S models, particularly ECMWF S2S, demonstrate improved performance in the central-eastern Pacific, potentially due to their effective simulation of the realistic air-sea interactions in these regions.

Extreme forecast

A primary target of subseasonal forecasts is extreme weather events, to better prepare for disasters like droughts and floods. This subsection focuses on the prediction skills for extreme precipitation events. Such events are identified when TP exceeds the 90th climatological
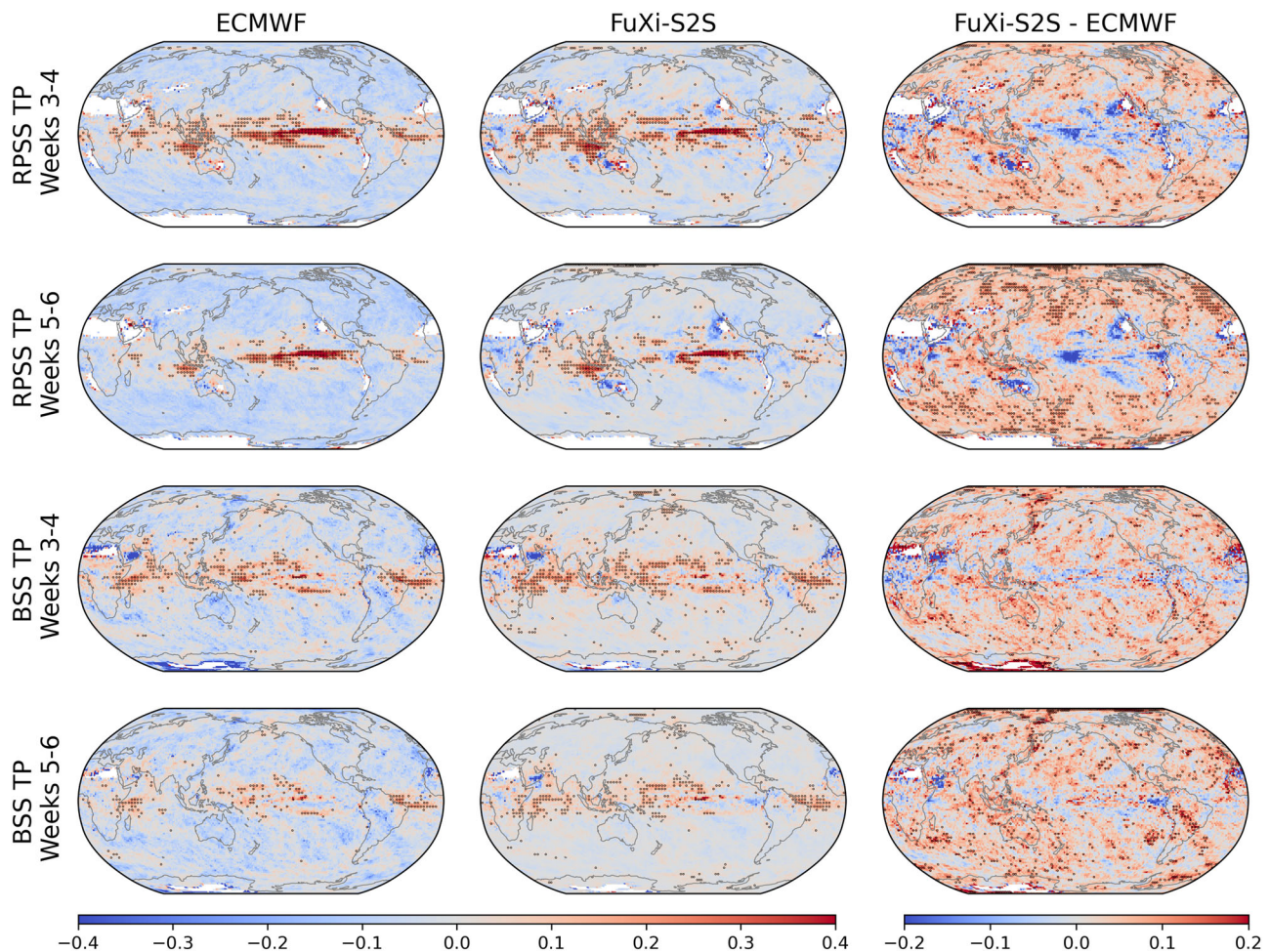
Fig. 2 | Maps displaying the average Ranked Probability skill Score (RPSS) (first and second rows) and Brier Skill Score (BSS) (third and fourth rows) without latitude weighting, comparing ECMWF subseasonal-to-seasonal (S2S) (first column) and FuXi-S2S (second column) forecasts. Additionally, the third column depicts the difference in RPSS and BSS between FuXi-S2S and ECMWF S2S for total precipitation (TP) at forecast lead times of weeks 3–4 (first and third rows) and weeks 5–6 (second and fourth rows), utilizing all testing data from 2017 to 2021.

Red contour lines in the first and second columns indicate areas with positive values of RPSS and BSS. Stippling on the map denotes areas where the skill score is statistically significant at the 97.5% confidence level. Specifically, in columns 1 and 2, stippling indicates regions where the skill scores of the ECMWF S2S and FuXi-S2S models significantly surpass those of climatology. In column 3, stippling highlights areas where the FuXi-S2S model significantly outperforms the ECMWF S2S.

percentile, a threshold that varies based on grid location, forecast initialization time, and forecast lead time.

The last two rows of Fig. 2 show the spatial distributions of the temporally-averaged Brier Skill Score (BSS)[52] for the extreme precipitation events, for ECMWF S2S and FuXi-S2S, and their differences over 3–4 and 5–6 week lead times. Similar to spatial pattern of RPSS, FuXi-S2S generally exhibits more regions with positive values of BSS than ECMWF S2S, suggesting more areas with skill relative to climatological forecasts. Similar to spatial pattern of RPSS, the BSS values are considerably higher over oceans than over land and decrease from lower latitudes to higher latitudes. Predominantly, the BSS differences favor FuXi-S2S in TP over land and in extra-tropical regions, marked by widespread red patterns. This suggests FuXi-S2S's dominance over ECMWF S2S in predicting extreme TP across land and extra-tropics, which is of great importance for disaster preparedness and early warning.

Supplementary Fig. 7 compares the latitude-weighted BSS between FuXi-S2S and ECMWF S2S, focusing on TP, T2M, Z500, and OLR in five geographical regions: global, in the extra-tropics (90°S–30° S and 30° N–90° N), in the tropics (30°S–30°N), over land, and over the ocean. Globally, FuXi-S2S outperforms ECMWF S2S in terms of BSS for TP, T2M, and OLR. Notably, in contrast to ECMWF S2S,

which exhibits consistently negative globally-averaged BSS values for TP across all lead times, FuXi-S2S demonstrates positive values for forecast lead times of 3, 3–4, and 5–6 weeks. In the extra-tropical regions, though the BSS scores are relatively lower in comparison to the global average, FuXi-S2S consistently exhibits superior performance compared to ECMWF S2S across all four variables. A similar pattern emerges in tropical regions, where FuXi-S2S demonstrates superior performance over ECMWF S2S for TP and OLR while achieving comparable accuracy in T2M and Z500. Over land areas, FuXi-S2S demonstrates consistently higher BSS values for TP and T2M, suggesting its superior ability to provide more accurate forecasts of extreme rainfall and high temperatures compared to ECMWF S2S.

MJO forecast

Recent studies have demonstrated the importance of accurately modeling various sources of subseasonal predictability, particularly the MJO[12,53,54], for improving subseasonal prediction skills. The MJO has a significant impact on global weather and climate, serving as a primary source of predictability at subseasonal timescales due to its quasi-periodic nature[55–58]. Accurate MJO prediction is essential for reliable subseasonal predictions. Although current state-of-the-art dynamical forecasts can predict the MJO up to 3–4 weeks in advance,
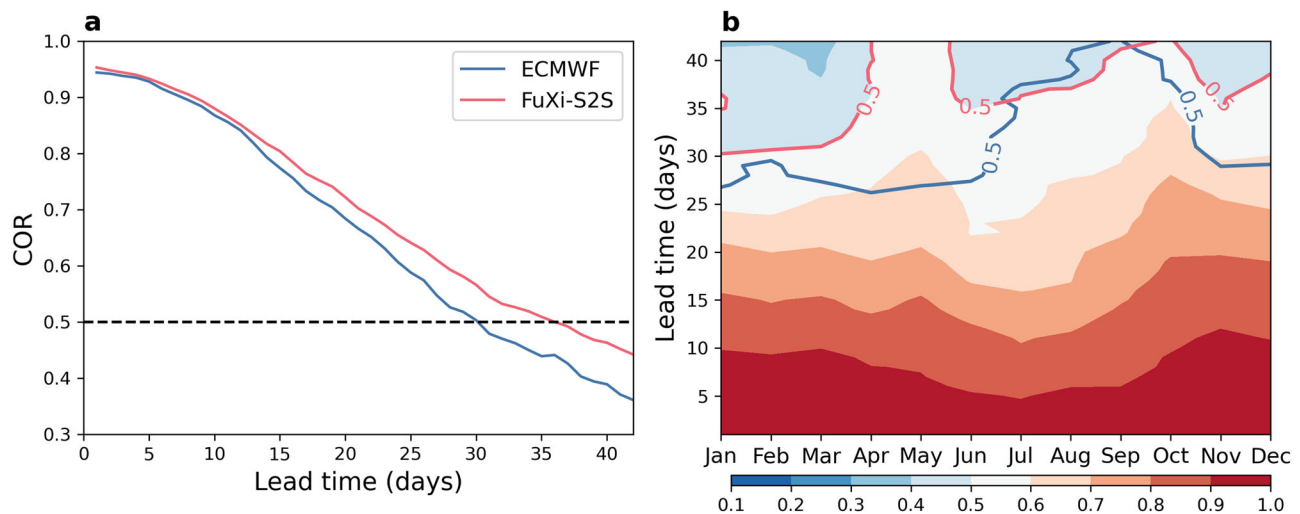
Fig. 3 | Comparison of real-time multivariate Madden–Julian Oscillation (MJO) (RMM) bivariate Correlation (COR) of the ensemble means between ECMWF subseasonal-to-seasonal (S2S) reforecasts (in blue) and FuXi-S2S forecasts (in red) using all testing data from 2017 to 2021. a Comparison of RMM bivariate COR as a function of forecast lead times. Dashed black line signifies the prediction skill threshold of COR = 0.5. **b** The RMM bivariate COR is depicted as a function of the month of initialization (x-axis) and forecast lead time (y-axis), with red and blue lines indicating the skillful MJO prediction days of ECMWF S2S (in blue) and FuXi-S2S (in red), respectively.

this falls short of the theoretical potential predictability of approximately 6–7 weeks[58–60]. In recent years, increasing efforts have focused on applying machine learning models to improve MJO forecasts, either by post-processing dynamical forecasts[61–63] or through direct forecasting[44,64,65]. However, only improving MJO predictions with machine learning models does not inherently ensure improved forecasts of related weather phenomena, such as tropical cyclones and monsoons, which also depend on accurate predictions of various weather parameters by the model. Therefore, continuous improvement in forecasting models is essential for advancing subseasonal prediction capabilities. This section specifically examines the performance of our FuXi-S2S model in MJO forecasts, although it is not explicitly optimized for this purpose.

In this study, we employed the real-time multivariate MJO (RMM) index[66], along with the commonly used metrics of bivariate correlation coefficient (COR), to evaluate the forecasting skill of the MJO. The RMM index used for verification was calculated using the Climate Prediction Center (CPC) OLR (CBO) data, in conjunction with the ERA5 zonal-wind component at 850 hPa and 200 hPa. Figure 3 presents the bivariate correlation (COR) skills of the RMM index for the ensemble mean of ECMWF S2S reforecasts and FuXi-S2S forecasts, averaged over the testing data spanning from 2017 to 2021. The results show a decrease in COR values as forecast lead times increase. Particularly, FuXi-S2S outperforms ECMWF S2S in MJO prediction, maintaining higher COR values for up to 42 days. When applying a COR threshold of 0.5 to determine skillful MJO forecast, FuXi-S2S extends the skillful forecast lead time from 30 days to 36 days, surpassing the performance of ECMWF S2S. Furthermore, the MJO prediction skills also depend on the seasonal cycle, as illustrated in Fig. 3. Both FuXi-S2S and ECMWF S2S demonstrate higher MJO prediction skills in September and October. Additionally, FuXi-S2S exhibits superior skills compared to ECMWF S2S during the boreal spring and winter, with skillful predictions extending beyond 42 days in April and May, which is the longest forecast lead time achievable by the FuXi-S2S model. Moreover, Supplementary Fig. 9 presents the COR and error for the amplitude and phase of the MJO. These are calculated using the ensemble mean of ECMWF S2S reforecasts and FuXi-S2S forecasts, averaged across over the 2017–2021 testing dataset. The results suggest that the FuXi-S2S model outperforms the ECMWF S2S model in predicting the MJO, primarily due to its superior capability in

forecasting the MJO phase. Additionally, FuXi-S2S demonstrates smaller amplitude errors, suggesting it more accurately maintains the amplitude of MJO events.

A two-dimensional phase-space diagram is commonly used to characterize the phase and amplitude of the MJO, using the x-axis and y-axis to represent the first and second principal components of Empirical Orthogonal Functions (EOFs) (RMM1 and RMM2), respectively. Supplementary Fig. 10 illustrates the forecast performance of four distinct MJO events with initialization dates of 27 June 2018, 3 November 2018, 18 April 2019, and 21 March 2021, as predicted by ECMWF S2S and FuXi-S2S. Data points on this two-dimensional phase-space diagram are plotted at 5-day intervals. The phase of the MJO is determined by the azimuth of the combined RMM indices 1 and 2 (RMM1 and RMM2), while its amplitude is represented by the radial distance from the origin. As visually shown in Supplementary Fig. 10, the counterclockwise movement of data points signifies the eastward propagation of MJO-associated convection, with the distance between successive points reflecting the propagation speed. In comparison to the observed MJO derived from CBO and ERA5 reanalysis data, both ECMWF S2S and FuXi-S2S exhibit slower propagation speeds and reduced amplitudes as the forecast lead time increases, particularly noticeable for MJO forecasts initialized on 21 March 2021. However, FuXi-S2S shows a more consistent alignment with observations across all MJO phases, especially in mitigating the negative amplitude biases in MJO forecasts when compared to ECMWF S2S.

The MJO originates from interactions of tropical convection and circulation but its effect is of global reach. Indeed, large TCC for Z500 over the extra-tropical Pacific is found along the path of the Pacific North/South American (PNA/PSA)[67,68] teleconnection pattern (Supplementary Fig. 2, rows 6 and 7). Compared to ECMWF S2S, improved MJO forecast in FuXi-S2S elevates TCC for these teleconnection patterns, especially along the PSA wave train in the Southern Hemisphere. Furthermore, the MJO is critical for stimulating these important teleconnection patterns, significantly affecting extra-tropical anomalies. Therefore, the accurate representation of MJO-related teleconnections is imperative for effective subseasonal forecasts. Supplementary Fig. 11 demonstrates that the FuXi-S2S model showcases enhanced skills in MJO prediction and realistic simulations of MJO teleconnections, which substantially contribute to its superior performance in subseasonal forecasts, particularly over extra-tropical regions.
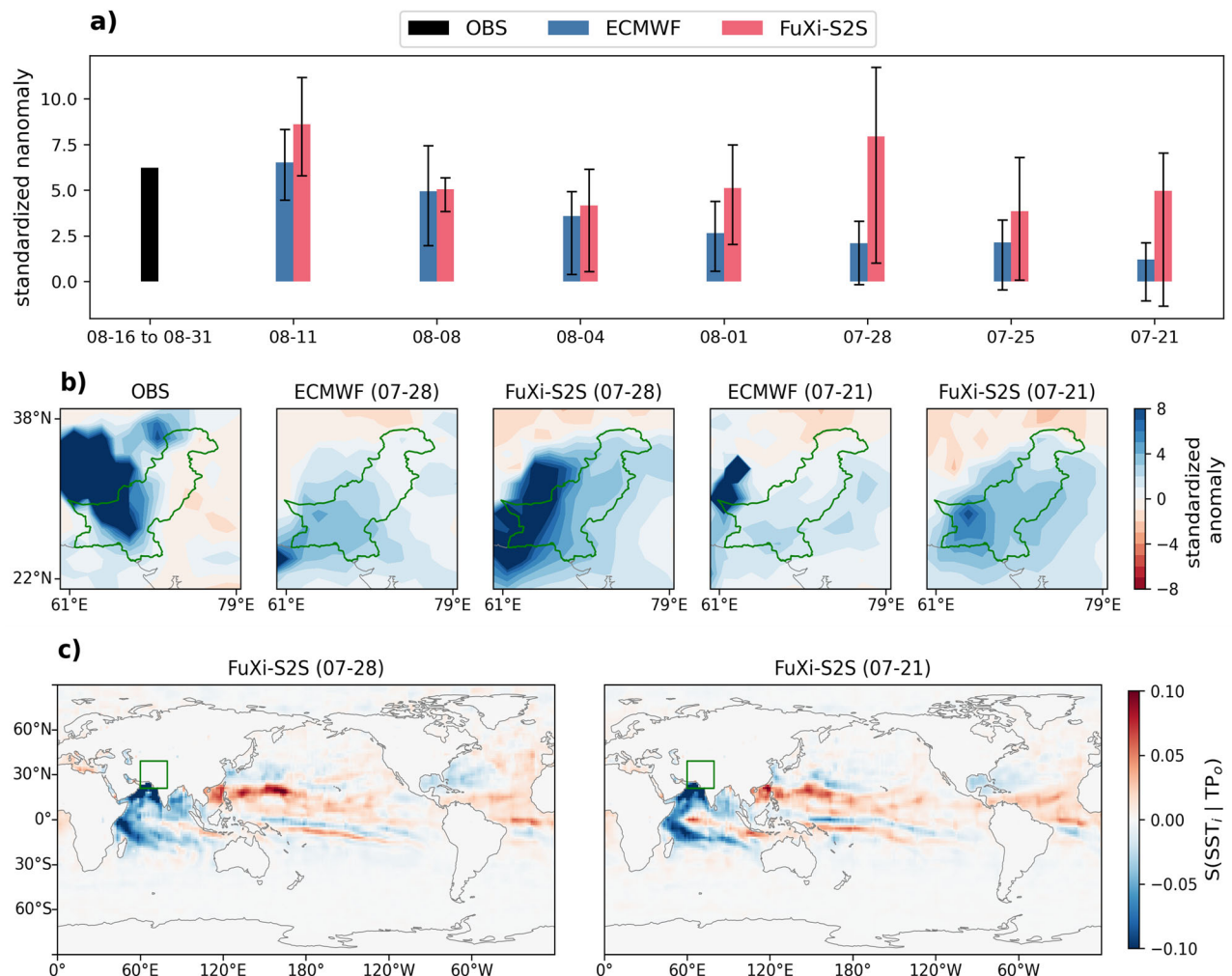
Fig. 4 | Comparative analysis for the 2022 Pakistan floods predictions between the ECMWF subseasonal-to-seasonal (S2S) and FuXi-S2S models as well as the precursor signals that contributed to accurate predictions by the FuXi-S2S model. Comparison of spatially and temporally averaged standardized total precipitation (TP) anomaly (**a**) over the 2 weeks from August 16th to August 31st, 2022, showcasing GPCP observations (in black) alongside predictions from ECMWF S2S real-time forecasts (in blue) and FuXi-S2S forecasts (in red), with initialization dates: August 11th (08-11, MM-DD), August 8th (08-08), August 4th (08-04), August 1st (08-01), July 28th (07-28), July 25th (07-25), and July 21st (07-21). The black lines on the bar of ECMWF S2S and FuXi-S2S forecasts represent the 25th and 75th percentiles. For the comparison of temporally averaged standardized TP anomaly maps (**b**), the first column represents GPCP observations, while the second and third columns display predictions from ECMWF S2S and FuXi-S2S, respectively, both initialized on July 28th, and the fourth and fifth columns correspond to predictions from ECMWF S2S and FuXi-S2S, respectively, with an initialization date of July 21st. Green contour indicates the border line of Pakistan. The saliency maps (**c**) were generated using the gradient of the negative standardized TP anomaly, averaged over the Pakistan region, in relation to the input SST. These maps correspond to forecasts initialized on July 28th (07-28, first column) and July 21st (07-21, second column). Here, the red and blue colors indicate the positive and negative correlations between the negative of standardized TP and variations in SST. The black lines on the bars in this figure represent the 25th and 75th percentiles of the ensemble forecasts for each start date for both ECMWF and FuXi-S2S models.

This study highlights FuXi-S2S proficiency in predicting the MJO. We envision that FuXi-S2S could serve as a pivotal tool in investigating other primary modes of subseasonal variability, such as the Boreal Summer Intraseasonal Oscillation (BSISO)[69], North Atlantic Oscillation (NAO)[70], and East Asia–Pacific (EAP) pattern[71]. Additionally, it would be worthwhile to explore how the prescribed fixed sea surface temperature (SST) or its absence impacts the forecast performance of the MJO. Savarin and Chen[72] demonstrated that either using a coupled atmosphere-ocean model or updating SST with observed values is essential for accurately modeling the eastward propagation of the MJO. However, this analysis is beyond the scope of the current study and will be addressed in future research.

Prediction of the 2022 Pakistan floods

In 2022, Pakistan experienced a series of exceptionally intense monsoon rainfall surges from early July to late August, resulting in total rainfall that reached a level approximately four standard deviations above the climatological mean[73]. This extreme rainfall event led to a significant humanitarian disaster, leaving over 2.1 million people homeless and resulting in 1730 fatalities. According to the World Bank, the economic damages and losses exceeded USD 30 billion[47]. Consequently, it is important to assess the ability of subseasonal forecasts to predict such extreme rainfall events.

Figure 4 illustrates the observed standardized TP anomaly alongside predictions that were initialized on different dates, generated by both the FuXi-S2S and ECMWF S2S models. These observations, taken from the Global Precipitation Climatology Project (GPCP), are spatially averaged over the Pakistan region (60–70° E in longitude and 25–35° N in latitude), and temporally over a 2-week period from August 16th to August 31st, 2022, corresponding to the period of most intense rainfall. The standardized anomaly for observed rainfall is ~6 standard deviations above the climatological mean. It is evident that

the ECMWF S2S model considerably underestimates rainfall intensity for forecasts initialized on July 21st, achieving only about one-third of the observed values. The ECMWF S2S forecasts gradually converge toward observations as the initialization dates approach the actual event. In contrast, FuXi-S2S exhibits superior forecast performance in predicting the intensity of extreme rainfall events earlier compared to ECMWF S2S. Specifically, FuXi-S2S predicts rainfall levels of at least 4 standard deviations above the climatological mean for forecasts initialized on July 21st, which is approximately 4 weeks in advance. Moreover, the spatial distributions of the standardized TP anomaly reveal that the FuXi-S2S predicted TP pattern more closely matches the observations.

Forecast skill typically improves with decreasing lead time, as in the ECMWF S2S model. The rainfall anomaly grows in FuXi-S2S forecasts initialized on July 28 (lead time of 18 days), albeit with a large forecast spread, possibly due to SST influence. Indeed, the saliency maps show that the FuXi-S2S forecasts initialized on July 28 and July 21 successfully captured predictable signals from SST anomalies in the tropical central Pacific and western Indian Ocean (Fig. 4c). At shorter lead times, the SST influence decreases while the effect of atmospheric initial conditions increases. The varying importance of SST and initial conditions may cause variability in the FuXi-S2S forecasts with lead time.

Discovery of precursor signals for the 2022 Pakistan floods prediction

Data-driven machine learning forecasting models, such as FuXi-S2S, often lack explicit integration of prior knowledge about the physical system they aim to predict. As a result, they are often referred to as "black boxes". Although FuXi-S2S has shown accuracy in previous subsections, the opacity of its predictive processes can diminish confidence in its reliability. Therefore, it is imperative to interpret FuXi-S2S, ensuring that their underlying reasoning is consistent with established understanding of weather systems. Here, we generated saliency maps to disentangle the key driving processes behind the FuXi-S2S model's prediction of the 2022 floods in Pakistan.

In this study, we utilized the negative absolute values of the TP anomaly, averaged across the Pakistan region (outlined by the green box in Fig. 4c), as a loss function. By implementing backward propagation of this loss function to calculate gradients, we obtained the saliency maps. These maps use red and blue colors to signify positive and negative correlations, respectively between the negative of standardized TP anomaly and SST. Specifically, blue (red) areas indicate that a decrease (increase) in SST is associated with an increase (decrease) in the negative of standardized TP anomaly, thereby leading to an increase (decrease) in TP anomaly. Analysis of these saliency maps facilitated the identification of potential precursor signals and sources of predictability that contributed to the occurrence of the extreme TP event. As illustrated in Fig. 4c, SST precursor signals, identified in forecasts initialized on different dates (July 28th and July 21st in 2022), show remarkable consistency. These signals indicate a consistent cooling of SST in the equatorial central Pacific and the tropical western Indian Ocean, along with warming in the tropical eastern Pacific. This spatial pattern aligns closely with findings from previous studies[47], which pinpointed the rapid development of a La Niña in the tropical Pacific and a negative phase of the Indian Ocean Dipole in the summer of 2022 as key precursor signals and driving forces of Pakistan's intense TP event. Our results confirm that the high predictive skill of the FuXi-S2S model can be attributed to its effective capture of the primary predictable sources of this event. Furthermore, these findings demonstrate the model's potential as a valuable tool for rapidly exploring the mechanisms behind extreme events and uncovering teleconnections within Earth's systems, thereby enhancing our physical understanding. Here, we focus on the gradient with respect to SST. Nevertheless, it is important to acknowledge the existence of

other significant precursor signals that may be associated with this extreme event, including U, V, and Z anomalies as noted in ref. 73. A more comprehensive examination of these factors is intended for future research.

Discussion

In this paper, we introduced FuXi-S2S, a machine learning-based subseasonal forecasting model. This model provides global forecasts of daily mean values for up to 42 days, with a daily temporal resolution and 1.5° spatial resolution encompassing five upper-air atmospheric variables across 13 pressure levels and 11 surface variables. The performance of FuXi-S2S was rigorously evaluated against ERA5 reanalysis data and compared with ECMWF S2S reforecasts. A comprehensive suite of metrics was employed for this evaluation, including the deterministic metrics of the ensemble mean, the probabilistic metrics of the ensemble forecast, and the capability to predict extreme events. Our results demonstrated that FuXi-S2S surpasses ECMWF S2S in forecast accuracy for the evaluated variables. Furthermore, FuXi-S2S significantly improves accuracy in predicting the MJO, extending the skillful MJO prediction from 30 days to 36 days. This improvement is particularly important given the MJO's influence on global climate patterns, and consequently, it improves the model's TP) forecast accuracy globally. Moreover, FuXi-S2S has shown utility in practical scenarios, such as its superior performance in predicting the extreme rainfall during the 2022 Pakistan floods earlier than the ECMWF S2S model. This early prediction capability is vital for improving disaster preparedness and response.

A key contributor to the superiority of FuXi-S2S is its innovative method of generating perturbations, which is essential for its successful ensemble forecasting. Unlike conventional models that employ random or meticulously calculated perturbations in initial conditions, FuXi-S2S incorporates background flow-dependent perturbations into its hidden features. These flow-dependent perturbations have shown to significantly enhance model's subseasonal forecast performance, as illustrated in Supplementary Fig. 1. FuXi-S2S, as a machine learning model, also distinguishes itself by its ability to generate large ensemble forecasts rapidly and efficiently, requiring significantly less time and computational resources than traditional models. Specifically, it can complete a comprehensive 42-day forecast with daily time steps in ~7 s using an Nvidia A100 GPU for a single member. Ensemble size is a critical determinant of the ensemble forecast skill. Research suggests that the optimal number of members for subseasonal forecasts potentially falls within the range of 100 to 200 members[21]. To ensure a fair comparison with the ECMWF S2S model, we have currently limited the FuXi-S2S model to a 51-member ensemble. However, it's important to note that FuXi-S2S is capable of generating larger ensembles with only a moderate increase in computational demands. Our Supplementary Fig. 15 illustrates that increasing the ensemble size to 101 members further enhances the forecast performance of FuXi-S2S compared to the 51-member ensemble.

Beyond its computational efficiency and superior accuracy, FuXi-S2S notably excels in identifying precursor signals and disentangling the complex processes underlying climate extremes, as demonstrated by its accurate prediction of the 2022 floods in Pakistan. Many subseasonal forecasting challenges stem from the limited understanding of these complex processes. Traditional physics-based models often rely on oversimplified representations of physical processes, which diminishes their forecast performance and analytical depth. In contrast, FuXi-S2S demonstrates proficiency in learning complex patterns and identifying subtle teleconnections from vast amounts of data. This approach resonates with Albert Einstein's insight, "You can't solve a problem with the ways of thinking that created it." In our study of the 2022 extreme rainfall event in Pakistan, we demonstrate that backward propagation and the resulting saliency maps successfully reveal that FuXi-S2S makes accurate forecasts by effectively capturing the key

Table 1 | A summary of all the upper-air and surface variable names and their abbreviations in this paper

Type	Full name	Abbreviation
Upper-air variables	geopotential	Z
	temperature	T
	u component of wind	U
	v component of wind	V
	specific humidity	Q
Surface variables	2 m temperature	T2M
	2 m dewpoint temperature	D2M
	sea surface temperature	SST
	outgoing longwave radiation	OLR
	10 m u wind component	U10
	10 m v wind component	V10
	100 m u wind component	U100
	100 m v wind component	V100
	mean sea-level pressure	MSL
	total column water vapor	TCWV
	total precipitation	TP

predictable sources associated with this event. Moreover, such gradient-based interpretation methods aid in explaining weather and climate forecasts made by machine learning models, such as the FuXi-S2S model[74]. Therefore, we advocate for a paradigm shift in the application of machine learning models like FuXi-S2S. The focus should not extend beyond enhancing forecast accuracy to include the development of a comprehensive framework for discovering previously unknown processes within the Earth's system[48,49]. We foresee a growing reliance on machine learning models like FuXi-S2S within the scientific community, acknowledging their essential role in advancing scientific discovery in Earth system science.

While FuXi-S2S offers a computationally efficient and accurate alternative to conventional NWP models for subseasonal forecasting, it also presents significant opportunities for improvement. For instance, the ECWMF S2S model runs at a spatial resolution of 36 km[75], which is considerably finer than the 1.5° resolution of FuXi-S2S. Currently, FuXi S2S predicts daily mean values up to 50 hPa and lacks critical weather parameters such as daily maximum and minimum temperatures, which are essential for some applications. Furthermore, given the known discrepancies between the ERA5 TP data and actual observations, as noted in refs. [76,77], GPCP observations have been utilized to evaluate the TP forecast performance for both ECMWF S2S and FuXi-S2S (refer to Supplementary Fig. 16). Anticipated future enhancements to the FuXi-S2S model include increasing the spatial resolution from 1.5° to 0.25°, incorporating additional weather parameters, extending the forecast beyond the current upper limit of 50 hPa, and employing more accurate TP data sources to enhance forecast accuracy.

Methods
Data
ERA5 stands as the fifth iteration of the ECMWF reanalysis dataset, offering a rich array of surface and upper-air variables. It operates at a remarkable temporal resolution of 1 h and a horizontal resolution of approximately 31 km, covering data from January 1950 to the present day[78]. Recognized for its expansive temporal and spatial coverage coupled with exceptional accuracy, ERA5 stands as the most comprehensive and precise reanalysis archive globally. In our study, we utilize daily statistics derived from the 1-hourly ERA5 dataset, which has a spatial resolution of 1.5° (comprising 121 × 240 latitude–longitude grid points) and a temporal resolution of 1 day. It serves as the sole data source for training the FuXi-S2S model.

Evaluating MJO predictions against MJO indices derived from satellite-observed OLR data is a common practice. Therefore, alongside the ERA5 reanalysis data, a newly developed OLR dataset called the Climate Prediction Center (CPC) OLR (CBO) has emerged. Spanning from 1991 to the present day, this dataset undergoes near real-time updates. While showing slight differences in magnitude compared to the U.S. National Oceanic and Atmospheric Administration (NOAA) Advanced Very High-Resolution Radiometer (AVHRR) OLR, the CBO dataset notably exhibits a high level of similarity in both pattern and magnitude of anomalies. In our research, we utilize the CBO data, which has a spatial resolution of 1° and a temporal resolution of 1 day. This data serves as the ground truth for OLR in the identification and verification of MJO events. Furthermore, for the assessment of rainfall in the Pakistan region, observed rainfall data are sourced from the GPCP dataset[79]. It is noteworthy that the MJO indices derived from ERA5 OLR data closely align with those derived from CBO OLR data.

The FuXi-S2S model forecasts a total of 76 variables, encompassing 5 upper-air atmospheric variables across 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa), and 11 surface variables. Among the upper-air atmospheric variables are geopotential (Z), temperature (T), u component of wind (U), v component of wind (V), and specific humidity (Q). The surface variables include 2 m temperature (T2M), 2 m dewpoint temperature (D2M), sea surface temperature (SST), OLR, 10 m u wind component (U10), 10 m v wind component (V10), 100 m u wind component (U100), 100 m v wind component (V100), mean sea-level pressure (MSL), total column water vapor (TCWV), and TP. OLR is known as the negative of top net thermal radiation (TTR) in ECMWF convention. Table 1 provides a comprehensive list of these variables along with their abbreviations. Variables such as U100 and V100 were selected for their potential utility in wind energy forecasting. The selection of the SST is based on prior research, which suggests that slowly evolving variables like SST are crucial for identifying predictable signals[80–82]. OLR was selected due to its significance in representing MJO events through OLR anomalies.

The model's training relies on 67 years of data spanning from 1950 to 2016, while evaluation involves a 5-year dataset from 2017 to 2021. The z-score normalization technique is employed to normalize all input and output variables, thereby ensuring uniformity in their mean and variance. For upper-air variables, the mean and standard deviation are calculated separately for different vertical levels, using only the training dataset. Additionally, the dataset for the year 2022 undergoes evaluation and comparison against the ECMWF real-time S2S forecasts, specifically concerning the catastrophic flooding in Pakistan. More detailed evaluations of TP and MJO predictions for the year 2022 can be found in the Supplementary Material.

In certain cases, subseasonal forecasts receive regular updates through the implementation of the latest model, incorporating research discoveries tailored for operational use[83]. For instance, the ECMWF S2S reforecasts, often termed hindcasts, which are generated on-the-fly by employing the most recent model version available at the time of forecast generation. In our research, we utilize the ECMWF S2S reforecasts generated from model cycle C47r3. These reforecasts encompass initialization dates over 20 years, ranging from January 3, 2002, to December 29, 2021. The ECMWF S2S reforecasts are initialized twice weekly, aligning with the real-time forecasts. Additionally, our comparative analysis involves employing the 51-member ECMWF real-time S2S forecast for the year 2022. For the analysis using testing data from 2017 to 2021, anomalies for all variables are defined as deviations from the climatological mean calculated over the 15-year period from 2002 to 2016. Meanwhile, for the analysis based on testing data in the year 2022, the climatological mean is calculated over the period from 2002 to 2021. Furthermore, a set of hindcasts from 2002 to 2016 is generated for FuXi-S2S, which are used to establish a
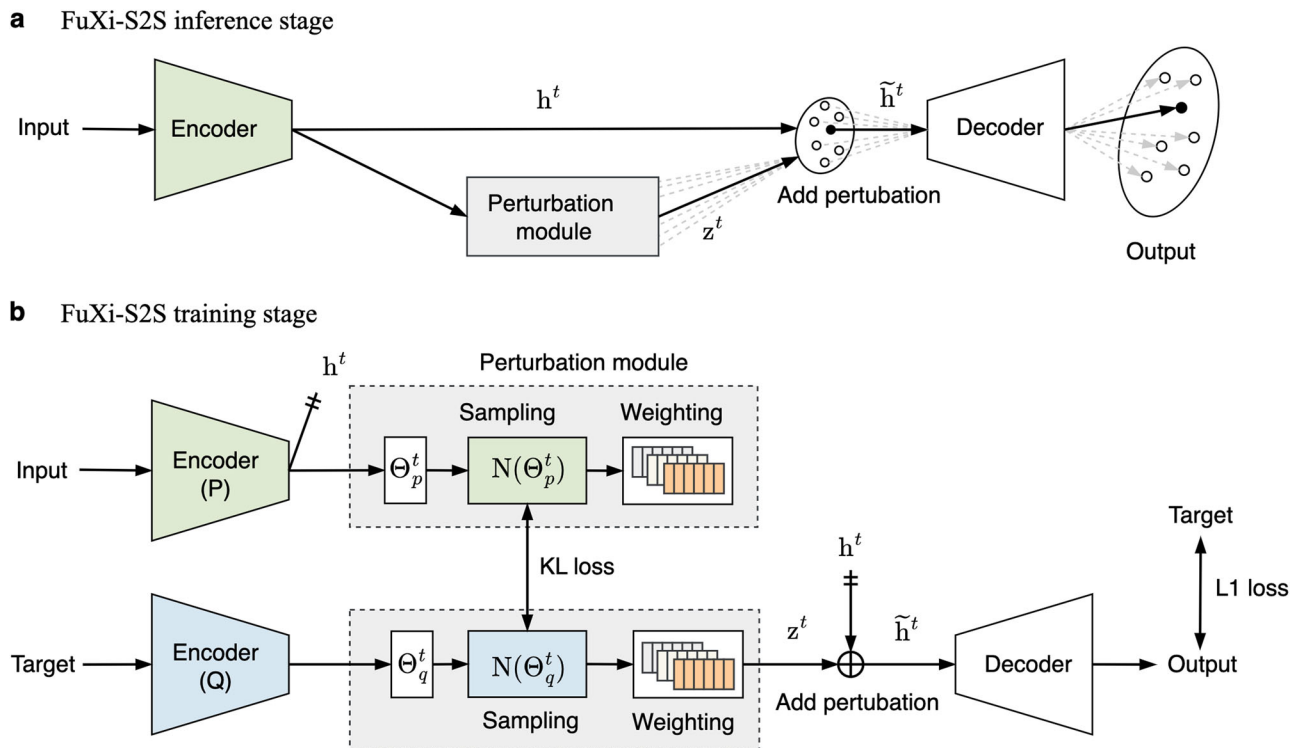
a FuXi-S2S inference stage

b FuXi-S2S training stage

Fig. 5 | Schematic diagram of the structures of the FuXi Subseasonal-to-Seasonal (FuXi-S2S) model. **a** Inference stage of the FuXi-S2S model. h^t represents the hidden feature generated by the Encoder from the input data. The perturbation vector z^t is generated by the perturbation module, resulting in the perturbed hidden feature $\tilde{h}^t$. **b** Training stage of the FuXi-S2S model. $N(\Theta_p^t)$ and $N(\Theta_q^t)$ are the low-rank multivariate Gaussian distributions generated by encoders P and Q, respectively. The Kullback–Leibler (KL) divergence loss measures the discrepancy between the distributions predicted by both encoders, $N(\Theta_p^t)$ and $N(\Theta_q^t)$.

climatology. This climatology is then subtracted from the FuXi-S2S forecasts for the testing data spanning from 2017 to 2021. This process facilitates the calculation of FuXi-S2S anomalies for evaluations.

To ensure equitable comparisons, we evaluate FuXi-S2S forecasts specifically on identical initialization dates corresponding to those utilized for both the ECMWF S2S reforecasts and forecasts. This approach facilitates a fair and direct assessment between FuXi-S2S and ECMWF S2S.

FuXi-S2S model

Most state-of-the-art machine learning models utilized in medium-range weather forecasting are built upon encoder-decoder[84] architectures[27–29,85]. These structures are favored due to their proficiency in processing and generating sequential and spatial data. Within these architectures, the encoder processes key features from the input data and transforms them into a compressed and abstract representation in the latent space. The decoder then utilizes this representation to generate weather forecasts. The primary objective of training these models is to minimize differences between the model's output and the target data. However, the standard encoder-decoder structures are inherently deterministic, producing identical forecasts for the same inputs, which limits their applicability in generating ensemble forecasts. To overcome this limitation, we introduce the FuXi-S2S model, drawing inspiration from Variational Autoencoders (VAEs)[86–88]. VAEs are inherently probabilistic, making them well-suited for tasks that require uncertainty quantification. Like VAEs, the FuXi-S2S model's encoder does not merely generate a static hidden feature from input data. Instead, it transforms input data into a Gaussian distribution in the latent space, which captures the probabilistic characteristics of the data, along with a static hidden feature. Then, the decoder combines samples from the Gaussian distribution with the static hidden feature to generate forecasts. This methodology

effectively captures the inherent uncertainty in the data, thereby enabling the generation of ensemble predictions under identical input conditions by repeatedly sampling from the Gaussian distribution. For better understanding, we draw analogies between these machine learning techniques and the conventional terminology in ensemble weather/subseasonal forecasting. In our model, the static hidden feature forms the basis for deterministic forecasts, while sampling from the Gaussian distribution serves as a perturbation module. This module introduces flow-dependent perturbations into the model's hidden feature, facilitating the generation of ensemble forecasts.

The FuXi-S2S model, illustrated in Fig. 5a, consists of three primary components: an encoder P, a perturbation module, and a decoder. The encoder, processing predicted weather parameters from two preceding time steps, with each time step representing 1 day, as FuXi-S2S is designed to forecast daily mean values. Specifically, it takes $\hat{X}^{t-1}$ and $\hat{X}^t$ as inputs into a two-dimensional (2D) convolution layer with a kernel size of two, which reduces the dimensions of the input data by half. Following this, the hidden feature h^t (with dimensions of $1536 \times 60 \times 120$) is derived from 12 repeated transformer blocks. The input to the encoder is a data cube that combines both upper-air and surface variables, with dimensions of $2 \times 76 \times 121 \times 240$. These dimensions represent two preceding time steps ($t-1$ and t), the number of input variables, and the latitude (H) and longitude (W) grid points, respectively. To account for the accumulation of forecast error over time, the forecast lead time (t) is also included in the encoder's input. Besides h^t, the encoder also generates a low-rank multivariate Gaussian distribution, $N(\Theta_p^t)$, characterized by a mean vector μ^t ($128 \times 60 \times 120$), a covariance matrix σ^t ($1536 \times 60 \times 120$), and a diagonal covariance matrix $diag^t$ ($128 \times 60 \times 120$). Intermediate perturbation vectors (z_p^t, dimension: $128 \times 60 \times 120$) are sampled from this Gaussian distribution ($N(\Theta_p^t)$). These vectors, after being weighted by a learned weight vector, yield the final perturbation vectors z^t (dimension: $1536 \times 60 \times 120$).

The decoder then processes the perturbed hidden features ($\tilde{h}^t = h^t + z^t$) through 24 transformer blocks and a fully connected layer, resulting in the final ensemble output $\hat{\mathbf{X}}^{t+1}$. The number of ensemble members generated equals the number of samples drawn from the Gaussian distribution $N(\Theta_p^t)$.

The FuXi-S2S model's training primarily focuses on constructing a Gaussian distribution that accurately represents the uncertainty in the model's predictions. A significant challenge in this process is the deviation of the Gaussian distribution derived from the model's predictions from the Gaussian distribution based on the target data, largely attributable to prediction errors. This challenge is addressed through knowledge distillation, which enables the transfer of information from real-world distributions to those predicted by the model. Within this framework, the encoder Q plays a crucial role, converting the target data into a Gaussian distribution. This distribution serves as a supervisor for the distribution generated by the encoder P, aiming to align both distributions closely by minimizing the Kullback–Leibler (KL) divergence loss (L_{KL}). This KL loss measures the discrepancy between the distributions predicted by both encoders. As illustrated in the Fig. 5b, during the training phase of the FuXi-S2S model, the encoder Q, which shares the network structure with the encoder P, processes a data cube containing target weather parameters from a preceding and the current time steps: $\mathbf{X}^t$ and $\mathbf{X}^{t+1}$. It predicts a low-rank multivariate Gaussian distribution ($N(\Theta_q^t)$) similar to the encoder P. Intermediate perturbation vectors are sampled from the encoder Q's distribution ($N(\Theta_q^t)$) during training (see Fig. 5b), and from the encoder P's distribution ($N(\Theta_p^t)$) during testing (see Fig. 5a). These vectors have dimensions of $128 \times 60 \times 120$. Additionally, an L1 loss is computed between the model's output ($\hat{\mathbf{X}}^{t+1}$) and the target $\mathbf{X}^{t+1}$. Therefore, the overall loss function at each autoregressive step is thus determined by the following equation:

$$L = \lambda L_{KL}(P^t, Q^t) + |\hat{\mathbf{X}}^{t+1} - \mathbf{X}^{t+1}| \tag{1}$$

where λ, a tune-able coefficient balancing L_{KL} and L1, is set to 1×10^{-4} in this study. The design of this loss function serves two purposes: the first term ensures the perturbation vector closely approximates the true data distribution, while the second term ensures the prediction unaffected by any perturbation vectors z^t.

In this study, we employ 51 ensemble members for subseasonal ensemble forecasting. As illustrated in Supplementary Fig. 15, the FuXi-S2S model, when enhanced with flow-dependent perturbations incorporated into its hidden features, demonstrates considerably improved forecast performance compared to the FuXi-S2S model that combines Perlin noise in the initial conditions with fixed perturbations added to the hidden features. Notably, the addition of Perlin noise results in only marginal improvements in forecast accuracy when the ensemble size is small. However, with larger ensemble sizes, such as the 51 members in this study, the addition of Perlin noise does not enhance forecast accuracy.

Similar to FuXi, we utilize an autoregressive, multi-step loss function to mitigate cumulative errors over long lead times, as outlined in Lam et al.[27]. The training process follows an autoregressive training regime and a curriculum training schedule, incrementally increasing the number of autoregressive steps from 1 to 17. Each autoregressive step undergoes 1000 gradient descent updates, resulting in a total number of 17,000 training steps. The training process utilizes 8 Nvidia A100 graphics processing units (GPUs), each employing a batch size of 1. Optimization is performed using the AdamW[89,90] optimizer with the following parameters: $\beta_1 = 0.9$ and $\beta_2 = 0.95$, an initial learning rate of 2.5×10^{-4}, and a weight decay coefficient of 0.1. The optimization hyperparameters used for training are summarized in Supplementary Table 1.

Saliency map

Recent developments in the field of XML have led to the emergence of various techniques[91], including saliency mapping. Saliency mapping quantifies the influence of a model's input on its output[46]. This method is characterized by the gradient intensities within the saliency maps; areas with higher gradients are considered critical by the model for making accurate predictions.

The generation of saliency maps primarily depends on backward propagation. This differs from standard model training as the propagation target can be adjusted depending on the specific goal of the analysis. Here, the saliency of the predicted anomaly relative to the input data is given by:

$$J(\mathbf{X}(c_o)) = -\sum_{i,j \in D} \frac{|f^n(\mathbf{X})(c_o, i, j) - \mu(c_o, i, j)|}{\sigma(c_o, i, j)} \tag{2}$$

$$S(c_i | c_o) = \frac{\partial J(\mathbf{X}(c_o))}{\partial \mathbf{X}(c_i)} \tag{3}$$

where f denotes the FuXi-S2S model and n is the number of forward steps, while μ and σ are the climatological mean and standard deviation, respectively. D specify the geographical area of interest. c_i and c_o represent the input and output variables. A well-trained model is expected to yield a saliency map that aligns well with the established physical understanding of weather systems. In our study, we construct an aggregated saliency map by averaging the individual maps generated from each of the 51 ensemble members.

Evaluation method

Prior to evaluation, each variable in the 42-day forecasts undergoes a detrending process to eliminate the linear trend. This step is essential for removing the linear long-term trends potentially affected by global warming[92]. For detrending, a linear regression model is fitted to estimate the weekly mean linear trend from both forecasts and observations over the hindcast period (2002–2016). For the testing period (2017–2021), this model takes the week of the year as input data to calculate the trend, which is then subtracted from both the forecasts and observations to obtain the detrended fields. Subsequently, the deterministic metrics of the ensemble mean are evaluated using the latitude-weighted TCC, which is calculated as follows:

$$TCC(c, \tau, i, j) = \frac{\sum_{t_0 \in D} \hat{\mathbf{A}}_{c,i,j}^{t_0 + \tau} \mathbf{A}_{c,i,j}^{t_0 + \tau}}{\sqrt{\sum_{t_0 \in D} (\hat{\mathbf{A}}_{c,i,j}^{t_0 + \tau})^2 \sum_{t_0 \in D} (\mathbf{A}_{c,i,j}^{t_0 + \tau})^2}} \tag{4}$$

where t_0 represents the forecast initialization time in the testing dataset D. H, and W denote the number of grid points in the latitude and longitude directions. The indices c, i, and j correspond to variables, latitude, and longitude coordinates, respectively. τ refers to the forecast lead time steps added to t_0. $\hat{\mathbf{A}}_{c,i,j}^{t_0 + \tau}$ and $\mathbf{A}_{c,i,j}^{t_0 + \tau}$ are the differences between the forecast or observation and the climatological mean, with the climatological mean derived from data spanning the years from 2002 and 2016.

To evaluate the ensemble forecast performance, we use the RPSS[51,52] which quantifies the comparison between the cumulative squared probability errors of a given forecast and a climatological forecast. The calculation of the RPSS metric necessitates prior determination of the ranked probability scores (RPS) for both the forecast ($RPS_{forecast}$) and the climatological forecast (RPS_{clim}) should be calculated first. The RPS aggregates the squared probability errors across K ($K = 3$ in this work) categories, such as tercile, arranged in ascending order. The tercile bounds are determined based on the average values over either 1-week or 2-week periods for each corresponding verification period. These calculations of tercile bounds are performed

separately for each forecast model and observation (ERA5 data). The metric assesses the accuracy with which the probability forecast predicts the actual observation category. The RPS score is derived from the sum of the squared differences between the cumulative categorical forecast probability and its observed counterpart, where $p_{O(i)} = 1$ denotes the observed category and $p_{O(i)} = 0$ represents other categories:

$$RPS_{forecast} = \sum_{k=1}^{K}(F_{forecast(k)} - F_{O(k)})\qquad(5)$$

$$RPS_{clim} = \sum_{k=1}^{K}(F_{clim(k)} - F_{O(k)})\qquad(6)$$

where $F_{forecast(k)} = \sum_{i=1}^{k} p_{forecast(i)}$, $F_{clim(k)} = \sum_{i=1}^{k} p_{clim(i)}$, $F_{O(k)} = \sum_{i=1}^{k} p_{O(i)}$ represent the kth components of the cumulative forecast, climatological, and observational distributions, respectively. And $p_{forecast(i)}$, $p_{clim(i)}$, $p_{O(i)}$ correspond to the forecasted, climatological, and observed probability of the event's occurrence in category i ($i \le k$). Crucially, the RPS is affected by both the forecast probabilities attributed to the observed category and the probabilities assigned to other categories. The RPS value varies between 0 and 1, where a lower value denotes a smaller forecast probability error, and thus a more accurate forecast. Specifically, a RPS value of 0 indicates a perfectly accurate categorical forecast. With the RPS values of both the forecast and the climatological forecast, the RPSS can be determined as:

$$RPSS = 1 - \frac{<RPS_{forecast}>}{<RPS_{clim}>}\qquad(7)$$

where the brackets <...> denote the average of the $RPS_{forecast}$ and RPS_{clim} values across all forecast-observation pairs. Since each forecast category is equally probable by design, the climatological forecast assumes a 33% probability of occurrence for each category. The RPSS metric serves a comparative measure against the climatological forecast. Its value ranges from $-\infty$ to 1, where 1 corresponds to a perfect forecast and higher values suggest better forecast performance. A positive RPSS value indicates superior accuracy over the climatological forecast, while a negative value suggests inferior accuracy. A value of zero suggests that the forecast has no added skill compared to the climatological forecast.

Additionally, we use the BSS[52] to evaluate the performance of extreme forecasts. The BSS, a widely used metric for assessing the quality of categorical probabilistic forecasts, can be considered as a special case of the RPSS with two forecast categories[93]. The BSS is computed using the following equation:

$$BSS = 1 - \frac{<BS_{forecast}>}{<BS_{clim}>}\qquad(8)$$

where $BS_{forecast}$ and BS_{clim} represent the Brier Scores (BS)[94] for the model's forecast and the climatological forecast, respectively. Similar to the RPS, the BS quantifies the mean squared difference between the predicted probabilities and observations (either 0 or 1) in binary probabilistic forecasts. In this study, the BSS is calculated for the ensemble mean of both FuXi-S2S and ECMWF S2S, using the 90th climatological percentiles as the threshold for extreme events. The BS ranges from 0 to 1, with lower values indicating a better agreement between ensemble forecasts and observations with 0 suggesting the best possible BS score. On the contrary, a higher BSS, up to a maximum of 1, indicates better performance. The BSS measures the improvement of a forecast's BS ($BS_{forecast}$) relative to that of a climatological forecast (BS_{clim}) as reference. A BSS of one indicates a perfect forecast, zero denotes no improvement over

climatology, and negative values suggest inferior performance compared to climatology.

The evolution of MJO is typically characterized using the Real-time Multivariate MJO (RMM) index, as originally developed by Wheeler and Hendon[66]. The RMM1 and RMM2 indices represent the first and second principal components of the combined EOF. This EOF is derived based on the daily mean values of OLR, zonal wind at 850 hPa (U850), and zonal wind at 200 hPa (U200), all averaged within the latitude range of 15° N and 15° S[95]. In this study, we use the EOFs derived by Wheeler and Hendon (2004)[66]. To obtain the predicted MJO indices, data from both the FuXi-S2S and ECMWF S2S models are firstly interpolated from a spatial resolution of 1.5°–2.5°, and projected onto the observed EOFs. After calculating the ensemble mean anomalies, the RMM for the ensemble mean of both modes was derived. The amplitude and phase of the MJO are respectively defined by the formulas: $RMMA = \sqrt{RMM1^2(t) + RMM2^2(t)}$ and $\theta = tan^{-1}\frac{RMM2^2(t)}{RMM1^2(t)}$. To assess the quality of the MJO forecasts, we calculate the bivariate COR using the following equation:

$$COR(\tau) = \frac{\sum_{t=1}^{N}[a_1(t)b_1(t,\tau) + a_2(t)b_2(t,\tau)]}{\sqrt{\sum_{t=1}^{N}[a_1^2(t) + a_2^2(t)]}\sqrt{\sum_{t=1}^{N}[b_1^2(t,\tau) + b_2^2(t,\tau)]}}\qquad(9)$$

where $a_1(t)$ and $a_2(t)$ are the observed RMM1 and RMM2 at time t derived from the ERA5 reanalysis dataset. Correspondingly, $b_1(t, \tau)$ and $b_2(t, \tau)$ represent the forecasts for time t with a lead time of τ days, respectively. N denotes the number of total predictions. We apply the threshold of COR = 0.5 for skillful prediction[95].

Additionally, we assessed the respective contributions of amplitude and phase to the prediction skills of the MJO by examining the COR and error metrics of ensemble mean forecasts for each component. The COR for amplitude ($COR_{amplitude}$) and phase (COR_{phase}) were calculated using the methods outlined by Wang et al.[96] as follows:

$$COR_{amplitude}(\tau) = \frac{\sum_{t=1}^{N} RMMA_{obs}(t) \times RMMA_{forecast}(t,\tau)}{\sqrt{\sum_{t=1}^{N} RMMA_{obs}^2(t)}\sqrt{\sum_{t=1}^{N} RMMA_{forecast}^2(t,\tau)}}\qquad(10)$$

$$COR_{phase}(\tau) = \frac{\sum_{t=1}^{N} RMMA_{obs}(t) \times cos(\theta_{forecast}(t,\tau) - \theta_{obs}(t))}{\sum_{t=1}^{N} RMMA_{obs}^2(t)}\qquad(11)$$

where $RMMA_{obs}$ and $RMMA_{forecast}$ represent the observed and predicted amplitudes of the MJO, respectively, while θ_{obs} and $\theta_{forecast}$ denote the observed and predicted phases. Additionally, we computed the average amplitude and phase errors ($ERROR_{amplitude}$ and $ERROR_{phase}$) as follows, based on the method described by Rashid et al.[95]:

$$ERROR_{amplitude}(\tau) = \frac{1}{N}\sum_{t=1}^{N}(RMMA_{forecast}(t,\tau) - RMMA_{obs}(t))\qquad(12)$$

$$ERROR_{phase}(\tau) = \frac{1}{N}\sum_{t=1}^{N} tan^{-1}\left(\frac{a_1(t)b_2(t,\tau) - a_2(t)b_1(t,\tau)}{a_1(t)b_1(t,\tau) + a_2(t)b_2(t,\tau)}\right)\qquad(13)$$

Further details about the COR and ERROR for the amplitude and phase are presented in the Supplementary Fig. 9.

Atmospheric predictability exhibits significant day-to-day variability, which in turn affects the potential accuracy of weather forecasts. To determine whether FuXi-S2S consistently outperforms ECMWF S2S despite this variability, we adopted a bootstrapping approach for significance testing. This method involves generating a large number of synthetic datasets, for example, 1000 in this work. For each day within these datasets, a forecast is randomly selected from either model A or model B. The forecast skill of each synthetic dataset is then evaluated

by comparing it with actual observation. If the performance of model A surpasses the 97.5th percentile of the skill distribution derived from the synthetic datasets, it can be considered "significantly better" than model B. In contrast, if its performance falls below the 2.5th percentile, it is regarded as "significantly worse". We also analyzed where the FuXi-S2S and ECMWF S2S models are significantly better or worse than the climatological forecasts, with model B representing these forecasts. Throughout the paper, significance testing has been applied to all bar plots and spatial maps of statistical metrics. For all the bar plots in the paper, a pale color is used when the FuXi-S2S model does not show a statistically significant improvement over the ECMWF S2S model. Additionally, we have marked areas on all spatial maps where the skill score is statistically significant with stippling.

Data availability

We downloaded a subset of the daily statistics from the ERA5 hourly data from the official website of Copernicus Climate Data (CDS) at https://cds.climate.copernicus.eu/cdsapp#!/software/app-c3s-daily-era5-statistics. The ECMWF S2S data were obtained from https://apps.ecmwf.int/datasets/data/s2s/. The 1° CPC OLR data are provided by the NOAA Physical Sciences Laboratory (PSL) from their website of https://psl.noaa.gov. Rainfall data from the Global Precipitation Climatology Project (GPCP) was obtained from the National Oceanic and Atmospheric Administration (NOAA), specifically the National Centers for Environmental Information (NCEI), which is accessible at https://www.ncei.noaa.gov/products/global-precipitation-climatology-project. The relevant data from each figure in the main manuscript and in the Supplementary Information are provided in https://zenodo.org/records/12662702[97].

Code availability

The source code employed for training and running FuXi-S2S models in this research is accessible within a specific Google Drive folder (https://drive.google.com/drive/folders/1z47CRQdKFZaOjtKQWSNZobC1_RePUVIK?usp=sharing)[98]. As the FuXi-S2S model and code are essential resources for this study. Currently, access to these resources is limited. For inquiries and access to the Google Drive link kindly reach out to Li Hao at the following email address: lihao_lh@fudan.edu.cn. Calculation of MJO index is based on the EOFs derived by Wheeler and Hendon (2004)[66]. The implementation of Perlin noise is based on publicly available from the GitHub repository: https://github.com/pvigier/perlin-numpy.

References

1. National Academies of Sciences. *Next generation earth system prediction: strategies for subseasonal to seasonal forecasts* (National Academies Press, Washington, DC, 2016).
2. White, C. J. et al. Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorol Appl.* **24**, 315–325 (2017).
3. Pegion, K. et al. The subseasonal experiment (subx): a multimodel subseasonal prediction experiment. *Bull. Am. Meteorol Soc.* **100**, 2043–2060 (2019).
4. White, C. J. et al. Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Am. Meteorol Soc.* **103**, E1448–E1472 (2022).
5. Domeisen, D. I. et al. Advances in the subseasonal prediction of extreme events: relevant case studies across the globe. *Bull. Am. Meteorol Soc.* **103**, E1473–E1501 (2022).
6. Lorenz, E. N. Forced and free variations of weather and climate. *J. Atmos. Sci.* **36**, 1367 – 1376 (1979).
7. Mariotti, A., Ruti, P. M. & Rixen, M. Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *NPJ Clim. Atmos. Sci.* **1**, 4 (2018).
8. Weyn, J. A., Durran, D. R., Caruana, R. & Cresswell-Clay, N. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.* **13**, e2021MS002502 (2021).
9. Han, J.-Y., Kim, S.-W., Park, C.-H. & Son, S.-W. Ensemble size versus bias correction effects in subseasonal-to-seasonal (s2s) forecasts. *Geosci. Lett.* **10**, 37 (2023).
10. Vitart, F. Evolution of ECMWF sub-seasonal forecast skill scores. *Q. J. R. Meteorol Soc.* **140**, 1889–1899 (2014).
11. Saha, S. et al. The NCEP climate forecast system version 2. *J. Clim.* **27**, 2185–2208 (2014).
12. Vitart, F. et al. The subseasonal to seasonal (s2s) prediction project database. *Bull. Am. Meteorol Soc.* **98**, 163–173 (2017).
13. Nowak, K., Webb, R., Cifelli, R. & Brekke, L. Sub-seasonal climate forecast rodeo. In *Proc. 2017 AGU Fall Meeting,* 11–15 (*New Orleans, LA,* 2017).
14. Monhart, S. et al. Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.* **123**, 7999–8016 (2018).
15. Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K. & Mackey, L. Improving subseasonal forecasting in the western us with machine learning. In *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 2325–2335 (2019).
16. Vitart, F. et al. Outcomes of the WMO prize challenge to improve subseasonal to seasonal predictions using artificial intelligence. *Bull. Am. Meteorol Soc.* **103**, E2878–E2886 (2022).
17. Mouatadid, S. et al. Adaptive bias correction for improved sub-seasonal forecasting. *Nat. Commun.* **14**, 3482 (2023).
18. Domeisen, D. I. et al. Advances in the subseasonal prediction of extreme events: relevant case studies across the globe. *Bull. Am. Meteorol Soc.* **103**, 1473–1501 (2022).
19. Demaeyer, J., Penny, S. G. & Vannitsem, S. Identifying efficient ensemble perturbations for initializing subseasonal-to-seasonal prediction. *J. Adv. Model. Earth Syst.* **14**, 1–30 (2022).
20. Buizza, R., Milleer, M. & Palmer, T. N. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol Soc.* **125**, 2887–2908 (1999).
21. Buizza, R. Introduction to the special issue on "25 years of ensemble forecasting". *Q. J. R. Meteorol Soc.* **145**, 1–11 (2019).
22. Leutbecher, M. Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteorol Soc.* **145**, 107–128 (2019).
23. Cohen, J. et al. S2s reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev. Clim. Change* **10**, e00567 (2019).
24. Richardson, D. S. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol Soc.* **127**, 2473–2489 (2001).
25. Hu, Y., Chen, L., Wang, Z. & Li, H. SwinVRNN: A data-driven ensemble forecasting model via learned distribution perturbation. *J. Adv. Model. Earth Syst.* **15**, e2022MS003211 (2023).
26. Kurth, T. et al. FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators. In *Proc. of the Platform for Advanced Scientific Computing Conference* (2022).
27. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, eadi2336 (2023).
28. Bi, K. et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023).
29. Chen, L. et al. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *Npj Clim. Atmos. Sci.* **6**, 1–11 (2023).
30. Zhong, X. et al. Fuxi-extreme: improving extreme rainfall and wind forecasts with diffusion model (2023).
31. Nguyen, T. et al. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. Preprint at https://arxiv.org/abs/2312.03876 (2023).

32. Haiden, T. et al. Evaluation of ECMWF forecasts, including the 2021 upgrade. ECMWF Technical Memorandum No. 884 (European Centre for Medium-Range Weather Forecasts, 2021). https://doi.org/10.21957/90pgicjk4.

33. Wang, C. et al. Coupled ocean-atmosphere dynamics in a machine learning earth system model. Preprint at https://arxiv.org/abs/2406.08632 (2024).

34. He, S., Li, X., DelSole, T., Ravikumar, P. & Banerjee, A. Sub-seasonal climate forecasting via machine learning: challenges, analysis, and advances. *Proc. AAAI Conf. Artif. Intell.* **35**, 169–177 (2021).

35. Kiefer, S. M., Lerch, S., Ludwig, P. & Pinto, J. G. Can machine learning models be a suitable tool for predicting central European cold winter weather on subseasonal to seasonal time scales? *Artif. Intell. Earth Syst.* **2**, 1–16 (2023).

36. Molteni, F., Buizza, R., Palmer, T. N. & Petroliagis, T. The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol Soc.* **122**, 73–119 (1996).

37. de Andrade, F., Coelho, C. A. & Cavalcanti, I. F. Global precipitation hindcast quality assessment of the subseasonal to seasonal (s2s) prediction project models. *Clim. Dyn.* **52**, 5451–5475 (2019).

38. Madden, R. A. & Julian, P. R. Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *J. Atmos. Sci.* **28**, 702–708 (1971).

39. Madden, R. A. & Julian, P. R. Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.* **29**, 1109–1123 (1972).

40. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).

41. McGovern, A. et al. Making the black box more transparent: understanding the physical implications of machine learning. *Bull. Am. Meteorol Soc.* **100**, 2175–2199 (2019).

42. Molnar, C., Casalicchio, G. & Bischl, B. Interpretable machine learning—a brief history, state-of-the-art and challenges. In *Proc. Joint European conference on machine learning and knowledge discovery in databases*, 417–431 (Springer, 2020).

43. Mamalakis, A., Ebert-Uphoff, I. & Barnes, E. A. Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. In *Proc. International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 315–339 (Springer, 2020).

44. Toms, B. A., Kashinath, K. & Yang, D. et al. Testing the reliability of interpretable neural networks in geoscience using the Madden–Julian oscillation. *Geosci. Model Dev.* **14**, 4495–4508 (2021).

45. Rasp, S. & Thuerey, N. Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: a new model for Weatherbench. *J. Adv. Model. Earth Syst.* **13**, e2020MS002405 (2021).

46. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proc. International Conference on Learning Representations* (2013).

47. Dunstone, N. et al. Windows of opportunity for predicting seasonal climate extremes highlighted by the Pakistan floods of 2022. *Nat. Commun.* **14**, 6544 (2023).

48. Faghmous, J. H. & Kumar, V. A big data guide to understanding climate change: the case for theory-guided data science. *Big Data* **2**, 155–163 (2014).

49. Karpatne, A. et al. Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331 (2017).

50. Chattopadhyay, A. & Hassanzadeh, P. Long-term instability of deep learning-based digital twins of the climate system: Cause and solution. In *Proc. APS March Meeting Abstracts* (2023).

51. Epstein, E. S. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol* **8**, 985–987 (1969).

52. Wilks, D. S. *Statistical methods in the atmospheric sciences* 3rd edn, Vol. 100 (2011).

53. Vitart, F. & Robertson, A. W. The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *NPJ Clim. Atmos. Sci.* **1**, 3 (2018).

54. Merryfield, W. J. et al. Current and emerging developments in subseasonal to decadal prediction. *Bull. Am. Meteorol Soc.* **101**, E869–E896 (2020).

55. Zhang, C. Madden–Julian oscillation. *Rev. Geophys.* **43** (2005).

56. Zhang, C. Madden–Julian oscillation: bridging weather and climate. *Bull. Am. Meteorol Soc.* **94**, 1849–1870 (2013).

57. Zhang, C. et al. Cracking the MJO nut. *Geophys. Res. Lett.* **40**, 1223–1230 (2013).

58. Neena, J. et al. Predictability of the Madden–Julian oscillation in the intraseasonal variability hindcast experiment (ISVHE). *J. Clim.* **27**, 4531–4543 (2014).

59. Kim, H., Vitart, F. & Waliser, D. E. Prediction of the Madden–Julian oscillation: a review. *J. Clim.* **31**, 9425–9443 (2018).

60. Jiang, X. et al. Fifty years of research on the madden-julian oscillation: Recent progress, challenges, and perspectives. *J. Geophys. Res. Atmos.* **125**, e2019JD030911 (2020).

61. Wu, J. & Jin, F.-F. Improving the MJO forecast of s2s operation models by correcting their biases in linear dynamics. *Geophys. Res. Lett.* **48**, 1–10 (2021).

62. Silini, R. et al. Improving the prediction of the Madden–Julian oscillation of the ECMWF model by post-processing. *Earth Syst. Dyn.* **13**, 1157–1165 (2022).

63. Kim, H., Ham, Y. G., Joo, Y. S. & Son, S. W. Deep learning for bias correction of MJO prediction. *Nat. Commun.* **12**, 3087 (2021).

64. Silini, R., Barreiro, M. & Masoller, C. Machine learning prediction of the Madden–Julian oscillation. *NPJ Clim. Atmos. Sci.* **4**, 57 (2021).

65. Delaunay, A. & Christensen, H. M. Interpretable deep learning for probabilistic MJO prediction. *Geophys. Res. Lett.* **49**, e2022GL098566 (2022).

66. Wheeler, M. C. & Hendon, H. H. An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. *Mon. Weather Rev.* **132**, 1917–1932 (2004).

67. Wallace, J. M. & Gutzler, D. S. Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Weather Rev.* **109**, 784–812 (1981).

68. Mo, K. C. & Ghil, M. Statistics and dynamics of persistent anomalies. *J. Atmos. Sci.* **44**, 877–902 (1987).

69. Zhu, B. & Wang, B. The 30-60-day convection seesaw between the tropical Indian and western Pacific oceans. *J. Atmos. Sci.* **50**, 184–199 (1993).

70. Walker, G. T. Correlations in seasonal variations of weather. viii, a further study of world weather. *Men. Indian Meteor. Dept.* **24**, 275–332 (1924).

71. Huang, R. H. Influence of the heat source anomaly over the western tropical Pacific for the subtropical high over East Asia. In *Proc. International Conference on the General Circulation of East Asia. Chendu, China, April 10–15, 1987,* 40–50 (1987).

72. Savarin, A. & Chen, S. S. Pathways to better prediction of the MJO: 2. impacts of atmosphere-ocean coupling on the upper ocean and MJO propagation. *J. Adv. Model. Earth Syst.* **14**, e2021MS002929 (2022).

73. Hong, C.-C. et al. Causes of 2022 Pakistan flooding and its linkage with China and Europe heatwaves. *NOJ Clim. Atmos. Sci.* **6**, 163 (2023).

74. Yang, R. et al. Interpretable machine learning for weather and climate prediction: a survey. Preprint at https://arxiv.org/abs/2403.18864 (2024).

75. Haiden, T. et al. Evaluation of ECMWF forecasts, including the 2018 upgrade. ECMWF Technical Memorandum No. 831 (European

Centre for Medium-Range Weather Forecasts, 2018). https://doi.org/10.21957/ldw15ckqi.

76. Nogueira, M. Inter-comparison of era-5, era-interim and GPCP rainfall over the last 40 years: process-based analysis of systematic and random differences. *J. Hydrol.* **583**, 1–17 (2020).

77. Lavers, D. A., Simmons, A., Vamborg, F. & Rodwell, M. J. An evaluation of era5 precipitation for climate monitoring. *Q. J. R. Meteorol. Soc.* **148**, 3152–3165 (2022).

78. Hersbach, H. et al. The era5 global reanalysis. *Q. J. R. Meteorol Soc.* **146**, 1999–2049 (2020).

79. Adler, R. F. et al. The global precipitation climatology project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. *Atmosphere* **9**, 138 (2018).

80. Albers, J. R. & Newman, M. Subseasonal predictability of the North Atlantic oscillation. *Environ. Res. Lett.* **16**, 1–10 (2021).

81. Yan, Y., Liu, B. & Zhu, C. Subseasonal predictability of South China Sea summer monsoon onset with the ECMWF s2s forecasting system. *Geophys. Res. Lett.* **48**, e2021GL095943 (2021).

82. Richter, J. H. et al. Quantifying sources of subseasonal prediction skill in cesm2. *NPJ Clim. Atmos. Sci.* **7**, 59 (2024).

83. Stan, C. et al. Advances in the prediction of MJO teleconnections in the s2s forecast systems. *Bull. Am. Meteorol Soc.* **103**, E1426–E1447 (2022).

84. Cho, K., Van Merriënboer, B., Bahdanau, D. & Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (2014).

85. Olivetti, L. & Messori, G. Advances and prospects of deep learning for medium-range extreme weather forecasting. *EGUsphere* **2023**, 1–20 (2023).

86. Doersch, C. Tutorial on variational autoencoders. *Preprint at* https://arxiv.org/abs/1606.05908 (2016).

87. Zhao, T., Zhao, R. & Eskenazi, M. Learning discourse-level diversity for neural dialog models using conditional variational auto-encoders. *Preprint at* https://arxiv.org/abs/1703.10960 (2017).

88. Kingma, D. P. & Welling, M. An introduction to variational auto-encoders. *Found. Trends Mach. Learn.* **12**, 307–392 (2019).

89. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations* (2015).

90. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations* (2017).

91. Samek, W., Wiegand, T. & Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. Preprint at https://arxiv.org/abs/1708.08296 (2017).

92. Weirich-Benet, E. et al. Subseasonal prediction of central European summer heatwaves with linear and random forest machine learning models. *Artif. Intell. Earth Syst.* **2**, e220038 (2023).

93. Weigel, A. P., Liniger, M. A. & Appenzeller, C. The discrete brier and ranked probability skill scores. *Mon. Weather Rev.* **135**, 118–124 (2007).

94. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1 (1950).

95. Rashid, H. A., Hendon, H. H., Wheeler, M. C. & Alves, O. Prediction of the Madden–Julian oscillation with the Poama dynamical prediction system. *Clim. Dyn.* **36**, 649–661 (2011).

96. Wang, S., Sobel, A. H., Tippett, M. K. & Vitart, F. Prediction and predictability of tropical intraseasonal convection: Seasonal dependence and the maritime continent prediction barrier. *Clim. Dyn.* **52**, 6015–6031 (2019).

97. Chen, L. et al. A machine learning model that outperforms conventional global subseasonal forecast models (Version 1.0) [Figure Dataset]. Zenodo. (2024). https://zenodo.org/records/12662702.

98. Chen, L. et al. A machine learning model that outperforms conventional global subseasonal forecast models (Version 1.0) [Dataset] [Software]. Zenodo. (2023). https://zenodo.org/records/10402083.

Acknowledgements

Author contributions

H.L., X.Z, L.C., and B.L. designed the project. L.C. designed and performed the model training. X.Z. and L.C. performed the analysis under supervision of H.L., B.L., J.W., Q.C., L.W., C.L., Z.H., and Y.Q. X.Z. and L.C. wrote and revised the manuscript. J.W., D.C., and S.X. contributed to interpreting results and discussions of associated dynamics.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-50714-1.

Correspondence and requests for materials should be addressed to Hao Li, Bo Lu or Yuan Qi.

Peer review information *Nature Communications* thanks Hannah Christensen and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.