nature portfolio

Peer Review File

A machine learning model that outperforms conventional global subseasonal forecast models

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Abstract of the paper "Meet the challenge of predictability desert: a machine learning model that outperforms conventional global subseasonal forecast models"

The authors introduce FuXi Subseasonal-to-Seasonal (FuXi-S2S), a machine learning based subseasonal forecasting model that provides global daily mean forecasts for up to 42 days. The authors have done a great deal of impressive work and many aspects of the paper are potentially very important and of high interest. However, the manuscript requires some improvements. Some general comments are:

(1) In the "Introduction" of the paper, the authors must present more clearly the following aspects: the background, the methods, and the main findings, as in the basic form of the manuscript, the Introduction offers information related only to some of the aspects and even so, their delimitation is unclear.
(2) "Introduction" section is weak. My concern is that the text does not contribute to the research question. I recommend showing other cases in the literature, and identifying the knowledge gap.
(3) “Discussion” section: Authors should also address why this study is important in terms of adding new knowledge. Also, they should explain the contributions and physical meaning of their findings. I request authors to improve this section.
(4) Figure 5b is not discussed. This part requires improvement to enhance its clarity and understandability.
(5) The authors used multiple datasets with different sources and temporal resolutions. How do you merge all these datasets into a common time series? Please explain how the correspondence is established.
(6) Have pre-processing methods been applied to the dataset? For example, it is common practice to remove outliers in real experiments.
(7) The paper does not provide specific details on the hyperparameters used.
(8) Table 1: Why were these meteorological variables chosen as input parameters for the model? Not all of these parameters are directly related to rainfall. Would the results be improved by selecting only the meteorological parameters that are related to rainfall? I request that the authors explain and provide evidence for their choice.
(9) Description of the FuXi-S2S model is very limited: The figure is nice, but in fact, it is very hard to figure out what they did. A more detailed treatment, even in a supplement, would be very important.

Reviewer #2 (Remarks to the Author):

Meet the challenge of predictability desert: a machine learning model that outperforms conventional global subseasonal forecast models

Lei Chen and co-authors

The authors present a data-driven subseasonal forecasting model called FuXi-S2S. The model is probabilistic, producing ensemble forecasts. The skill of the model is compared to subseasonal forecasts from ECMWF, widely regarded to be the leading subseasonal forecasting centre worldwide. FuXi-S2S produces forecasts which generally match the skill of ECMWF – even slightly better for some variables. The authors go on to assess the skill of predicting the Madden-Julian Oscillation, showing that FuXi-S2S has very good skill here – extending the predictable window by 6 days. This is even though FuXi-S2S was not explicitly trained to produce MJO forecasts. They finally consider an example of an extreme flooding event, including an assessment of sources of predictability for that event. I must emphasise here that the subseasonal timescale is a very difficult timescale to address, so it is impressive that a data-driven model can produce skilful predictions, and even marginal improvements are of interest. This is the first (to my knowledge) data-driven S2S model that moves beyond predicting low-dimensional indices, so the work is of great significance. The technical approach taken to generate the ensemble is also very interesting, involving learning how to sample from the latent space optimally. I would recommend publication in Nature Communications, though I have two main criticisms for the paper, which should be addressed before being accepted.

Firstly, there is no significance testing. Some of the improvements are marginal, and it is important to assess which improvements are nevertheless significant. I include advice on how to do this testing below.

Secondly, the authors have a tendency to be slightly sensationalist in their language. Some of the improvements are minor. One gets the impression that the case studies and examples are cherry-picked to some extent to show off FuXi to the best advantage. This language doesn't seem necessary to me – we all know this is a very challenging timescale, and the results are impressive as they are. I have suggested specific changes to the wording in places below.

Major Comments

Significance testing.

The results are let down by a lack of significance testing throughout. Please rectify this. Each pair of bars in the bar charts (fig 1, SF1, 3, 4, 5, 8, 9, 11, 12) could be made bold for a pair where the difference is significant, and pale colouration if not significant.

It would be usual to show stippling on a map (Fig 2, SF 2, 7, 8) if the skill score is significant, and leave without stippling if it is not.

I include comments on how best to carry out this significance testing, in case they are of use.

A t-test is not appropriate here, and will likely indicate insignificant results. This is because you have paired data. The predictability of the atmosphere varies substantially from day such that the potential skill of a forecast can also vary substantially from day to day. The interesting thing is whether FuXi has slightly better skill than ECMWF for each day, even if the distributions of skill over all the days are very strongly overlapping because of the varying underlying predictability. A bootstrapping approach can account for this. To do this, create a large number (e.g. 1000) of synthetic datasets: to populate each day of the synthetic dataset, randomly choose between the forecast from model A and from model B for that day. Then assess the skill score for each of these 1000 synthetic datasets by comparing with observations. If the skill score for model A is greater than the 97.5% of the distribution of skill scores from the synthetic datasets then you could consider that model to be "significantly better" than model B, while less than 2.5% would be "significantly worse", for example.

For the maps, it would be good to see where FuXi-S2S is significantly better (or worse) than ECMWF, but also where each of the FuXi and ECMWF models are significantly better (or worse) than model B = climatology.

Sensationalist language.

I suggest making the following specific changes

P5 "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 marginally inferior performance at lead times of 6 and 5-6 weeks." Either include the qualifier "marginally" in front of both "superior" and "inferior" or in front of neither. The improvement and degradation are the same approximate size

P5 "Moreover, FuXi-S2S also outperforms ECMWF over a large part of extra-tropical regions for both T2M and Z500." Please add qualifier "though it is generally less skilful in the tropics".

P6 Section 2.2. Ensemble forecast. It's worth mentioning (especially following the significance testing) that the skill is negative compared to climatology for both ECMWF and FuXi over large regions of the globe, and particularly in the extra tropics. This indicates just how hard the problem is – that climatology can do better than either a dynamical or ML forecast model. So it's not just a case of "considerably higher skills in tropical regions" – it's more that there *is* skill in the tropics whereas there is no skill in the extra tropics.

P7 “Notably, FuXi-S2S shows consistently higher RPSS values than ECMWF S2S across all variables, namely TP, T2M, Z500, and OLR, especially over extratropical averages.” This is not true: TP in the tropics is more skilful for ECMWF than FuXi for lead times of 3, 4, 5 and 6 weeks, yet FuXi does better for the two-week averages. This is a notable and slightly confusing result. Why can FuXi do better on the 2-week averages than on the single weeks? The degradation seems to be dominated by the central-eastern Pacific, masking good performance elsewhere (e.g.as demonstrated later by the Pakistan flood case)

P11 “The ECMWF S2S forecasts gradually converge toward observations as the initialization dates approach the actual event. In contrast, FuXi-S2S exhibits superior forecast performance” this is an interesting phrasing, as the smooth convergence of ECMWF is actually quite a desirable property. Why do you see such jumpiness in the FuXi-S2S forecasts? What is the ensemble doing here? It would be interesting to see 25 and 75th percentiles of the ensemble for each start date for ECMWF and FuXi-S2S

Minor comments

P2 Introduction. This was comprehensive and easy to follow – very nice

P4 “It compares the performance of FuXi-S2S with that of the 11-member ECMWF S2S reforecasts from the model cycle C47r3”. Please add qualifier “over the same period”

P8 L7 “generally predicts more regions” -> “generally exhibits more regions”

P8 L8 “more areas with higher skills relative to climatological forecasts” -> “more areas with skill relative to climatological forecasts”

P9 paragraph 2. I found this description of the MJO evaluation confusing – more detail would be beneficial. In particular, clarify which multivariate EOFs you project onto (do all models use the observed EOFs as a basis? And clarify that you compute the ensemble mean anomalies first, and then compute the RMM of the ensemble mean (which I think is what you are doing, as opposed to computing the RMM of each ensemble member and then averaging).

P10 Fig 3 caption. Why is the RMM COR globally averaged and latitude weighted? It’s just an index.

P10 Para 1. I’d like to see the skill of FuXi-S2S compared to ECMWF at predicting the phase and skill at predicting RMM amplitude – perhaps as an extra supp mat figure. What is giving this improvement in skill? Sounds from this discussion like the amplitude is the main source of improvement.

P11 “Supplementary Figure 7” I found it hard to understand exactly what this figure showed. Are these composites given that the MJO starts in phase 4? Or is it composites given that it is in phase

four in each of week 3, or week 4, etc.

P13 the interpretation of precursors was very nice to see.

P16 On the creation of S2S anomalies. You describe the approach taken for ECMWF – what do you do for FuXi-S2S – do you also produce a set of hindcasts for FuXi and use them to construct a climatology to be subtracted off? This would be the fairest approach, allowing lead-time (and model) dependent biases to be removed from FuXi-S2S as they are for the ECMWF forecasts

P17 Section 4.2 What is the timestep used by FuXi-S2S – have I just missed this detail?

P19 L4 "Similar to FuXi, an autoregressive, we utilize an autoregressive ..." – typo? Remove the first "an autoregressive"

P20 Section 4.4 On the detrending – what exactly are you detrending? Is this across the 42-day forecasts? Or somehow over the hindcast period?

P20, 7 lines from the bottom "The tercile bounds are determined ..." – are these determined for each model separately? Please clarify

Reviewer #2 (Remarks on code availability):

I have checked that the record exists, but I have not attempted to download the code and run it, etc.

If the paper passes onto another round of reviews, I'm happy to ask someone in my group to check the code.

Reviewer #3 (Remarks to the Author):

1 Summary of the Content

In this manuscript, the authors explore the use a vision transformer-based model (FuXi-S2S) to predict the 24-hour averaged 3-dimensional atmosphere with emphasis on the 2-6 week forecast range, commonly known as subseasonal-to-seasonal (S2S) time frame. Presently, the use of data-driven models for the S2S time frame have not been well explored compared to the medium-range (1-14 day) time frame. Using a novel probabilistic encoder-processor-decoder architecture (note this is different than the original FuXi model), FuXi-S2S is able to achieve skillful S2S forecasts. FuXi-S2S uses its inherit probabilistic nature to generate ensemble members, of which are used to quantify forecast uncertain and provide an ensemble mean forecast. When compared to a state-of-the-art numerical-based S2S system (ECMWF S2S), FuXi-S2S outperforms it for several deterministic (temporal anomaly correlation coefficient) and probabilistic metrics (Ranked

Probability skill Score and Brier Skill Score) for a select number of variables, mainly total precipitation and outgoing long-wave radiation (OLR).

Another strength of FuXi-S2S is the ability to predict the Madden–Julian Oscillation (MJO) with skill out to 36 days, an improvement of 6 days over the ECMWF S2S system. FuXi-S2S includes several new variables compared to the original FuXi model including 100-meter winds and OLR. Additionally, to the best of my knowledge, this is the first data-driven weather model that explicitly aims to machine learn the coupling between the sea surface temperature field and atmosphere. Overall, the manuscript is robust and represents an exciting step for data-driven S2S.

2 Overall Feedback

This paper demonstrates an important step for date-driven weather forecasting as the S2S time frame has long been difficult for NWP-based models. My only major concern is the assessment of the model quality is somewhat overstated. There are a few exaggerations which if toned down will improve the paper.

1. As mentioned above, the FuXi-S2S improvement over NWP is overstated in some sections. First the horizontal resolution is 1.5 degrees, which is several time more coarse than other state-of-the-art data driven models (Graphcast, Pangu, original FuXi) and the ECWMF S2S system. While the results are promising and for at least the show variables and metrics in this paper outperforming ECWMF's S2S system, it is worth noting
ECWMF's S2S is run at significantly high resolution. It is also worth noting for the total precipitation (TP) metrics, ERA5 is used as verification for the results shown in Figure 1 and 2. There are known disagreements between ERA5 TP and observations (Lavers et al., 2022). If observations or GCPC products were used at the verification the results in the paper may change.

a) "FuXi-S2S forecasts demonstrate superior deterministic and ensemble forecasts." The only comparisons in the paper are ensemble means using deterministic and probabilistic metrics using the ensemble forecasts. Neither of these are deterministic forecasts (e.g., the ensemble mean is not a deterministic forecast). To show that has better deterministic forecasts, a single forecast for a date using FuXi-S2S would need to be compared to other models like IFS, GraphCast, Pangu, etc. There are a number of references to the ensemble mean as a deterministic forecast I suggest rewording them.

b) "Our results demonstrated that FuXi-S2S comprehensively surpasses ECMWF S2S in forecast accuracy". This paper does not show that. Further diagnostics like the score-card used in GraphCast's paper would need to be shown. There only 4 variables are shown I would tone down this sentence.

c) "FuXi-S2S has shown remarkable 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 is just one case study. I suggest tuning down this sentence (e.g. remove remarkable).

d) “Compared to conventional NWP ensemble forecasting methods, which often struggle with constructing initial condition perturbations due to the complexities of multi-variate interactions and the need to maintain dynamic balance and ensemble spread in simulations [37], our approach of introducing perturbations directly into the model’s latent space offers a novel and effective alternative.”. This raises a good question, how is the ensemble spread in FuXi-S2S? A skill to spread ratio or something similar would answer that question. If ensemble system is very under or over disruptive I suggest toning down the language about how well this machine learned method of perturbations is compared to traditional methods used for NWP. I will also note, NWP ensemble generation has come a long way since the 1996
(the year the paper referenced in this sentence was published).

2. A brief discussion of the trades offs between the training of the original FuXi architecture (e.g., U-Transformer) vs the model in the paper would be beneficial to the reader. I assume there is some trade off training this architecture compared to standard U-transformer. For example does this architecture require longer training time, consumer more GPU memory, require any model-parallelisms, etc?

3. In my opinion, one of the most novel portions of this research in the prediction of sea surface temperatures (SST). I think including an assessment of the skill of SST forecasts (e.g. ensemble mean RMSE as a function of lead time) would be an interesting addition to the manuscript, either in the main paper or in the
supplementary materials.

4. Related to my previous comment, the sea surface temperature field plays a large role for atmospheric modeling in the S2S time frame. There is also emerging research suggestion a relationship between the MJO and SST (Savarin and Chen, 2022). In future work wok it would be interesting to see how the MJO forecast performance changes if the SST field was persisted or not included at all.

5. A known problem issue with date-drive weather models is spectra biases (Chattopadhyay and Hassanzadeh, 2023) (also see https://sites.research.google/weatherbench/spectra/). This is particularly problematic for vision transformers including the original FuXi model. I am curious if the authors can comment if there have observed similar artifacts in this model. If not that would be in great interest to the readers and field.

6. The authors did a great job with the saliency map and the associated discussion. I am excited to see this line of research in future work for ML-based weather forecasting.

3 Specific Remarks
• “Recent advancements in machine learning for medium- range weather forecasting [25–30] 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 [31]”. I suggest adding Nguyen et. al. 2023 (https://arxiv.org/abs/2312.03876) as they have similar resolution and are an ensemble

system that shows forecasts out to 15 days.
• Section 4.1 paragraph 2. I do not understand the motivation to use a different OLR product to calculate MJO when ERA5 is used for training. Especially ERA5 was used to verify other portions of the paper.
• “Specifically, it can complete a comprehensive 42-day forecast with daily time steps in approximately 7 seconds using an Nvidia A100 GPU. ” Is this a single member or a full ensemble?
• “Each autogressive step undergoes 1000 iterations of training.” 1000 epochs or 1000 training steps?
• Section 2.1 title “Deterministic forecast”. This section title is misleading, the forecast is not deterministic (e.g. a single forecast) it is the ensemble mean. I suggest changing this to something like “Deterministic Metrics”.
• Section 2.2 title “Ensemble forecast”. This section title should probably be changed to “Probabilistic metrics” to better reflect proposed revision to section 2.1.
• Figure 4. The current date formatting (MMDD) is unconventional and quite difficult to read. I highly recommend change to MM-DD or even MM-DD-YYYY. It took me a minute to even realize they were dates.
• Table 1. There is reference to both outgoing longwave radiation (OLR) and top net thermal radiation (TTR) as variables being predicted. It is my understanding that negative TTR is the exact same as OLR. What is the reason both are included in this table? It implies they are different.
2

References

Chattopadhyay, A., and P. Hassanzadeh, 2023: Long-term instabilities of deep learning-based digital twins of the climate system: The cause and a solution. 2304.07029.

Lavers, D. A., A. Simmons, F. Vamborg, and M. J. Rodwell, 2022: An evaluation of era5 precipitation for climate monitoring. Quarterly Journal of the Royal Meteorological Society, 148 (748), 3152–3165, doi:https: //doi.org/10.1002/qj.4351, URL https://doi.org/10.1002/qj.4351.

Savarin, A., and S. S. Chen, 2022: Pathways to better prediction of the mjo: 2. impacts of atmosphere-ocean coupling on the upper ocean and mjo propagation. Journal of Advances in Modeling Earth Systems, 14 (6), e2021MS002 929, doi:https://doi.org/10.1029/2021MS002929, URL https://doi.org/10.1029/2021MS002929.

Reviewer #3 (Remarks on code availability):

I did not see an input.nc file in the Zenodo project. The code looks runnable besides that.

Reply to Reviewer #1

We appreciate the reviewer's meticulous review and valuable feedback. We have made amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Abstract of the paper "Meet the challenge of predictability desert: a machine learning model that outperforms conventional global subseasonal forecast models"

The authors introduce FuXi Subseasonal-to-Seasonal (FuXi-S2S), a machine learning based subseasonal forecasting model that provides global daily mean forecasts for up to 42 days. The authors have done a great deal of impressive work and many aspects of the paper are potentially very important and of high interest. However, the manuscript requires some improvements. Some general comments are:

(1) In the "Introduction" of the paper, the authors must present more clearly the following aspects: the background, the methods, and the main findings, as in the basic form of the manuscript, the Introduction offers information related only to some of the aspects and even so, their delimitation is unclear.

Response:

We are grateful for the reviewer's suggestion to delineate the different aspects in the "Introduction" more distinctly. In response, we have restructured and elaborated the "**Introduction**" as follows.

The **background** is presented in the first two paragraphs of the Introduction. In this section, we highlight the significance of subseasonal forecasting, the existing gap between necessity and capability, and the limitations of traditional NWP models.

The methodology is elaborated in the third paragraph of the Introduction, which has been extensively revised to provide details on the main configurations of our FuXi-S2S model, with a special focus on the innovative perturbation module.

The key findings are summarized in the concluding paragraph, which includes the comprehensive statistical metrics, predictions for the 2022 Pakistan floods and the ability of the FuXi-S2S model to identify potential precursor signals to these events.

(2) "Introduction" section is weak. My concern is that the text does not contribute to the research question. I recommend showing other cases in the literature, and identifying the knowledge gap.

Response:

As suggested by the reviewer, we have reworked the sentences in the "Introduction" section to accentuate the existing knowledge gaps: "*Machine learning models have made significant strides in medium-range weather forecasting, but their success in*

subseasonal forecasting has been less pronounced [8, 33, 34]. 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, leading to a rapid reduction in ensemble spread. The inadequate representation of the complexities limits the performance of these prior machine learning-based subseasonal forecasting models, which does not yet rival that of traditional EPS based on NWP models."

(3) "Discussion" section: Authors should also address why this study is important in terms of adding new knowledge. Also, they should explain the contributions and physical meaning of their findings. I request authors to improve this section.

Response:

As suggested by the reviewer, we expanded "Discussion" section as follows:
"*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 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 [73]*"
Reference:
[73] Yang, R., et al.: Interpretable Machine Learning for Weather and Climate Prediction: A Survey. Preprint at https://arxiv.org/abs/2403.18864 (2024)

(4) Figure 5b is not discussed. This part requires improvement to enhance its clarity and understandability.

Response:

We are thankful for the reviewer's feedback concerning the clarity and comprehensibility of Figure 5b. In response to this, we have incorporated additional sentences (listed below) into the discussion of Figure 5b in Section 4.2. This will enhance the explanation and understanding of the figure. We hope this addresses the reviewer's concern effectively.

"*As illustrated in the Figure 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:* X_t *and* X_{t+1}*.*"

"Intermediate perturbation vectors are sampled from the encoder Q's distribution (N(Θtq)) during training (see Figure 5b), …"

(5) The authors used multiple datasets with different sources and temporal resolutions. How do you merge all these datasets into a common time series? Please explain how the correspondence is established.

Response:

As per the reviewer's suggestion, we have clarified our data usage in the study. The primary dataset utilized in this study is the ERA5 reanalysis dataset. More specifically, we employed the daily statistics derived from the 1-hourly ERA5 dataset, resulting in a temporal resolution of 1 day. **The ERA5 dataset served as the sole data source for training the FuXi-S2S model.** We have included an additional sentence in Section 4.1 to reflect this information:

"*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.*"

We value the reviewer's feedback and have clarified our use of additional datasets. Other datasets, such as the satellite-observed outgoing longwave radiation (OLR) dataset and the GPCP precipitation dataset, are utilized solely for verification purposes. We have also included an additional sentence in Section 4.1 to describe the temporal resolution of CBO data:

"*In our research, we employ the CBO data, which has a spatial resolution of 1° and a temporal resolution of 1 day. This data serves the ground truth for OLR in the identification and verification of MJO events.* "

(6) Have pre-processing methods been applied to the dataset? For example, it is common practice to remove outliers in real experiments.

Response:

We applied the z-score preprocessing method to the dataset for data normalization. The ERA5 reanalysis data, which has undergone a rigorous quality control process to address outliers, is employed for model training. We added the following sentences in the subsection 4.1 "Data":

"*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."

(7) The paper does not provide specific details on the hyperparameters used.

Response:

We appreciate the feedback and would like to clarify that the details concerning the hyperparameters used have now been provided at the end of Section 4.2 as follows:

"*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. The training process utilizes 8 Nvidia A100 graphics processing units (GPUs), each employing a batch size of 1. Optimization is performed using the AdamW [87, 88] optimizer with the following parameters: $\beta1=0.9$ and $\beta2=0.95$, an initial learning rate of $2.5\times10-4$, and a weight decay coefficient of 0.1. The optimization hyperparameters used for training are summarised in Supplementary Table 1.*"

Also, we have added a table in the supplementary material:

Supplementary Table 1: Model training hyperparameters

Optimizer	AdamW
LR decay schedule	Cosine
Number of GPUs used	8
Batch size	1 per GPU
Peak LR	2.5e-4
Weight decay	0.1
Total training steps	17,000

(8) Table 1: Why were these meteorological variables chosen as input parameters for the model? Not all of these parameters are directly related to rainfall. Would the results be improved by selecting only the meteorological parameters that are related to rainfall? I request that the authors explain and provide evidence for their choice.

Response:

We selected certain meteorological variables that are useful for applications beyond rainfall prediction. To clarify, we have added more explanations and motivations for the choice of input and output parameters in the model. The following sentences have been added to subsection 4.1, 'Data':

"*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 [79–81]. OLR was selected due to its significance in representing MJO events through OLR anomalies.*"

[79] Albers, J.R., Newman, M.: Subseasonal predictability of the north Atlantic oscillation. Environmental Research Letters 16(4), 1–10 (2021)
[80] Yan, Y., Liu, B., Zhu, C.: Subseasonal predictability of south china sea summer monsoon onset with the ecmwf s2s forecasting system. Geophysical Research Letters 48(24), 2021–095943 (2021)
[81] Richter, J.H., et al.: Quantifying sources of subseasonal prediction skill in cesm2. npj Climate and Atmospheric Science 7(1), 59 (2024)

(9) Description of the FuXi-S2S model is very limited: The figure is nice, but in fact, it is very hard to figure out what they did. A more detailed treatment, even in a supplement, would be very important.

Response:

Following the reviewer's suggestion, a detailed description of the FuXi-S2S model has been added to Section 4.2 as follows:

"*The FuXi-S2S model, illustrated in Figure 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 one 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 2, which reduces the dimensions of the input data by half. Following this, the hidden feature* h^t *(with dimensions of 1536 × 60 × 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 × 76 × 121 × 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 ht, the encoder also generates a low-rank multivariate Gaussian distribution,* $N(\Theta^t_p)$*, characterized by a mean vector* μ^t *(128 × 60 × 120), a covariance matrix* σ^t *(1536 × 60 × 120), and a diagonal covariance matrix* $diag^t$ *(128 × 60 × 120). Intermediate perturbation vectors (*z^t_p*, dimension: 128 × 60 × 120) are sampled from this Gaussian distribution (*$N(\Theta^t_p)$*). These vectors, after being weighted by a learned weight vector, yield the final perturbation vectors zt (dimension: 1536×60×120). 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* $\hat{X}^{t+1}$*. The number of ensemble members generated equals the number of samples drawn from the Gaussian distribution* $N(\Theta^t_p)$*.*"

Reply to Reviewer #2

We appreciate the reviewer's meticulous review and valuable feedback. We have made amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

The authors present a data-driven subseasonal forecasting model called FuXi-S2S. The model is probabilistic, producing ensemble forecasts. The skill of the model is compared to subseasonal forecasts from ECMWF, widely regarded to be the leading subseasonal forecasting centre worldwide. FuXi-S2S produces forecasts which generally match the skill of ECMWF – even slightly better for some variables. The authors go on to assess the skill of predicting the Madden-Julian Oscillation, showing that FuXi-S2S has very good skill here – extending the predictable window by 6 days. This is even though FuXi-S2S was not explicitly trained to produce MJO forecasts. They finally consider an example of an extreme flooding event, including an assessment of sources of predictability for that event. I must emphasise here that the subseasonal timescale is a very difficult timescale to address, so it is impressive that a data-driven model can produce skilful predictions, and even marginal improvements are of interest. This is the first (to my knowledge) data-driven S2S model that moves beyond predicting low-dimensional indices, so the work is of great significance. The technical approach taken to generate the ensemble is also very interesting, involving learning how to sample from the latent space optimally. I would recommend publication in Nature Communications, though I have two main criticisms for the paper, which should be addressed before being accepted.

Firstly, there is no significance testing. Some of the improvements are marginal, and it is important to assess which improvements are nevertheless significant. I include advice on how to do this testing below.

Response: As suggested by the reviewer, we have incorporated significance testing through a bootstrapping approach with 1000 iterations. Additionally, the related figures have been updated to display the significance of improvement.

Secondly, the authors have a tendency to be slightly sensationalist in their language. Some of the improvements are minor. One gets the impression that the case studies and examples are cherry-picked to some extent to show off FuXi to the best advantage. This language doesn't seem necessary to me – we all know this is a very challenging timescale, and the results are impressive as they are. I have suggested specific changes to the wording in places below.

Response: Thank you for the helpful comments. Accordingly, we have revised the language in our manuscript to address the issue you raised, particularly the specific comments provided below.

Major Comments

Significance testing.

The results are let down by a lack of significance testing throughout. Please rectify this. Each pair of bars in the bar charts (fig 1, SF1, 3, 4, 5, 8, 9, 11, 12) could be made bold for a pair where the difference is significant, and pale colouration if not significant.

Response:

As suggested by the reviewer, we conducted significance testing on all bar charts (Figures 1, SF 4, 3, 6, 7, 12, 13, 1, and 15) using the bootstrapping method with 1000 iterations. Results that were not statistically significant are now represented using a pale color scheme. Additionally, the original figures labeled 1, SF1, 3, 4, 5, 8, 9, 11, 12 have been renumbered as Figures 1, SF 4, 3, 6, 7, 12, 13, 1, and 15, respectively. In Figure 15, we use a cross-line on the bar plot to denote statistically significant improvements of the 51-member and 101-member FuXi-S2S forecasts over the ECMWF S2S reforecasts.

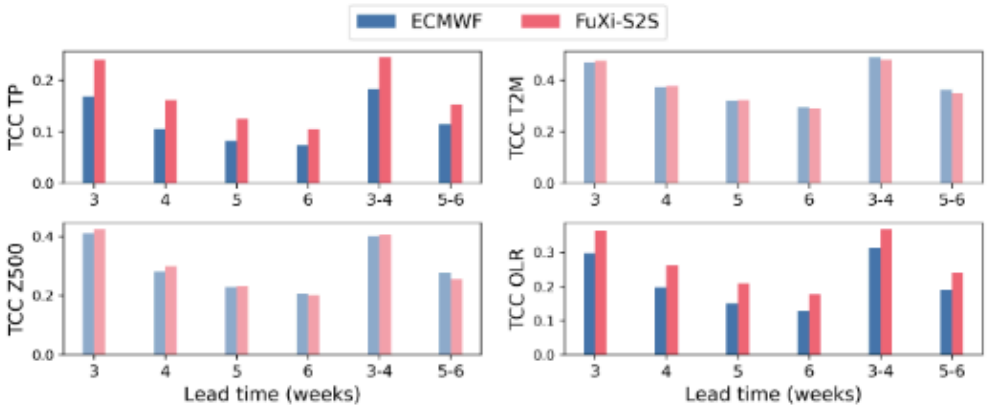
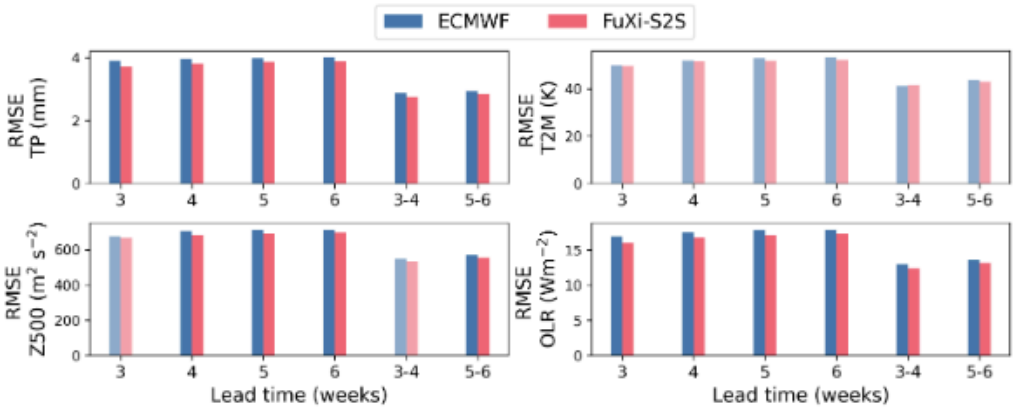
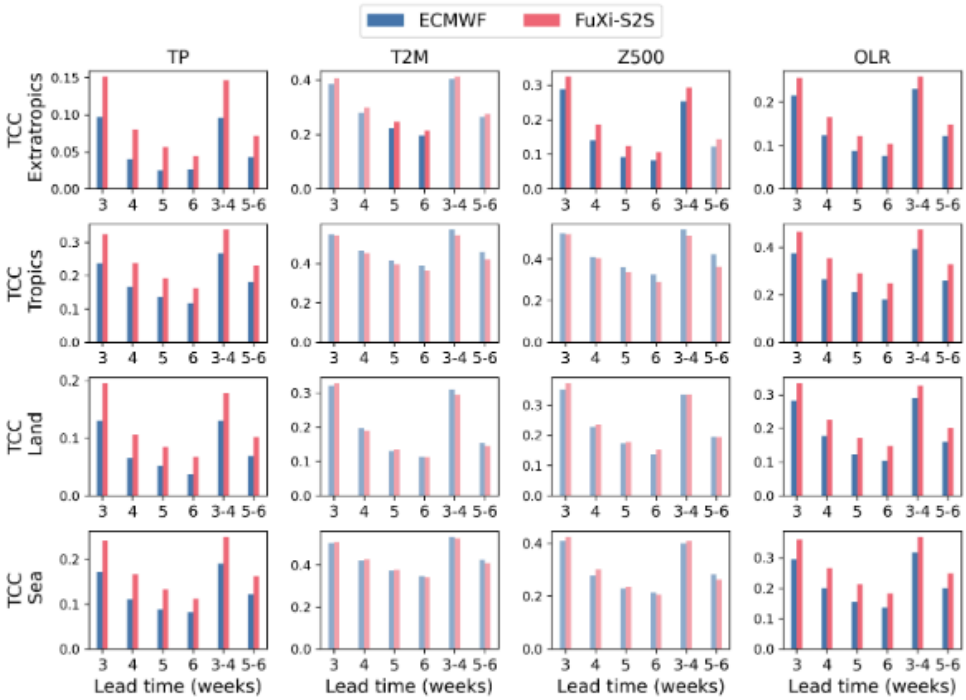


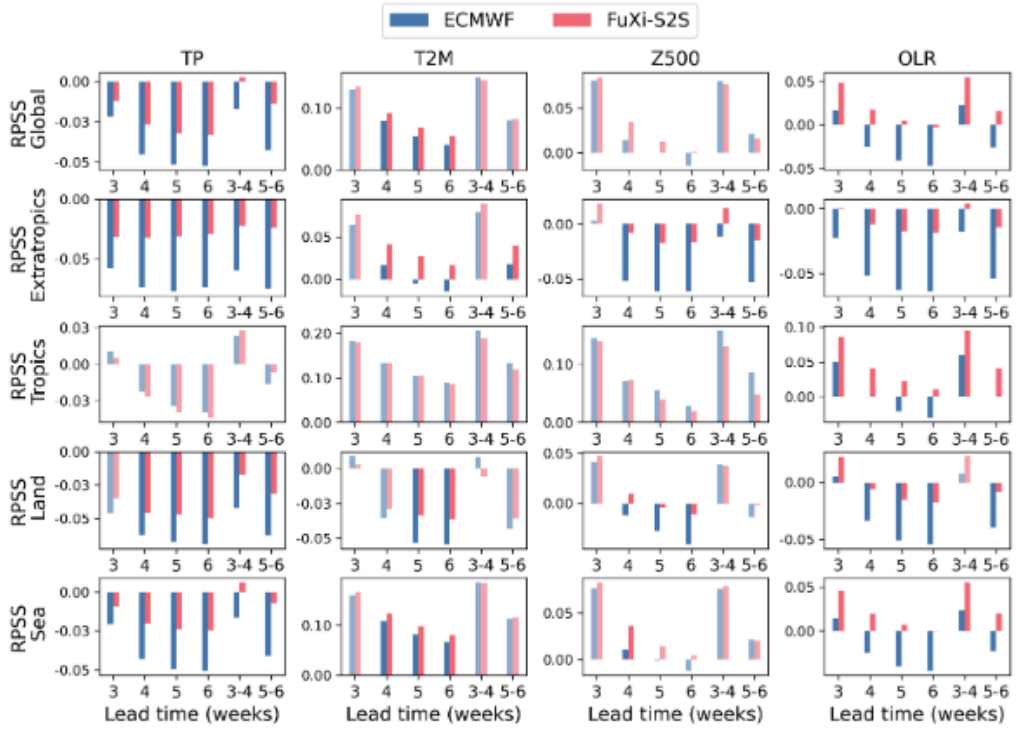
Fig. 1: Comparison of globally-averaged and latitude-weighted temporal anomaly correlation coefficient (TCC) of the ensemble mean between ECMWF S2S reforecasts (in blue) and FuXi-S2S forecasts (in red) for TP, T2M, Z500, and OLR. Rows 1 and 2 represent the performance across these variables, utilizing all testing data from the period spanning from 2017 to 2021. A bootstrapping approach, repeated 1000 times, is used for significant testing. When the FuXi-S2S forecasts fail to show a statistically significant improvement over the ECMWF S2S reforecasts at the 97.5% confident level, a pale color scheme is used to denote these results.



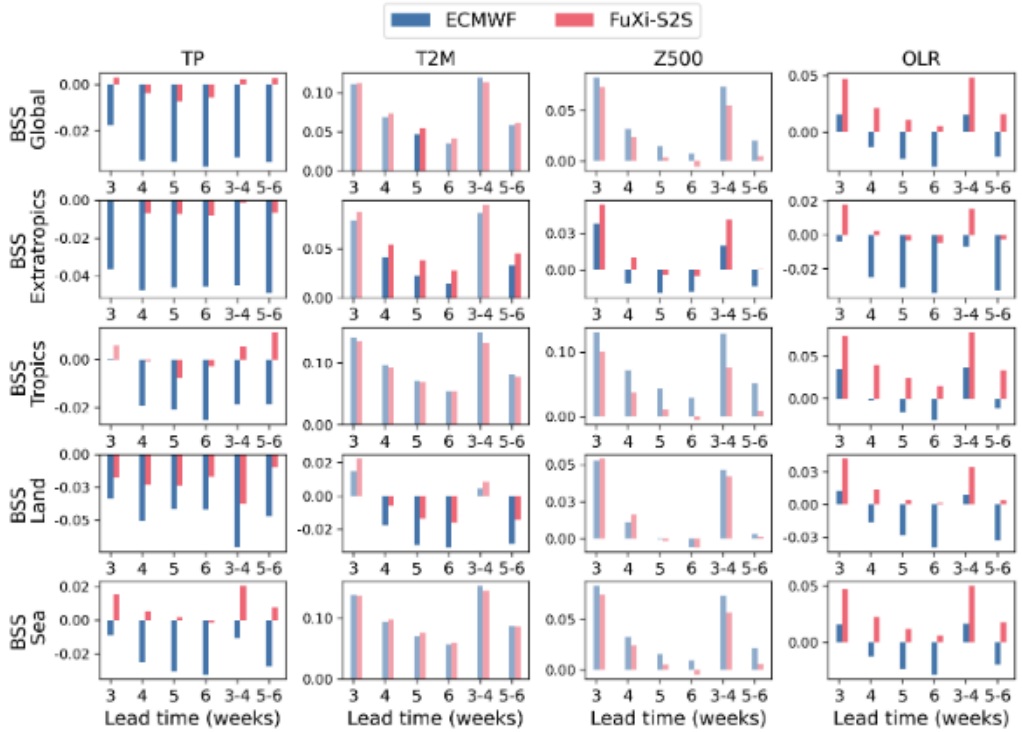
Supplementary Figure 4: Comparison of the globally-averaged and latitude-weighted RMSE of the ensemble mean between ECMWF S2S reforecasts (in blue) and FuXi-S2S forecasts (in red) for TP, T2M, Z500, and OLR, using all testing data between 2017 and 2021. A bootstrapping approach, repeated 1000 times, 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. It is important to note that TP here refers to 24-hour accumulated precipitation.



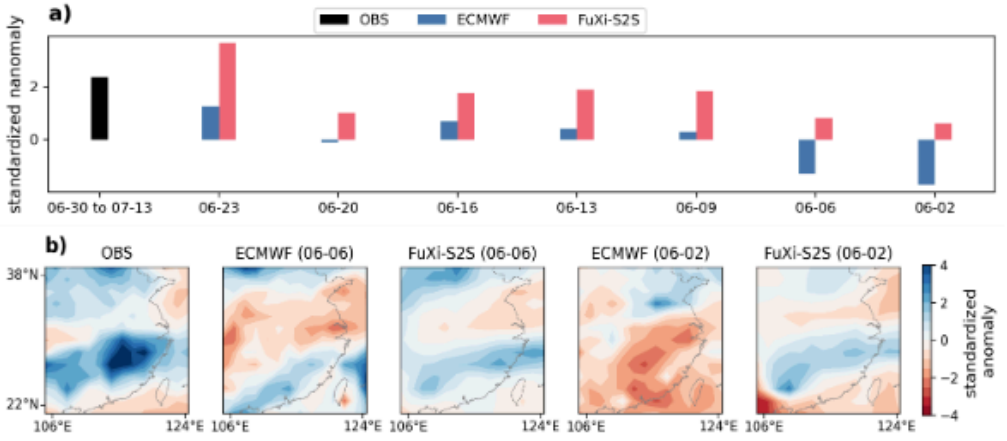
Supplementary Figure 3: Comparison of the latitude-weighted TCC of the ensemble mean of ECMWF S2S (in blue) forecasts and FuXi-S2S forecasts (in red) for TP (first column), T2M (second column), Z500 (third column), and OLR (fourth column) averaged over extra-tropics (90°S - 30°S and 30°N - 90°N, first row), tropics (30°S - 30°N, second row), land (third row), and sea (fourth row), using all testing data between 2017 and 2021. 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.



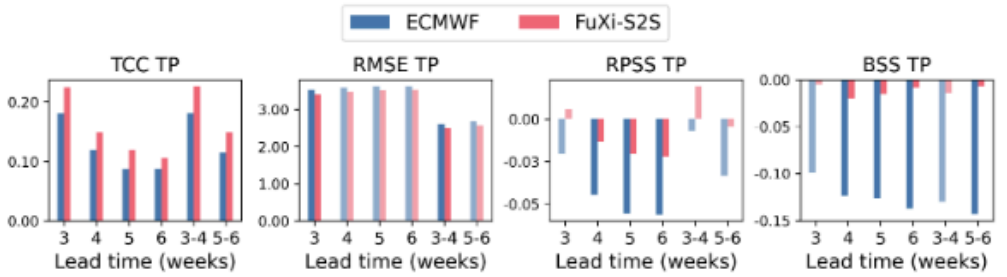
Supplementary Figure 6: Comparison of the latitude-weighted RPSS of ECMWF S2S (in blue) forecasts and FuXi-S2S forecasts (in red) for TP (first column), T2M (second column), Z500 (third column), and OLR (fourth column) averaged over extra-tropics (90°S - 30°S and 30°N - 90°N, first row), tropics (30°S - 30°N, second row), land (third row), and sea (fourth row), using all testing data between 2017 and 2021. 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.



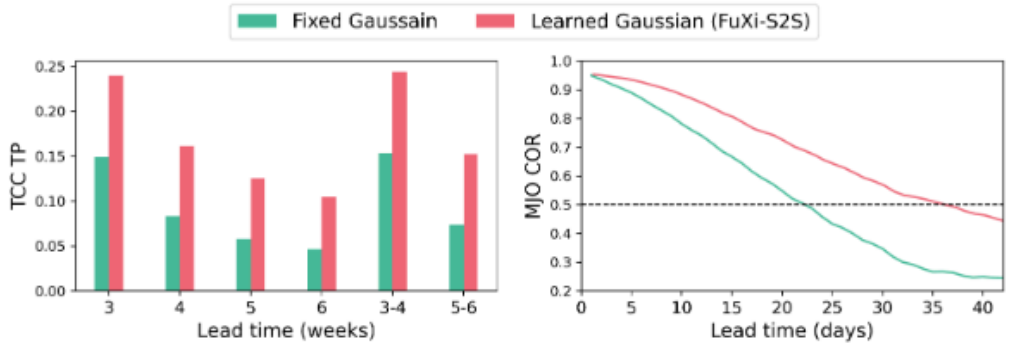
Supplementary Figure 7: Comparison of the latitude-weighted BSS of the ensemble mean of ECMWF S2S (in blue) forecasts and FuXi-S2S forecasts (in red) for TP (first column), T2M (second column), Z500 (third column), and OLR (fourth column) averaged over extra-tropics (90°S - 30°S and 30°N - 90°N, first row), tropics (30°S - 30°N, second row), land (third row), and sea (fourth row), using all testing data between 2017 and 2021. 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.



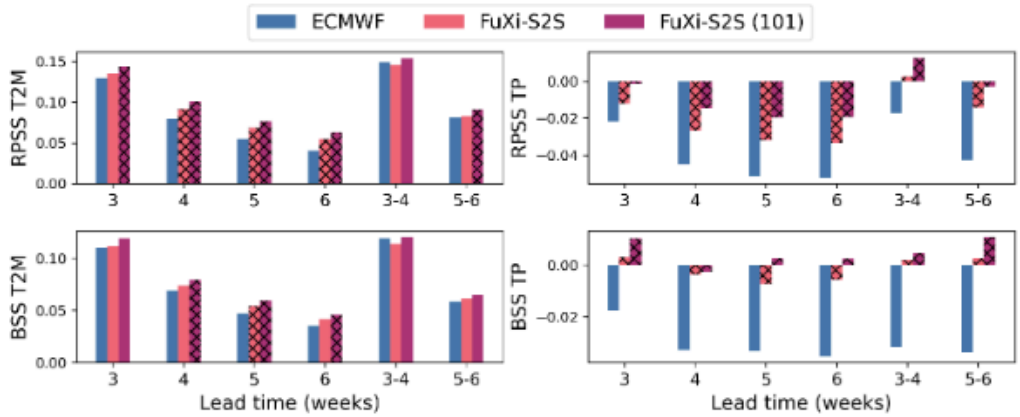
Supplementary Figure 12: Comparison of the spatially and temporally averaged standardised TP anomaly (a) for the 2 weeks from June 30th to July 13th, 2020 for GPCP observation (in black) and the predictions from ECMWF S2S reforecasts (in blue) and FuXi-S2S forecasts (in red), with initialization dates: June 23rd (06-23, MM-DD), June 20th (06-20), June 16th (06-16), June 13th (06-13), June 9th (06-09), June 6th (06-06), and June 2nd (06-02). Comparison of the temporally averaged standardised TP anomaly maps (b) for GPCP observation (first column) and predictions from ECMWF S2S (second column) and FuXi-S2S (third column), with initialization dates on June 6th (06-06, first row), and June 2nd (06-02, second row).



Supplementary Figure 13: Comparison of the globally-averaged latitude-weighted TCC (first column), RMSE (second column), RPSS (third column), and BSS (fourth column) between ECMWF S2S real-time forecasts (in blue) and FuXi-S2S forecasts (in red) for TP, using testing data from 2022. When the FuXi-S2S forecasts do not demonstrate a statistically significant improvement over the ECMWF S2S reforecasts, a pale color scheme is used to denote these results. It is important to note that TP here refers to 24-hour accumulated precipitation. 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.



Supplementary Figure 1: Comparison of the FuXi-S2S model (in red) and FuXi-S2S with fixed Gaussian perturbations (in green), utilizing all testing data from 2017 to 2021. The first column is the comparison of the globally-averaged latitude-weighted TCC. The second column is the comparison of the globally-averaged latitude-weighted RMM bivariate COR of the FuXi-S2S (in red) and FuXi-S2S with fixed Gaussian noise (in light red) using testing data from 2017 to 2021. 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.



Supplementary Figure 15: Comparison of the globally-averaged latitude-weighted RPSS (first row) and BSS (second row) between ECMWF S2S reforecasts (in blue), 51-member FuXi-S2S forecasts (in red), and 101-member FuXi-S2S forecasts (in purple) for T2M and TP, using testing data from 2017 to 2021. When the 51-member FuXi-S2S forecasts or 101-member FuXi-S2S forecasts demonstrate a statistically significant improvement over the ECMWF S2S reforecasts at the 97.5% confidence level, a cross-line on the bar plot is used to denote these results.

It would be usual to show stippling on a map (Fig 2, SF 2, 7, 8) if the skill score is significant, and leave without stippling if it is not.

Response:

Following the reviewer's suggestion, we have added stippling on the map (see Fig 2, SF 2) to indicate areas where the skill score is statistically significant. The original

Supplementary Figures 7 and 8, which are renumbered as Supplementary Figures 9 and 11, do not represent spatial maps of errors. Therefore, no singicance testing has been conducted for these two figures.

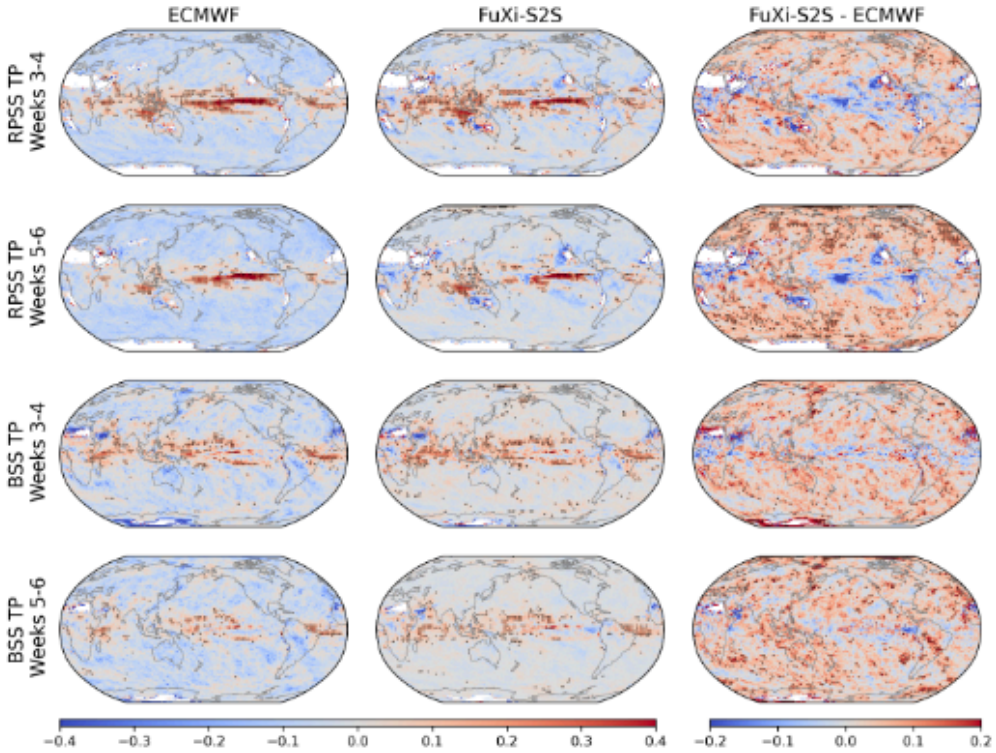
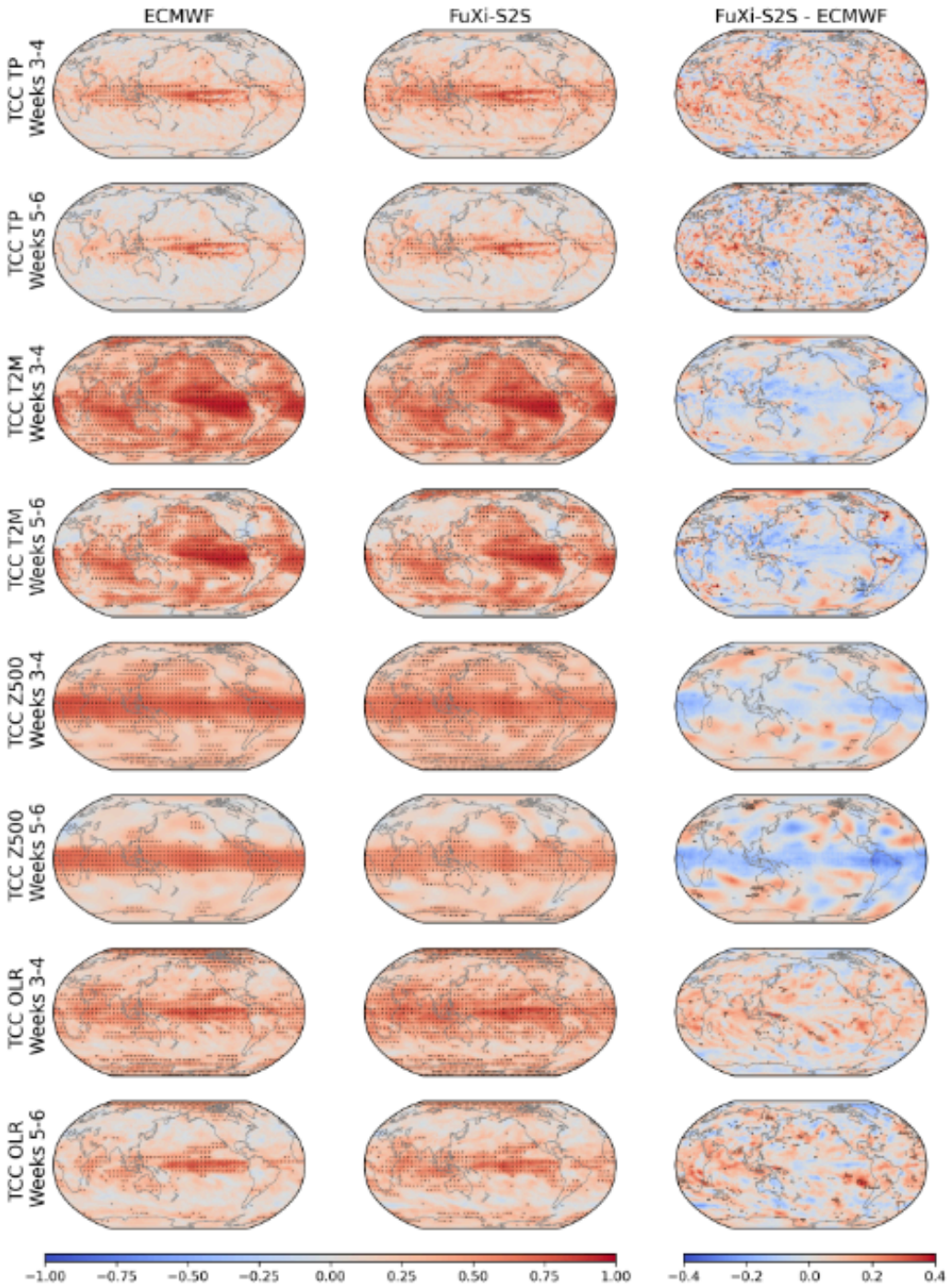


Fig. 2: Maps displaying the average Ranked Probability skill Score (RPSS) (first and second rows) and Brier Skill Score (BSS) (third and fourth rows) without latitude weighting, comparing ECMWF S2S (first column) and FuXi-S2S (second column) forecasts. Additionally, the third column depicts the difference in RPSS and BSS between FuXi-S2S and ECMWF S2S for 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 surpasses those of climatology. In column 3, stippling highlights areas where the FuXi-S2S model significantly outperforms the ECMWF S2S.



Supplementary Figure 2: Spatial map of average TCC without latitude weighting of ECMWF S2S (first column) and FuXi-S2S (second column), and the differences in TCC between FuXi-S2S and ECMWF S2S (third column) for TP (first and second rows), T2M (third and fourth rows), Z500 (fifth and sixth rows), and OLR (seventh and eighth rows) at forecast lead times of weeks 3-4 (first, third, fifth, and seventh rows), weeks 5-6 (second, fourth, sixth, and eighth rows), using all testing data between 2017 and 2021. 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 surpasses those of climatology. In column 3, stippling highlights areas where the FuXi-S2S model significantly outperforms the ECMWF S2S.

I include comments on how best to carry out this significance testing, in case they are of use.

A t-test is not appropriate here, and will likely indicate insignificant results. This is because you have paired data. The predictability of the atmosphere varies substantially from day such that the potential skill of a forecast can also vary substantially from day to day. The interesting thing is whether FuXi has slightly better skill than ECMWF for each day, even if the distributions of skill over all the days are very strongly overlapping because of the varying underlying predictability. A bootstrapping approach can account for this. To do this, create a large number (e.g. 1000) of synthetic datasets: to populate each day of the synthetic dataset, randomly choose between the forecast from model A and from model B for that day. Then assess the skill score for each of these 1000 synthetic datasets by comparing with observations. If the skill score for model A is greater than the 97.5% of the distribution of skill scores from the synthetic datasets then you could consider that model to be "significantly better" than model B, while less than 2.5% would be "significantly worse", for example.

Response:

Following the reviewer's suggestion, we have employed the bootstrapping method, performing 1000 iterations, to conduct significance testing on all the bar charts and maps presented in this paper. Additionally , we have added a description of significance testing in the Section 4.4:

"*Atmospheric predictability exhibits significant day-to-day variability, which in turn affects the potential accuracy of weather forecasts. To determine whether FuXi-S2S consistently outperform 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 map of statistical metrics. For all the bar plots in the paper, a pale color is used when the FuXi-S2S model do 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.*"

Sensationalist language.

I suggest making the following specific changes

P5 "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 marginally inferior performance at lead times of 6 and 5-6 weeks." Either include the qualifier "marginally" in front of both "superior" and "inferior" or in front of neither. The improvement and degradation are the same approximate size

Response: Following the reviewer's suggestion, we have excluded the word "marginally" in front of of both "superior" and "inferior".

P5 "Moreover, FuXi-S2S also outperforms ECMWF over a large part of extra-tropical regions for both T2M and Z500." Please add qualifier "though it is generally less skilful in the tropics".

Response: As suggested by the reviewer, we have changed the sentence to "Moreover, FuXi-S2S also outperforms ECMWF in a majority of extra-tropical regions for both T2M and Z500, although its performance is generally less skilful in the tropical areas."

*P6 Section 2.2. Ensemble forecast. It's worth mentioning (especially following the significance testing) that the skill is negative compared to climatology for both ECMWF and FuXi over large regions of the globe, and particularly in the extra tropics. This indicates just how hard the problem is – that climatology can do better than either a dynamical or ML forecast model. So it's not just a case of "considerably higher skills in tropical regions" – it's more that there *is* skill in the tropics whereas there is no skill in the extra tropics.*

Response: We agree with you and have modified the sentences by following the reviewer's suggestion, as follows: "*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. Moreover, the RPSS values are notably higher over oceans compared to land areas.*"

P7 "Notably, FuXi-S2S shows consistently higher RPSS values than ECMWF S2S across all variables, namely TP, T2M, Z500, and OLR, especially over extratropical averages." This is not true: TP in the tropics is more skilful for ECMWF than FuXi for lead times of 3, 4, 5 and 6 weeks, yet FuXi does better for the two-week averages. This is a notable and slightly confusing result. Why can FuXi do better on the 2-week averages than on the single weeks? The degradation seems to be dominated by the

central-eastern Pacific, masking good performance elsewhere (e.g.as demonstrated later by the Pakistan flood case)

Response:

We have revised the sentences as "*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 extratropical averages. However, in the tropics, ECMWF S2S outperforms FuXi-S2S at lead times of 3 to 6 weeks for one-week averages, whereas FuXi-S2S surpasses ECMWF S2S for two-week averages. This discrepancy in performance likely arises from the fact that one-week averages filter out variability with periods shorter than two weeks, while two-week averages attenuate variability with periods shorter than four weeks. Thus, the skill differences between the one-week and two-week averages may reflect FuXi-S2S's enhanced ability in capturing lower-frequency variability. Furthermore, a previous study [35] 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.*"

Reference:

[35] de Andrade, F., Coelho, C.A., Cavalcanti, I.F.: Global precipitation hindcast quality assessment of the subseasonal to seasonal (s2s) prediction project models. Climate Dynamics 52(9), 5451–5475 (2019)

P11 "The ECMWF S2S forecasts gradually converge toward observations as the initialization dates approach the actual event. In contrast, FuXi-S2S exhibits superior forecast performance" this is an interesting phrasing, as the smooth convergence of ECMWF is actually quite a desirable property. Why do you see such jumpiness in the FuXi-S2S forecasts? What is the ensemble doing here? It would be interesting to see 25 and 75th percentiles of the ensemble for each start date for ECMWF and FuXi-S2S

Response:

As suggested by the reviewer, we have added the 25th and 75th percentiles of the ensemble for each start date for ECMWF and FuXi-S2S in Figure 4. Regarding the jumpiness in the FuXi-S2S forecasts, we have included the following sentences to explain this phenomenon:

"*Forecast skill typically improves with descreasing 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, possible due to SST influence. Indeed, the saliency maps show that the FuXi-S2S forecasts initialized on July 28 and July 21 successfully captured predictabable signals from SST anomalies in the tropical central Pacific and western Indian Ocean (Figure 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."

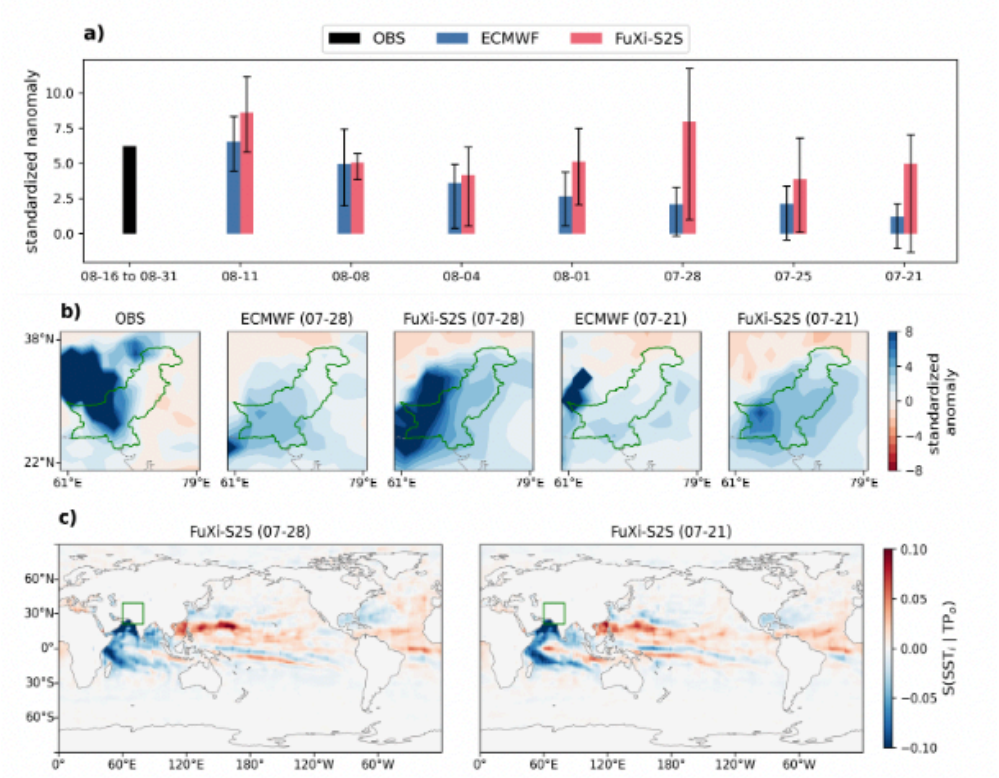


Fig. 4: Comparison of spatially and temporally averaged standardized TP anomaly (a) over the two weeks from August 16th to August 31st, 2022, showcasing GPCP observations (in black) alongside predictions from ECMWF S2S real-time forecasts (in blue) and FuXi-S2S forecasts (in red), with initialization 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 Figure 4 represent the 25th and 75th percentiles of the ensemble forecasts for each start date for both ECMWF and FuXi-S2S models.

Minor comments

P2 Introduction. This was comprehensive and easy to follow – very nice

Response: Thank you for your positive feedback on the Introduction section of our paper. We are pleased to hear that you found it comprehensive and easy to follow.

P4 "It compares the performance of FuXi-S2S with that of the 11-member ECMWF S2S reforecasts from the model cycle C47r3". Please add qualifier "over the same period"

Response: As suggested by the reviewer, we have appended the qualifier "*over the same period*" to the end of the sentence.

P8 L7 "generally predicts more regions" -> "generally exhibits more regions"

Response: As suggested by the reviewer, we have changed the sentence to "*generally exhibits more regions*".

P8 L8 "more areas with higher skills relative to climatological forecasts" -> "more areas with skill relative to climatological forecasts"

Response: As suggested by the reviewer, we have changed the sentence to "*more areas with skill relative to climatological forecasts*".

P9 paragraph 2. I found this description of the MJO evaluation confusing – more detail would be beneficial. In particular, clarify which multivariate EOFs you project onto (do all models use the observed EOFs as a basis? And clarify that you compute the ensemble mean anomalies first, and then compute the RMM of the ensemble mean (which I think is what you are doing, as opposed to computing the RMM of each ensemble member and then averaging).

Response: Point taken. To clarify this, we have revised the sentences as "*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°to a 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.*"

P10 Fig 3 caption. Why is the RMM COR globally averaged and latitude weighted? It's just an index.

Response: Thank you for highlighting this issue. We acknowledge this was a typo and have now removed the phrase "*globally-averaged and latitude-weighted*" from the caption of Figure 3.

P10 Para 1. I'd like to see the skill of FuXi-S2S compared to ECMWF at predicting the phase and skill at predicting RMM amplitude – perhaps as an extra supp mat figure. What is giving this improvement in skill? Sounds from this discussion like the amplitude is the main source of improvement.

Response:

As suggested by the reviewer, we have calculated and analyzed the bivariate correlation coefficient and errors associated with the MJO phase and amplitude. Additinoally, we have included a comparison of the FuXi-S2S and ECMWF S2S models for these metrics in Supplementary Figure 9. Corresponding descriptions have been incorporated into the main text as follows:

"*Moreover, Supplementary Figure 9 presents the bivariate correlation coefficient (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.*"

Also, we have added the following sentences and a Figure in the Supplementary material:

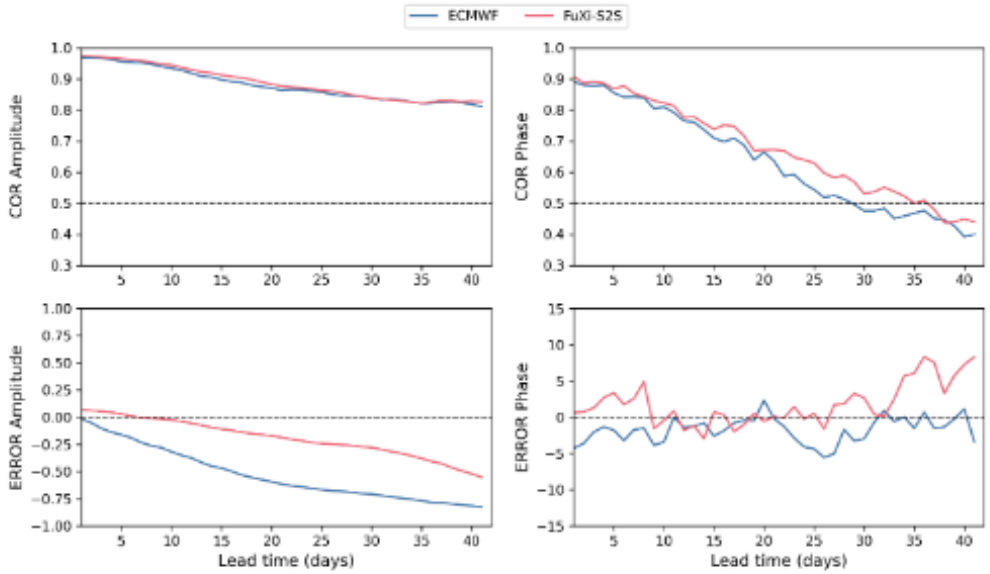
"*Supplementary Figure 9 presents a comparative analysis of the correlation (COR) and error (ERROR) metrics for the amplitude and phase of the MJO. These metrics are derived from the ensemble mean of ECMWF S2S reforecasts and FuXi-S2S forecasts, averaged over all the testing data from 2017 to 2021. Among them, the COR reflects the accuracy of evolution, and ERROR indicates the systematic bias. The analysis reveals that COR values decline with increasing forecast lead times, with a more pronounced decrease observed for the MJO phase compared to the amplitude. Throughout the 42-day forecast period, the COR for MJO amplitude remains consistently above 0.8. The differences in amplitude COR between the ECMWF S2S and FuXi-S2S models are negligible. In contrast, FuXi-S2S consistently outperforms ECMWF S2S in phase COR, maintaining higher values over the entire forecast duration. Specifically, the COR for the MJO phase drops below 0.5 at 28 days for ECMWF S2S, whereas for FuXi-S2S, it remains above this threshold until 34 days. Additionally, negative error values for the MJO amplitude indicate that the amplitude is on average smaller in both the ECMWF S2S and FuXi-S2S simulations compared to ERA5 data, aligning with findings from previous studies [2, 3]. FuXi-S2S exhibits smaller errors than ECMWF S2S, suggesting it better maintains the amplitude of MJO events. Regarding the error in the MJO phase, both models show comparable values up to 32 days, indicating the small systematic phase speed error in both models. However, after 32 days, FuXi-S2S shows larger phase errors than ECMWF S2S.*

Overall, the superior performance of FuXi-S2S in predicting MJO compared to that of ECMWF S2S is primarily due to its enhanced ability to predict the MJO phase."

References:

[2] Vitart, F., Woolnough, S., Balmaseda, M.A., Tompkins, A.M.: Monthly forecast of the madden–julian oscillation using a coupled gcm. Monthly Weather Review 135(7), 2700–2715 (2007). https://doi.org/10.1175/MWR3415.1

[3] Vitart, F., Molteni, F.: Simulation of the madden–julian oscillation and its teleconnections in the ecmwf forecast system. Quarterly Journal of the Royal Meteorological Society 136(649), 842–855 (2010). https://doi.org/10.1002/qj.623



Supplementary Figure 9: Comparison of correlation (COR) (first row) and error (ERROR) (second row) in the amplitude (first column) and phase (second column) of the MJO of the ensemble mean between ECMWF S2S reforecasts (in blue) and FuXi-S2S forecasts (in red) using all testing data from 2017 to 2021. Dashed black line signifies the prediction skill threshold of COR=0.5 and ERROR=0.

We have also added a description of how the COR and errors are calculated for the amplitude and phase of the MJO in Subsection 4.4:

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 ($\mathrm{COR}_{amplitude}$) and phase (COR_{phase}) were calculated using the methods outlined by Wang et al. [95] as follows:

$$\mathrm{COR}_{amplitude}(\tau) = \frac{\sum_{t=1}^{N} \mathrm{RMMA}_{obs}(t) \times \mathrm{RMMA}_{forecast}(t,\tau)}{\sqrt{\sum_{t=1}^{N} \mathrm{RMMA}_{obs}^{2}(t)} \sqrt{\sum_{t=1}^{N} \mathrm{RMMA}_{forecast}^{2}(t,\tau)}} \quad (10)$$

$$\mathrm{COR}_{phase}(\tau) = \frac{\sum_{t=1}^{N} \mathrm{RMMA}_{obs}(t) \times cos(\theta_{forecast}(t,\tau) - \theta_{obs}(t))}{\sum_{t=1}^{N} \mathrm{RMMA}_{obs}^{2}(t)} \quad (11)$$

where RMMA_{obs} and $\mathrm{RMMA}_{forecast}$ 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 ($\mathrm{ERROR}_{amplitude}$ and ERROR_{phase}) as follows, based on the method described by Rashid et al. [94]:

$$\mathrm{ERROR}_{amplitude}(\tau) = \frac{1}{N} \sum_{t=1}^{N} (\mathrm{RMMA}_{forecast}(t,\tau) - \mathrm{RMMA}_{obs}(t)) \quad (12)$$

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

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

[94] Rashid, H.A., Hendon, H.H., Wheeler, M.C., Alves, O.: Prediction of the madden–julian oscillation with the poama dynamical prediction system. Climate Dynamics 36, 649–661 (2011)

[95] 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. Climate Dynamics 52, 6015–6031 (2019)

P11 "Supplementary Figure 7" I found it hard to understand exactly what this figure showed. Are these composites given that the MJO starts in phase 4? Or is it composites given that it is in phase four in each of week 3, or week 4, etc.

Response: Yes, those composites in the original "Supplementary Figure 7" correspond to the forecasts when MJO starts in phase 4 at forecast initialization times. To clarify this, we have changed the cpation of the "Supplementary Figure 7" as *"Composite map of Z500 anomalies derived from ERA5 reanalysis data (first column), ECMWF S2S reforecasts (second column) and FuXi-S2S forecasts (third column). These maps cover forecast lead times of weeks 3, 4, 5, and 6, represented in the first, second, third, and*

fourth rows, respectively. All maps use testing data between 2017 and 2021, corresponding to initial forecast periods when the MJO is in phase 4 of its lifecycle."

P13 the interpretation of precursors was very nice to see.

Response: Thank you for your positive feedback regarding our interpretation of the precusors.

P16 On the creation of S2S anomalies. You describe the approach taken for ECMWF – what do you do for FuXi-S2S – do you also produce a set of hindcasts for FuXi and use them to construct a climatology to be subtracted off? This would be the fairest approach, allowing lead-time (and model) dependent biases to be removed from FuXi-S2S as they are for the ECMWF forecasts

Response:

Indeed, we employed the fairest approach: the anomalies for each forecast were calculated by subtracting forecasts from the climatology of the respective forecast. And we have made it clear by adding the following sentence:

"*Furthermore, a set of hindcasts from 2002 to 2016 is generated for FuXi-S2S, which are used to to establish a 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.*"

P17 Section 4.2 What is the timestep used by FuXi-S2S – have I just missed this detail?

Response: The timestep used by FuXi-S2S is one day. To clarify, we have revised the sentence as "*The encoder, processing predicted weather parameters from two preceding time steps, with each time step representing one day, as FuXi-S2S is designed to forecast daily mean values.*".

P19 L4 "Similar to FuXi, an autoregressive, we utilize an autoregressive …" – typo? Remove the first "an autoregressive"

Response: Thanks for highlighting this typo, we have removed the first "an autoregressive" in the sentence.

P20 Section 4.4 On the detrending – what exactly are you detrending? Is this across the 42-day forecasts? Or somehow over the hindcast period?

Response: To clarify the detrending process we did, we have revised the sentences as

"*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 [90]. 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.*"

P20, 7 lines from the bottom "The tercile bounds are determined …" – are these determined for each model separately? Please clarify

Response: Yes, we calculated the tercile for each model separately. As suggested by the reviewer, we have made it clear by adding the following sentence "*These calculations of tercile bounds are performed separately for each forecast model and observation (ERA5 data).*" following "*The tercile bounds are determined …*".

Reviewer #2 (Remarks on code availability):

I have checked that the record exists, but I have not attempted to download the code and run it, etc.

If the paper passes onto another round of reviews, I'm happy to ask someone in my group to check the code.

Response:

We appreciate the efforts and wiliness of the reviewer to download and evaluate our model, which is accessible via the following link:

https://drive.google.com/drive/folders/1xeZqomNfwOVZDavWTXj8oRxidpemuwk8?usp=sharing.

Everybody is welcome to utilize the model freely, including conducting research using the FuXi-S2S dataset.

Reply to Reviewer #3

We would like to thank the reviewer for careful review and constructive comments. We have revised the paper and implemented all the suggestions. Our reply follows with original comments quoted in italic font.

1 Summary of the Content

In this manuscript, the authors explore the use a vision transformer-based model (FuXi-S2S) to predict the 24-hour averaged 3-dimensional atmosphere with emphasis on the 2-6 week forecast range, commonly known as subseasonal-to-seasonal (S2S) time frame. Presently, the use of data-driven models for the S2S time frame have not been well explored compared to the medium-range (1-14 day) time frame. Using a novel probabilistic encoder-processor-decoder architecture (note this is different than the original FuXi model), FuXi-S2S is able to achieve skillful S2S forecasts. FuXi-S2S uses its inherit probabilistic nature to generate ensemble members, of which are used to quantify forecast uncertain and provide an ensemble mean forecast. When compared to a state-of-the-art numerical-based S2S system (ECMWF S2S), FuXi-S2S outperforms it for several deterministic (temporal anomaly correlation coefficient) and probabilistic metrics (Ranked Probability skill Score and Brier Skill Score) for a select number of variables, mainly total precipitation and outgoing long-wave radiation (OLR).

Another strength of FuXi-S2S is the ability to predict the Madden–Julian Oscillation (MJO) with skill out to 36 days, an improvement of 6 days over the ECMWF S2S system. FuXi-S2S includes several new variables compared to the original FuXi model including 100-meter winds and OLR. Additionally, to the best of my knowledge, this is the first data-driven weather model that explicitly aims to machine learn the coupling between the sea surface temperature field and atmosphere. Overall, the manuscript is robust and represents an exciting step for data-driven S2S.

2 Overall Feedback

This paper demonstrates an important step for date-driven weather forecasting as the S2S time frame has long been difficult for NWP-based models. My only major concern is the assessment of the model quality is somewhat overstated. There are a few exaggerations which if toned down will improve the paper.

Response:

We appreciate the reviewer's feedback and have toned down the sections in question, as detailed in the response to the specific comments below.

1. As mentioned above, the FuXi-S2S improvement over NWP is overstated in some sections. First the horizontal resolution is 1.5 degrees, which is several time more coarse than other state-of-the-art data driven models (Graphcast, Pangu, original FuXi) and the ECWMF S2S system. While the results are promising and for at least the show variables and metrics in this paper outperforming ECWMF's S2S system, it is worth noting ECWMF's S2S is run at significantly high resolution. It is also worth noting for the total precipitation (TP) metrics, ERA5 is used as verification for the results shown in Figure 1 and 2. There are known disagreements between ERA5 TP and observations (Lavers et al., 2022). If observations or GCPC products were used at the verification the results in the paper may change.

Response:

Following the suggestion, we have revised as follows:

We have reworded the sentence in Abstract from "*When compared to the state-of-the-art ECMWF Subseasonal-to-Seasonal (S2S) reforecasts using a conventional model, the FuXi-S2S forecasts demonstrate superior deterministic and ensemble forecasts.*" to "*When compared to the state-of-the-art ECMWF Subseasonal-to-Seasonal (S2S) reforecasts using a conventional model, the FuXi-S2S forecasts demonstrate superior performance in both the ensemble mean and ensemble forecasts for total precipitation (TP) and outgoing longwave radiation (OLR)*".

We reworded the sentence in Discussion from "*Our results demonstrated that FuXi-S2S comprehensively surpasses ECMWF S2S in forecast accuracy*." to "*Our results demonstrated that FuXi-S2S surpasses ECMWF S2S in forecast accuracy for the evaluated variables*.".

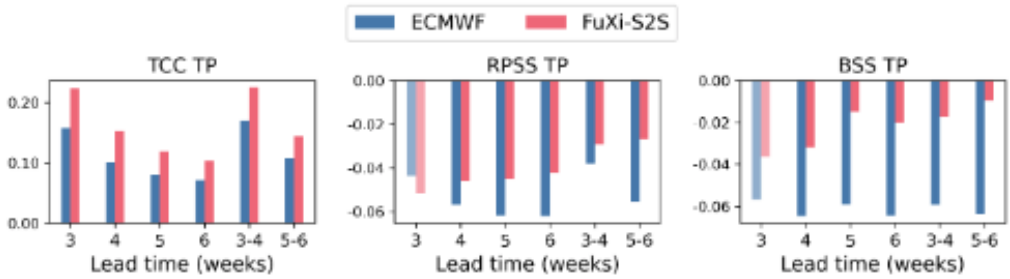
Further, we have added the sentence in Discussion:
"*While FuXi-S2S offers a computationally efficient and accurate alternative to conventional NWP models for subseasonal forecasting, it also presents significant opportunities for improvement. For instance, the ECWMF S2S model runs at a spatial resolution of 36 km [74], 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 [75, 76], GPCP observations have been utilized to evaluate the TP forecast performance for both ECMWF S2S and FuXi-S2S (refer to Supplementary Figures 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.".

References:
[74] Haiden, T., et al.: Evaluation of ECMWF forecasts, including the 2018 upgrade (2018)
[75] 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. Journal of Hydrology 583, 1–17 (2020)
[76] Lavers, D.A., Simmons, A., Vamborg, F., Rodwell, M.J.: An evaluation of era5 precipitation for climate monitoring. Q. J. R. Meteorol. Soc. 148(748), 3152–3165 (2022). https://doi.org/10.1002/qj.4351

Latsly, we have repeated total precipitation verification using GPCP data as ground truth. The results, as shown in Supplementary Figure 16, are generally consistent with those obtained using the ERA5 dataset. Therefore, we have kept the results verified against the ERA5 dataset in the main manuscript and added an additional section in the supplementary material that describes the comparative analysis of ECMWF S2S reforecasts and FuXi-S2S forecasts against the GPCP TP data.

"*Supplementary Figure 16 presents a comparative analysis of the globally-averaged and latitude-weighted TCC, RPSS, and BSS of the ensemble mean between ECMWF S2S reforecasts and FuXi-S2S forecasts for TP, based on testing data between 2017 and 2021. Unlike prior analyses, this evaluation employs the GPCP dataset as the reference, rather than the ERA5 dataset. Consistent with the results shown in Figure 1 of the main text and Supplementary Figure 4, where ERA5 serves as the verification target, the FuXi-S2S model generally outperforms ECMWF S2S at most forecast lead times, achieving higher TCC, RPSS, and BSS values than ECMWF S2S. However, an exception is noted in week 3, where ECMWF S2S exhibits superior RPSS values. Notably, since the FuXi-S2S is trained on TP data from the ERA5 dataset, its performance slightly diminishes when evaluated against the GPCP dataset. This reduction in performance is likely due to the discrepancies between the GPCP and ERA5 datasets. Considering the known differences between the ERA5 TP data and actual observations, as highlighted in [9,10], exploration of more accurate TP data sources is planned to enhance the forecast accuracy of the FuXi-S2S model.*"



Supplementary Figure 16: Comparison of the globally-averaged and latitude-weighted TCC, RPSS, and BSS between ECMWF S2S reforecasts (in blue) and FuXi-S2S forecasts (in red) for TP, using all testing data between 2017 and 2021. Notably, verification is conducted with the GPCP dataset, rather than ERA5 dataset. 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.

a) "FuXi-S2S forecasts demonstrate superior deterministic and ensemble forecasts." The only comparisons in the paper are ensemble means using deterministic and probabilistic metrics using the ensemble forecasts. Neither of these are deterministic forecasts (e.g., the ensemble mean is not a deterministic forecast). To show that has better deterministic forecasts, a single forecast for a date using FuXi-S2S would need to be compared to other models like IFS, GraphCast, Pangu, etc. There are a number of references to the ensemble mean as a deterministic forecast I suggest rewording them.

Response:

As suggested by the reviewer, we have replaced "deterministic forecast" with "ensemble mean forecast" and revised the following sentences accordingly.

We have reworded the sentence in Abstract from "When compared to the state-of-the-art ECMWF Subseasonal-to-Seasonal (S2S) reforecasts using a conventional model, the FuXi-S2S forecasts demonstrate superior deterministic and ensemble forecasts." to "When compared to the state-of-the-art ECMWF Subseasonal-to-Seasonal (S2S) reforecasts using a conventional model, the FuXi-S2S forecasts demonstrate superior performance in both the ensemble mean and ensemble forecasts for total precipitation (TP) and outgoing longwave radiation (OLR)."

We have reworded the sentence in Introduction from "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 [5,34]." to "*Remarkably, FuXi-S2S outperforms the ECMWF Subseasonal to Seasonal (S2S) ensemble, which is recognized as the most skillful S2S modeling system, in producing both deterministic forecasts of the ensemble mean and probabilistic forecasts [5,36]*".

We have reworded the sentence in Section 2.2 from "*Deterministic forecasts based on the ensemble mean exhibit limited skill, with TCC values below 0.5 for all subseasonal forecast lead times.*" to "*Deterministic metrics, evaluated using the ensemble mean, exhibit limited predictive skill, with TCC values below 0.5 for all subseasonal forecast lead times.*"

Wehave reworded the sentence from "*A comprehensive suite of metrics was employed for this evaluation, including the deterministic forecast skill of the ensemble mean, the probabilistic skill of the ensemble forecast, and the capability to predict extreme events.*" to "*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*."

We reworded the sentence from "*Subsequently, the deterministic forecast performance of the ensemble mean is evaluated using the latitude-weighted TCC,*" to "*Subsequently, the deterministic metrics of the ensemble mean is evaluated using the latitude-weighted TCC,*"

b) "Our results demonstrated that FuXi-S2S comprehensively surpasses ECMWF S2S in forecast accuracy". This paper does not show that. Further diagnostics like the score-card used in GraphCast's paper would need to be shown. There only 4 variables are shown I would tone down this sentence.

Response: As suggested by the reviewer, we have toned down this sentence: "*Our results demonstrated that FuXi-S2S surpasses ECMWF S2S in forecast accuracy for the evaluated variables.*"

We agree with the reviewer on the importance of a comprehensive analysis involving a broader range of variables, similar to the score-card used in GraphCast's paper. However, subseasonal forecasts generally exhibit significantly lower predictability compared to medium-range forecasts, with most related studies evaluating only total precipitation and 2-meter temperature. Therefore, our analysis primarily focused on these two commonly assessed variables. Evaluating additional variables in subseasonal forecasts poses significant challenges, as it necessitates running hindcasts to establish the climatology needed for calculating anomalies. Moving forward, we aim to enhance our subseasonal forecast model and expand the list of variables assessed.

c) "FuXi-S2S has shown remarkable 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 is just one case study. I suggest tuning down this sentence (e.g. remove remarkable).

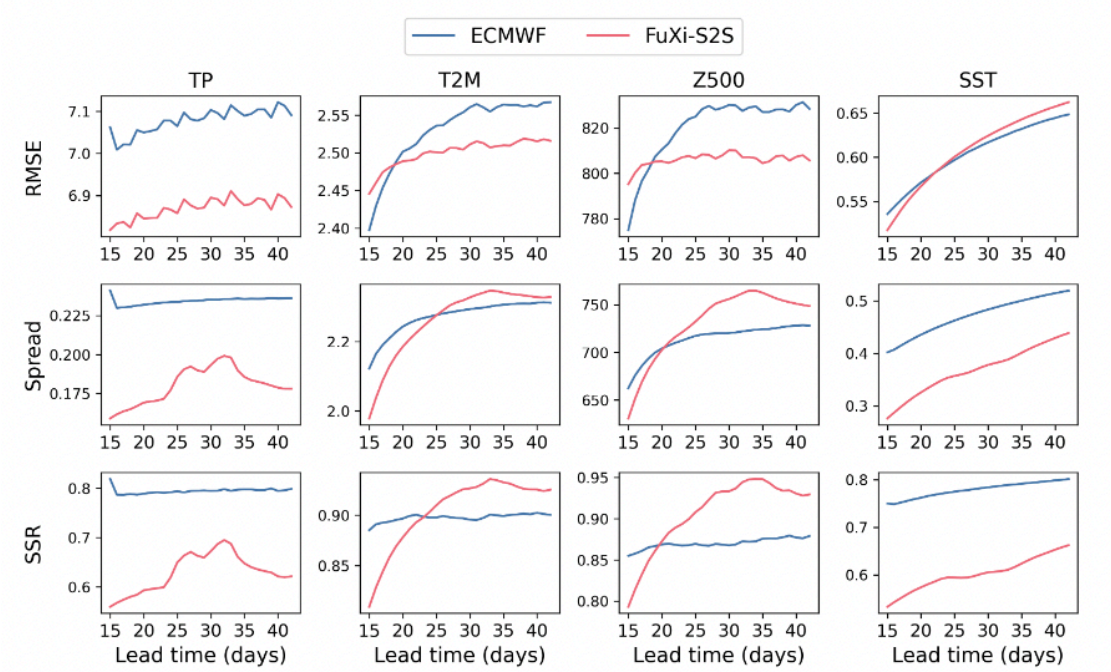
Response: As suggested by the reviewer, we have revised the sentence by removing the word "remarkable".

d) "Compared to conventional NWP ensemble forecasting methods, which often struggle with constructing initial condition perturbations due to the complexities of multi-variate interactions and the need to maintain dynamic balance and ensemble spread in simulations [37], our approach of introducing perturbations directly into the model's latent space offers a novel and effective alternative.". This raises a good question, how is the ensemble spread in FuXi-S2S? A skill to spread ratio or something similar would answer that question. If ensemble system is very under or over disruptive I suggest toning down the language about how well this machine learned method of perturbations is compared to traditional methods used for NWP. I will also note, NWP ensemble generation has come a long way since the 1996 (the year the paper referenced in this sentence was published).

Response: As suggested by the reviewer, we have calculated the ensemble spread and spread skill ratio (SSR) for the TP, T2M, Z500, and SST. As shown in Supplementary Figure 7, both the ECMWF S2S and FuXi-S2S models are underdispersive, as indicated by SSRs below 1 for all variables. However, the FuXi-S2S model exhibits larger spreads for T2M and Z500 compared to the ECMWF S2S model, with SSRs are closer to 1. This suggests that FuXi-S2S ensemble methods are more effective. Additionally, we have added a section that presents a detailed comparison of the ensemble spread and SSR between ECMWF S2S reforecasts and FuXi-S2S forecasts. We have also moderated the language regarding comparison of FuXi-S2S to traditional methods in subseasonal forecasts, acknowledging that there is still room for improvement in the FuXi-S2S model.

"*Supplementary Figure 7 compares the globally-averaged and latitude-weighted RMSE, ensemble spread, and spread skill ratio (SSR) between ECMWF S2S reforecasts and FuXi-S2S forecasts for TP, T2M, Z500, and sea surface temperature (SST). These metrics are derived from daily mean forecasts, calculated using all available testing data from 2017 to 2021 as a function of forecast lead times. For TP, FuXi-S2S consistently outperforms ECMWF S2S in terms of RMSE. For SST, FuXi-S2S initially shows slightly superior performance compared to ECMWF S2S for the forecast lead times of 15 to 20 days, but its performance declines relative to ECMWF S2S thereafter. In terms of ensemble spread, FuXi-S2S generally shows smaller spread than ECMWF S2S for both TP and SST. However, their SSR values are consistently lower than those of ECMWF S2S across all forecast lead times, suggesting that the ensemble spread of ECMWF S2S more accurately predicts forecast skill for these variables. For T2M, FuXi-S2S demonstrates SSR values closer to 1 compared to ECMWF S2S during the forecast lead times from 20 to 42 days, indicating a higher reliability of ensemble spread. Overall, both ECMWF S2S and FuXi-S2S have SSR values below 1 for all 4 evaluated variables across all forecast lead times, suggesting underdispersion. This*

result suggests that there is still room for improvement in the FuXi-S2S to achieve SSR closer to 1."



Supplementary Figure 7: Comparison of the globally-averaged, latitude-weighted RMSE, ensemble spread, and SSR of ECMWF S2S reforecasts (in blue), and FuXi-S2S forecasts (in red) for TP, T2M, and SST as a function of forecast lead times. This analysis includes all testing data between 2017 and 2021, using the daily mean forecasts to calculate these metrics.

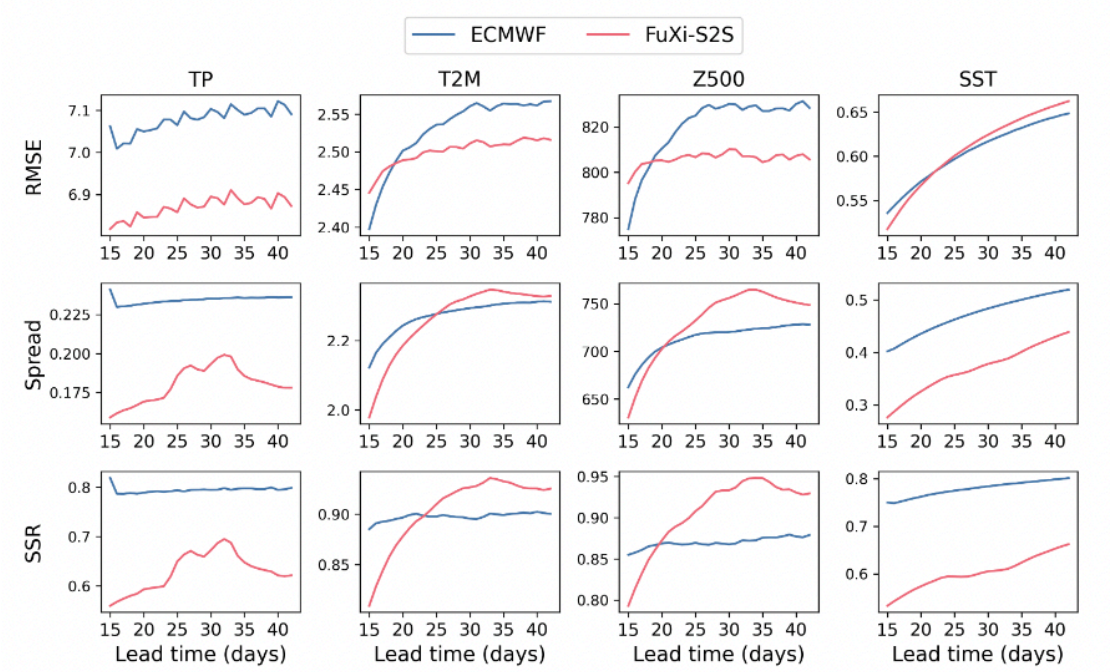
2. A brief discussion of the trades offs between the training of the original FuXi architecture (e.g., U-Transformer) vs the model in the paper would be beneficial to the reader. I assume there is some trade off training this architecture compared to standard U-transformer. For example does this architecture require longer training time, consumer more GPU memory, require any model-parallelisms, etc?

Response: As suggested by the reviewer, we have added a brief discussion of the difference between the original FuXi model architecture and the model used in this paper:

3. In my opinion, one of the most novel portions of this research in the prediction of sea surface temperatures (SST). I think including an assessment of the skill of SST forecasts (e.g. ensemble mean RMSE as a function of lead time) would be an interesting addition to the manuscript, either in the main paper or in the supplementary materials.

Response: Following the suggestion, we have included the evaluations of SST in the Supplementary Figure 7. FuXi-S2S initially exhibits marginally superior performance compared to ECMWF S2S for forecast lead times of 15 to 20 days; however, its performance diminishes relative to ECMWF S2S thereafter. We have also added a detailed description of these findings in the supplementary material:

"*Supplementary Figure 7 compares the globally-averaged and latitude-weighted RMSE, ensemble spread, and spread skill ratio (SSR) between ECMWF S2S reforecasts and FuXi-S2S forecasts for TP, T2M, Z500, and sea surface temperature (SST). These metrics are derived from daily mean forecasts, calculated using all available testing data from 2017 to 2021 as a function of forecast lead times. For TP, FuXi-S2S consistently outperforms ECMWF S2S in terms of RMSE. For SST, FuXi-S2S initially shows slightly superior performance compared to ECMWF S2S for the forecast lead times of 15 to 20 days, but its performance declines relative to ECMWF S2S thereafter. In terms of ensemble spread, FuXi-S2S generally shows smaller spread than ECMWF S2S for both TP and SST. However, their SSR values are consistently lower than those of ECMWF S2S across all forecast lead times, suggesting that the ensemble spread of ECMWF S2S more accurately predicts forecast skill for these variables. For T2M, FuXi-S2S demonstrates SSR values closer to 1 compared to ECMWF S2S during the forecast lead times from 20 to 42 days, indicating a higher reliability of ensemble spread. Overall, both ECMWF S2S and FuXi-S2S have SSR values below 1 for all 4 evaluated variables across all forecast lead times, suggesting underdispersion. This result suggests that there is still room for improvement in the FuXi-S2S to achieve SSR closer to 1.*"



Supplementary Figure 7: Comparison of the globally-averaged, latitude-weighted RMSE, ensemble spread, and SSR of ECMWF S2S reforecasts (in blue), and FuXi-S2S forecasts (in red) for TP, T2M, and SST as a function of forecast lead times. This analysis includes all testing data between 2017 and 2021, using the daily mean forecasts to calculate these metrics.

4. Related to my previous comment, the sea surface temperature field plays a large role for atmospheric modeling in the S2S time frame. There is also emerging research suggestion a relationship between the MJO and SST (Savarin and Chen, 2022). In future work wok it would be interesting to see how the MJO forecast performance changes if the SST field was persisted or not included at all.

Response:

We agree with the reviewer that SST is important for forecasts at subseasonal timescales. Investigating the impact of using prescribed fixed SST or excluding SST on the forecast performance of the MJO would be both interesting and valuable. However, this topic is beyond the scope of the current study. To clarify this point, we have added the following sentences to Section 2.4.

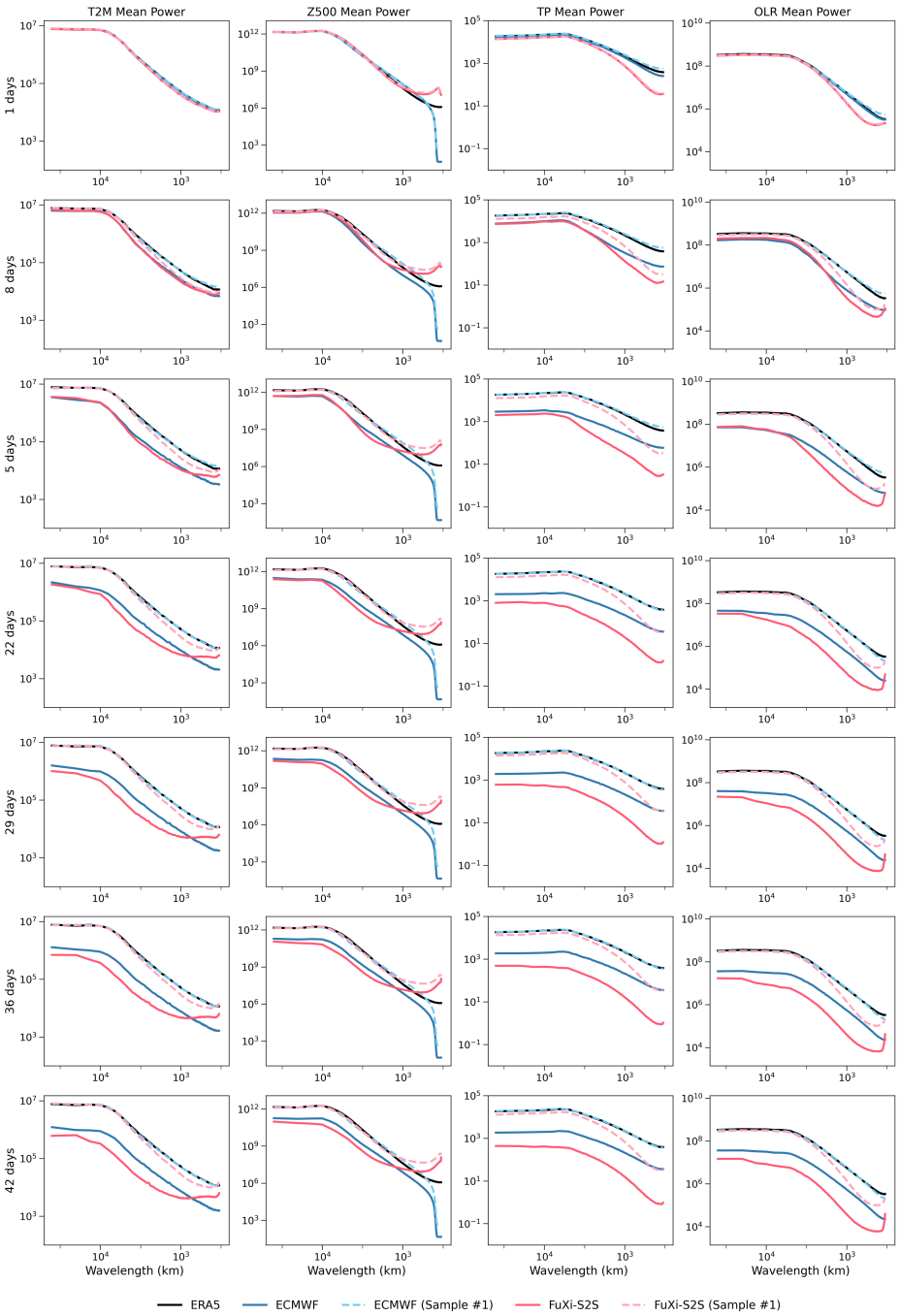
"*Additionally, it would be worthwhile to explore how the prescribed fixed sea surface temperature (SST) or its absence impacts the forecast performance of the MJO. Savarin and Chen [71] demonstrated that the use of a coupled atmosphere-ocean model or the updating of SST with observed values is essential 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*"

[71] 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. Journal of Advances in Modeling Earth Systems 14(6), 2021–002929 (2022)

5. A known problem issue with date-drive weather models is spectra biases (Chattopadhyay and Hassanzadeh, 2023) (also see https://sites.research.google/weatherbench/spectra/). This is particularly problematic for vision transformers including the original FuXi model. I am curious if the authors can comment if there have observed similar artifacts in this model. If not that would be in great interest to the readers and field.

Response: Sure. we have addressed this issue by including plots of the energy spectra along with corresponding descriptions in the supplementary material.

"*Supplementary Figure 4 presents the energy spectra of T2M, Z500, TP, and OLR at seven forecast lead times: 1, 8, 15, 22, 29, 36, and 42 days). This figure demonstrates the effectiveness of the models by showcasing the energy levels across various scales and lead times. The spectra are calculated and presented for both the ensemble mean and a randomly selected ensemble member from the ECMWF S2S reforecasts and FuXi-S2S forecasts. The ERA5 spectra remain consistent across increasing forecast lead times, serving as a baseline to evaluate whether the forecasts become increasingly smoother as the forecast lead times increases. Remarkably, at longer wavelengths, a randomly selected member from either the FuXi-S2S or ECMWF S2S models shows closer alignment with the ERA5 benchmark, suggesting that both models proficiently predict the dominant, larger-scale motions. However, at shorter wavelengths, the FuXi-S2S model initially matches the ERA5 spectra but shows a gradual reduction in energy as forecast lead times increase, indicating increasingly smoother forecasts at smaller scales. In contrast, an ECMWF S2S ensemble member maintains consistent agreement at these smaller scales. Regarding the ensemble mean, both the ECMWF S2S reforecasts and FuXi-S2S forecasts generally exhibit lower energy spectra levels at most forecast lead times compared to both ERA5 data and individual ensemble members. As the lead time increases, the ensemble mean of both models demonstrate a decline in performance at smaller scales, a degradation more significant than that observed in individual ensemble members of the FuXi-S2S model. The notably lower energy levels in the FuXi-S2S model, particularly at longer forecast lead times compared to the ECMWF S2S and ERA5 data, underscore a critical area for model improvement to enhance forecast accuracy and smoothness.*"



Supplementary Figure 4: Energy spectra of T2M (first row) and TP (second row) for ERA5 (in black), ECMWF S2S (in blue) reforecasts and FuXi-S2S forecasts (in red) at forecast lead times of 15 days (first column), 22 days (second column), 29 days (third column), 36 days (fourth column), and 42 days (fifth column), using all testing data between 2017 and 2021.

6. The authors did a great job with the saliency map and the associated discussion. I am excited to see this line of research in future work for ML-based weather forecasting.

Response: We appreciate the positive feedback from the reviewer.

3 Specific Remarks

• "Recent advancements in machine learning for medium- range weather forecasting [25–30] 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 [31]". I suggest adding Nguyen et. al. 2023 (https://arxiv.org/abs/2312.03876) as they have similar resolution and are an ensemble system that shows forecasts out to 15 days.

Response: As recommended by the reviewer, we have included the citation for Nguyen et. al. 2023 (https://arxiv.org/abs/2312.03876).

• Section 4.1 paragraph 2. I do not understand the motivation to use a different OLR product to calculate MJO when ERA5 is used for training. Especially ERA5 was used to verify other portions of the paper.

Response:

Evaluating MJO predictions against MJO indices, which are derived from satellite-observed OLR data, is a standard procedure. For this analysis, we chose the Climate Prediction Center (CPC) OLR (CBO) dataset. Additionally, we evaluated the MJO using ERA5 OLR data. This assessment confirmed the conclusion stated in the main text: FuXi-S2S effectively extends the skillful MJO prediction period from 30 to 36 days. To provide further clarity on this matter, we have incorporated additional sentences into Section 4.1:

"*Evaluating MJO predictions against MJO indices derived from satellite observed OLR data is a common practice.*"

"*It is noteworthy that the MJO indices derived from ERA5 OLR data closely align with those derived from CBO OLR data.*"

• "Specifically, it can complete a comprehensive 42-day forecast with daily time steps in approx- imately 7 seconds using an Nvidia A100 GPU. " Is this a single member or a full ensemble?

Response: We are grateful for the reviewer's observation, which pointed out a lack of clarity in our description. To unequivocally indicate that it pertains to a single member, we have accordingly amended the sentence: "*Specifically, it can complete a comprehensive 42-day forecast with daily time steps in approximately 7 seconds using an Nvidia A100 GPU for a single member.*" This modification ensures the specificity and clarity of our statement regarding the computational capabilities of the system.

• "Each autogressive step undergoes 1000 iterations of training." 1000 epochs or 1000 training steps?

Response: To enhance clarity, we revised the sentence to: "Each autoregressive step undergoes 1000 gradient descent updates."

• Section 2.1 title "Deterministic forecast". This section title is misleading, the forecast is not deterministic (e.g. a single forecast) it is the ensemble mean. I suggest changing this to something like "Deterministic Metrics".

Response: As suggested by the reviewer, we have changed the title of Section 2.1 from "Deterministic forecast" to "Deterministic metrics".

• Section 2.2 title "Ensemble forecast". This section title should probably be changed to "Probabilistic metrics" to better reflect proposed revision to section 2.1.

Response: As suggested by the reviewer, we have changed the title of Section 2.2 from "Ensemble forecast" to "Probabilistic metrics".

• Figure 4. The current date formatting (MMDD) is unconventional and quite difficult to read. I highly recommend change to MM-DD or even MM-DD-YYYY. It took me a minute to even realize they were dates.

Response: As suggested by the reviewer, we have revised the date formatting in Figure 4 from MMDD to MM-DD. The Figure is shown below:

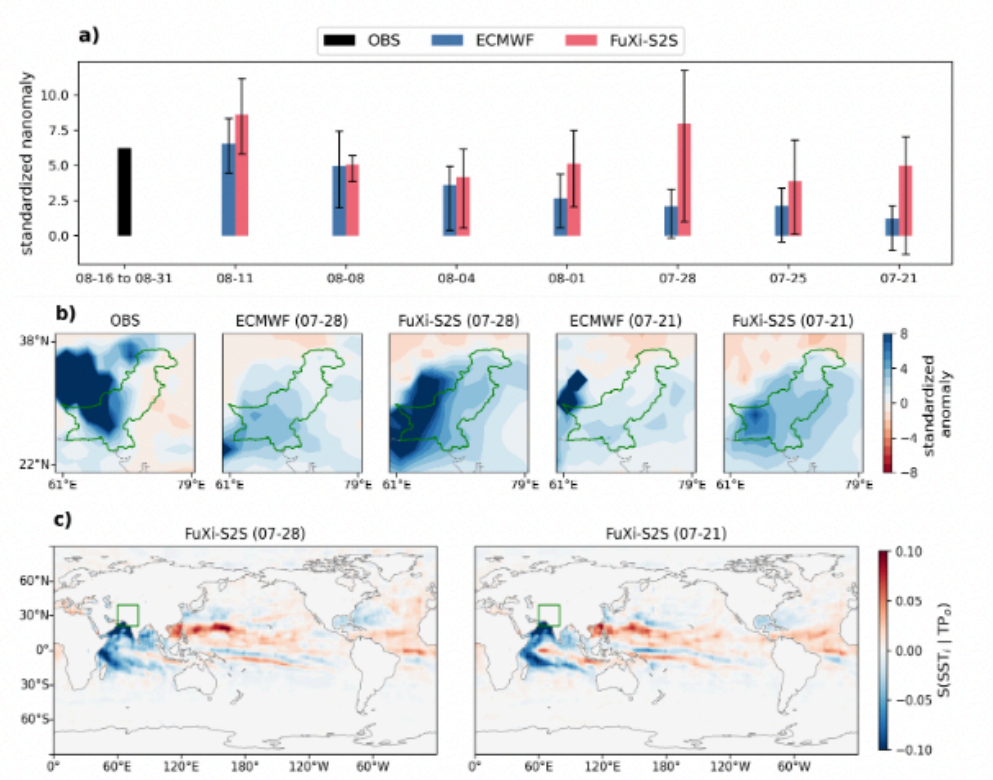


Fig. 4: Comparison of spatially and temporally averaged standardized TP anomaly (a) over the two weeks from August 16th to August 31st, 2022, showcasing GPCP observations (in black) alongside predictions from ECMWF S2S real-time forecasts (in blue) and FuXi-S2S forecasts (in red), with initialization 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 Figure 4 represent the 25th and 75th percentiles of the ensemble forecasts for each start date for both ECMWF and FuXi-S2S models.

• Table 1. There is reference to both outgoing longwave radiation (OLR) and top net thermal radiation (TTR) as variables being predicted. It is my understanding that negative TTR is the exact same as OLR. What is the reason both are included in this table? It implies they are different.

Response:

We appreciate the reviewer's comment that the negative TTR is equivalent to OLR. To avoid any potential confusion, we have removed TTR from Table 1.

References

Chattopadhyay, A., and P. Hassanzadeh, 2023: Long-term instabilities of deep learning-based digital twins of the climate system: The cause and a solution. 2304.07029.

Lavers, D. A., A. Simmons, F. Vamborg, and M. J. Rodwell, 2022: An evaluation of era5 precipitation for climate monitoring. Quarterly Journal of the Royal Meteorological Society, 148 (748), 3152–3165, doi:https: //doi.org/10.1002/qj.4351, URL https://doi.org/10.1002/qj.4351.

Savarin, A., and S. S. Chen, 2022: Pathways to better prediction of the mjo: 2. impacts of atmosphere-ocean coupling on the upper ocean and mjo propagation. Journal of Advances in Modeling Earth Systems, 14 (6), e2021MS002 929, doi:https://doi.org/10.1029/2021MS002929, URL https://doi.org/10.1029/2021MS002929.

Reviewer #3 (Remarks on code availability):

I did not see an input.nc file in the Zenodo project. The code looks runnable besides that.

Response: Thanks for highlighting this issue. We have uploaded the input.nc file to the Zenodo project.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have adequately addressed my comments in the revised version of the manuscript. In particular, the manuscript has been improved in readability and focuses more on the research goal. Therefore, I have no further comments.

Reviewer #2 (Remarks to the Author):

Dear Authors,

Many thanks for your detailed response to my comments and suggestions. I am very happy with the changes you have made and have no further suggestions. I therefore recommend your paper for publication in its current state.

Reviewer #2 (Remarks on code availability):

I have checked the code is present, but I'm afraid I don't have the capacity to check it runs, and there was no-one available in my group to do this check by the deadline.

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for addressing my main concerns with the paper and toning down the sensationalist language. The addition of the SSR plot, SST forecast skill, and spectral plots to the supplementary materials will be of great interest to readers. The addition of the significance testing to the main manuscript improves the overall quality of the paper. Overall the manuscript is worthy of publication in the current form. My only suggestions are to add a reference to a newly developed data-driven seasonal model to the introduction/background section (https://arxiv.org/abs/2406.08632) and do a final proofread for any grammatical mistakes.

Reviewer #3 (Remarks on code availability):

The authors added the requested input files and the code should now be runnable.

Reply to Reviewer #1

The authors have adequately addressed my comments in the revised version of the manuscript. In particular, the manuscript has been improved in readability and focuses more on the research goal. Therefore, I have no further comments.

Response:

We appreciate the reviewer's review and valuable feedback, which has been helpful in enhancing the overall quality of the paper.

Reply to Reviewer #2 (Remarks to the Author):

Dear Authors,

Many thanks for your detailed response to my comments and suggestions. I am very happy with the changes you have made and have no further suggestions. I therefore recommend your paper for publication in its current state.

Response: We appreciate the reviewer's review and valuable feedback, which has been helpful in enhancing the overall quality of the paper.

Reply to Reviewer #2 (Remarks on code availability):

I have checked the code is present, but I'm afraid I don't have the capacity to check it runs, and there was no-one available in my group to do this check by the deadline.

Response:

We welcome reviewers to examine our code at their convenience and invite any questions regarding its running.

Reply to Reviewer #3 (Remarks to the Author):

I would like to thank the authors for addressing my main concerns with the paper and toning down the sensationalist language. The addition of the SSR plot, SST forecast skill, and spectral plots to the supplementary materials will be of great interest to readers. The addition of the significance testing to the main manuscript improves the overall quality of the paper. Overall the manuscript is worthy of publication in the current form. My only suggestions are to add a reference to a newly developed data-driven seasonal model to the introduction/background section (https://arxiv.org/abs/2406.08632) and do a final proofread for any grammatical mistakes.

Response: We appreciate the reviewer's review and valuable feedback, which has been helpful in enhancing the overall quality of the paper. As suggested by the reviewer, we added the following sentence in the "Introduction" section to reference the newly developed data-driven seasonal model:
"Machine learning models have achieved made significant strides in medium-range weather forecasting and seasonal forecasting [33],…"
[33] Wang, C., Pritchard, M.S., Brenowitz, N., Cohen, Y., Bonev, B., Kurth, T., Durran, D., Pathak, J.: Coupled Ocean-Atmosphere Dynamics in a Machine Learning Earth System Model. Preprint at https://arxiv.org/abs/2406.08632 (2024

Reviewer #3 (Remarks on code availability):

The authors added the requested input files and the code should now be runnable.

Response: We welcome reviewers to examine our input files and code at their convenience and invite any questions regarding its running.