Improving the Predictability of the Madden-Julian Oscillation at Subseasonal Scales with Gaussian Process Models

Haoyuan Chen[1], Emil Constantinescu[2], Vishwas Rao[2], Cristiana Stan[3].

[1]Department of Industrial and Systems Engineering, Texas A&M University,
College Station, TX 77843
[2]Mathematics and Computer Science Division, Argonne National Laboratory,
Lemont, IL 60439
[3]Department of Atmospheric, Oceanic and Earth Sciences, George Mason University,
Fairfax, VA 22030

Key Points:

- Propose a probabilistic framework for MJO prediction using Gaussian process models and empirical correlations
- Nonparametric model has a better prediction skill than ANN for the 5 forecast lead days in terms of correlation and overall in terms of RMSE
- The Gaussian process model provides the confidence intervals for the forecast at subseasonal scales (3 weeks) on average

Corresponding author: Emil Constantinescu, `emconsta@anl.gov`

Abstract

The Madden–Julian Oscillation (MJO) is an influential climate phenomenon that plays a vital role in modulating global weather patterns. In spite of the improvement in MJO predictions made by machine learning algorithms, such as neural networks, most of them cannot provide the uncertainty levels in the MJO forecasts directly. To address this problem, we develop a nonparametric strategy based on Gaussian process (GP) models. We calibrate GPs using empirical correlations and we propose a posteriori covariance correction. Numerical experiments demonstrate that our model has better prediction skills than the ANN models for the first five lead days. Additionally, our posteriori covariance correction extends the probabilistic coverage by more than three weeks.

Plain Language Summary

The Madden–Julian Oscillation, or MJO, is a significant weather pattern that affects weather, influencing rainfall, temperature, and even storm frequency and intensity. When the MJO is active, it can affect the weather globally. To better predict weather changes with 3-4 weeks in advance , we rely on the ability to predict the MJO's activity. Data-driven methods such as the ones that rely on deep neural networks have been recently employed to make such predictions. By examining existing MJO patterns, neural networks attempt to predict upcoming ones. However, while neural networks are robust enough to predict the MJO's activity, they do not provide confidence intervals for those predictions. To address this shortcoming, we use a model known as the "Gaussian process" or GP. This statistical tool is distinctive because it not only provides predictions but also quantifies the level of confidence in them.

1 Introduction

The Madden–Julian Oscillation (MJO) (Madden & Julian, 1971, 1972) is the dominant mode of intraseasonal variability of the tropics (Zhang, 2013). In the tropics, the MJO exerts its influence on weather and modulates cyclone activity (Maloney & Hartmann, 2000; Camargo et al., 2009) and El Nino Southern Oscillation (ENSO; Bergman et al., 2001; Lybarger & Stan, 2019). The MJO influence extends outside of the tropics and is one of the important sources of potential predictability on the subseasonal-to-seasonal (S2S) time scales in the extratropics (Stan et al., 2017). Originating in the equatorial Indian Ocean, the MJO propagates eastward along the equator alternating between phases of active and suppressed convection. Traditionally, the amplitude and phase of the MJO have been described by using various MJO indices derived from outgoing longwave radiation (OLR) alone (e.g., OLR MJO Index, OMI; real-time OLR MJO index, ROMI) or in combination with the zonal wind at 850 hPa and 200 hPa (Real-time Multivariate MJO, RMM). The RMM index (Wheeler & Hendon, 2004) consists of a pair (in quadrature) of principal component (PC) time series known as RMM1 and RMM2 ($\text{RMM} = \sqrt{\text{RMM1}^2 + \text{RMM2}^2}$). RMM1 and RMM2 are the first two PCs of combined OLR and zonal winds in the lower (850 hPa) and upper (200 hPa) troposphere averaged between 15S and 15N.

Despite the MJO's pivotal role in the climate system, significant gaps remain in our understanding of its underlying mechanisms. Consequently, climate models struggle to accurately reproduce the observed characteristics of the MJO (G. Chen et al., 2022), and forecast systems face limitations in predicting the MJO with skill beyond a two-week lead time (Kim et al., 2018; Lim et al., 2018; Kim et al., 2019).

Recent advancements in machine learning (ML) applications in predicting geoscientific phenomena spanning from weather to climate (He et al., 2021; Molina et al., 2023) hold the promise of enhancing the skill of deterministic (Love & Matthews, 2009; Toms et al., 2019; Silini et al., 2021; Suematsu et al., 2022; Martin et al., 2022; Hagos et al., 2022) and probabilistic (Delaunay & Christensen, 2022) forecast of the MJO. Improvement in the forecast skill has been achieved also by applying ML techniques for correcting the forecasts of dynamical models (Kim et al., 2021; Silini et al., 2022). The majority of ML models used for MJO prediction are based on artificial neural networks (ANNs). The work of Delaunay and Christensen (2022) uses deep convolutional neural networks (CNNs) to quantify the uncertainty. We note that the probabilistic method in (Delaunay & Christensen, 2022) is not fully data driven. A wide array of ANN architectures has been devised for MJO prediction models. Toms et al. (2019) employed two hidden layers comprising fully connected networks, while Love and Matthews (2009) and Martin et al. (2022) utilized a single hidden layer of fully connected networks. Suematsu et al. (2022) employed recurrent neural networks as a form of reservoir computing, whereas Silini et al. (2022) employed them as autoregressive neural networks.

In terms of input variables, some of the ML models for MJO prediction utilize a selected set of atmospheric state variables, including the OLR and zonal winds, to predict one of the MJO indices (Toms et al., 2019; Delaunay & Christensen, 2022). Others focus solely on the atmospheric state variables required for constructing and predicting the MJO index (Martin et al., 2022). Certain models use the MJO index as both input and output (Love & Matthews, 2009; Suematsu et al., 2022; Silini et al., 2021, 2022) or combine it with other climate indices (Hagos et al., 2022). Some studies suggest that increasing the number of input variables can enhance MJO forecast skill. Nonetheless, models utilizing only the MJO index as a predictor exhibit comparable forecast skill, highlighting the significance of the ML model's characteristics. Thus, the prediction of the MJO can be regarded as a nonparametric problem, while most existing ANN models fall under parametric ML techniques. An alternative avenue for exploration lies in *Gaussian processes* (GPs), which represent a nonparametric learning approach that could be harnessed for MJO prediction. The GP approach has been applied to modeling geophysical datasets such as the prediction of tide height (Roberts et al., 2013). However, this approach is not autoregressive.

Currently, only one of the ML models proposed for MJO forecasting offers the capability to quantify the forecast uncertainty. The model developed by Delaunay and Christensen (2022) predicts both the forecast mean and variance of RMM indices, providing insight into forecast reliability by using a combined model- and data-driven strategy. The model assumes a bivariate Gaussian distribution on the CNN (LeCun et al., 1995). The CNN is trained by maximizing the log-likelihood for each of the forecast lead times. Specifically, the CNN input is a series of daily gridded maps that include zonal wind at 200 hPa and 850 hPa, OLR, sea surface temperature, specific humidity at 400 hPa, geopotential at 850 hPa, and downward longwave radiation at the surface; and the output is the mean and variance of the forecast of RMM1 and RMM2. The output variance represents the intrinsic chaotic (aleatoric) uncertainty in the prediction. In addition, the epistemic uncertainty is estimated by using a Monte Carlo dropout method to produce an ensemble of forecasts. We note, however, that this model assumes no correlation between RMM1 and RMM2 and relies only on the past day t to predict the mean and variance on day $t+\tau$. It overlooks the lag correlation between RMM1 and RMM2 as outlined in CLIVAR (2009) and the potential influences of the values between day t and day $t+\tau$ on the day $t+\tau$. Additionally, interpreting uncertainties derived from neural network (NN) models can be challenging because the influence of weights θ on the NNs is not always clear and NNs may not inherently reflect probabilities. Moreover, the quality of the uncertainty estimates provided by Monte Carlo dropouts depends on choices of architecture designs, and effective design of training procedures is necessary to obtain satisfactory results (Verdoja & Kyrki, 2020). Additionally, the recent short and medium range weather forecasting models such as FourCastNet (Pathak et al., 2022), GenCast (Price et al., 2023), and Aardvark (Vaughan et al., 2024) are not amenable for forecasting MJO.

To address these gaps, we present a novel data-driven and autoregressive probabilistic model for forecasting the MJO amplitude and phase that depends only on the past MJO observations. This model harnesses the power of GPs, enabling us not only to make predictions but also to quantify the inherent uncertainties associated with these forecasts. A

GP is an extension of the multivariate Gaussian distributions to infinite dimensions. In practical terms, this means that given an input vector, the process will return a probability distribution of the observation vector based on the input. As a result, GPs provide a natural way to quantify uncertainty in predictions. GPs offer greater interpretability and transparency compared to NNs. This clarity stems from the GP's covraiance kernel, which provides more readily understandable insights into the model behavior than the complex array of parameters found in NNs (Stein, 1999; Myren & Lawrence, 2021). As statistical models, GPs provide insight into how predictions are made, and the covariance function of a GP reveals the relationships among input features and their impact on predictions. Furthermore, GPs typically involve fewer hyperparameters to tune when compared with NNs, leading to increased computational efficiency.

Specifically, the contributions of this paper are as follows:

- Introduction of a probabilistic framework for the MJO based on GP models that are trained using empirical correlations to improve forecast accuracy.
- Development of a nonparametric strategy utilizing GP models to directly provide uncertainty levels in MJO forecasts that do not rely on ensemble prediction.
- Proposal of a posteriori covariance correction extending probabilistic MJO coverage over three weeks.
- Enhancement of interpretability and transparency compared to neural network models, alongside improved computational efficiency due to fewer hyperparameters.

The paper is organized as follows. In Section 2 we present the data utilized in this study and describe our methodology for forecasting the MJO. In Section 3 we elaborate on the metrics used for analyzing the performance of the proposed model compared to observations and dynamical forecast systems. Section 4 showcases the results we have obtained in this work. In Section 5 we discuss our findings and present directions for future work.

2 Methodology

2.1 Data

The daily MJO RMM index dataset [1] used in the study is provided by the Bureau of Meteorology. RMM1 and RMM2 values are available from June 1, 1974, to the most recent date. Because of missing values before 1979, we select the January 1, 1979, to December 31, 2023, range for our study. The dataset is divided into three subsets: *i*) the training set used to determine the parameters of the prediction and corresponding variance, January, 1, 1979 to December 31, 2016; *ii*) the validation set used to obtain the corrected variance with increasing lags, January 1, 2007 to December 31, 2011; and *iii*) the test set used to verify the results, January 1, 2012 to December 31, 2023. The start of predictions in the validation set (t_v = Jan–01–2007) and test set (t_0 = Jan–01–2012) are part of the model input. The training dataset is further divided into $n = 10,000$ samples of length $L = 40, 60$ days.

2.2 GP model

In this work we obtain the probability distribution of predicted RMM indices. The entire algorithm for our method is described in the diagram shown in Figure 1. The details related to time series prediction and Gaussian process are provided in Appendix A and the mathematical framework of the proposed method in Appendix B.

[1] `http://www.bom.gov.au/climate/mjo/graphics/rmm.74toRealtime.txt`

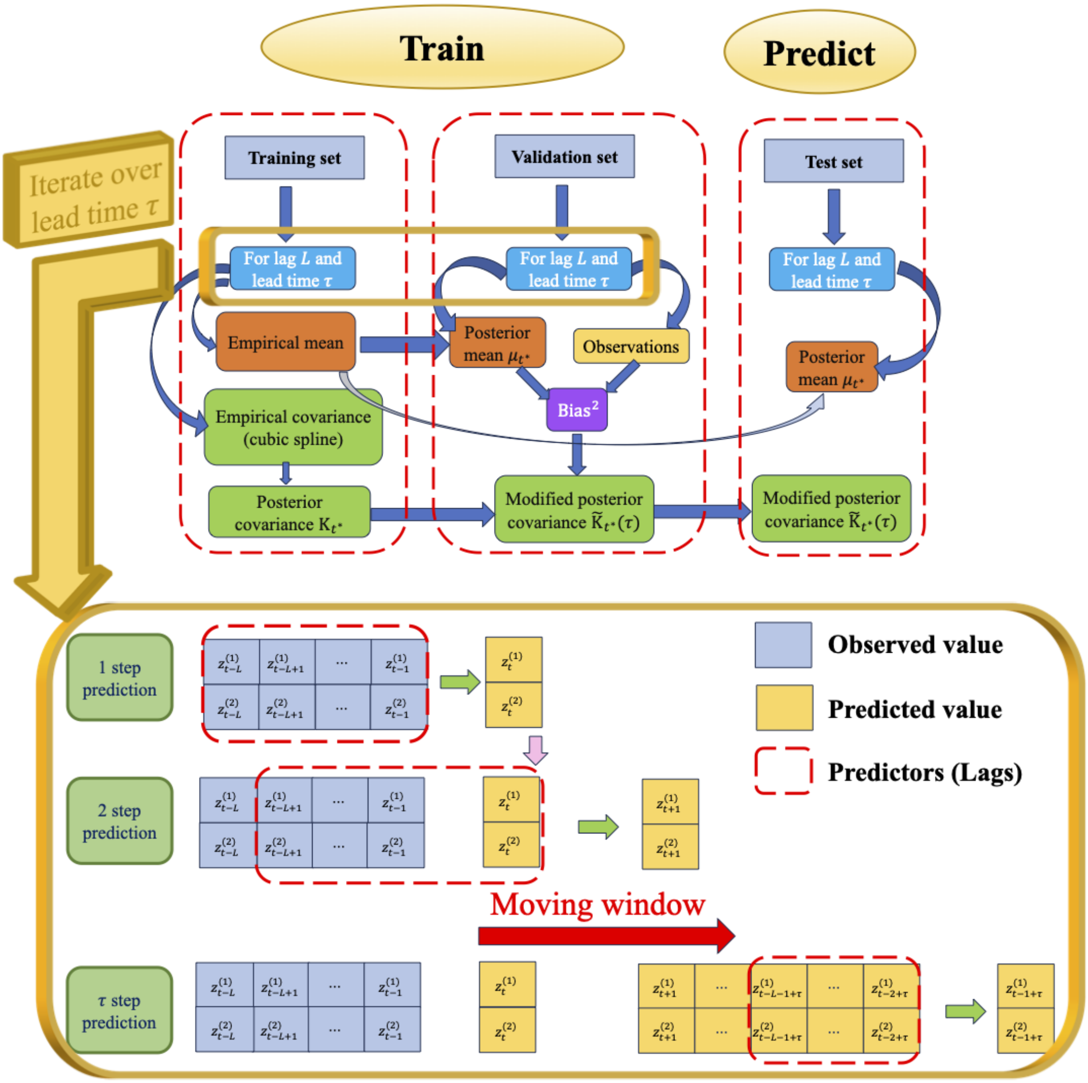
Train
Predict
Iterate over lead time τ
Training set
Validation set
Test set
For lag L and lead time τ
Empirical mean
Posterior mean μ_{t^*}
Observations
Bias^2
Empirical covariance (cubic spline)
Posterior covariance K_{t^*}
Modified posterior covariance $\tilde{K}_{t^*}(\tau)$
1 step prediction
2 step prediction
τ step prediction
Moving window
Observed value
Predicted value
Predictors (Lags)

Figure 1. Flowchart of the entire algorithm. *Top:* Diagram of the GP model for the MJO forecast. The blue arrows indicate the order of operations in the algorithm. t^* represents the predicted timestamp, Bias^2 is the square of the bias between the predicted values and the true observations. *Bottom:* Iterated method for the multistep time series forecasting for two outputs with lag $= L$, lead time $= \tau$ ($\tau > L$). $z_t^{(1)}$, $z_t^{(2)}$ are the values of RMM1 and RMM2 at time t. The green arrows indicate one-day-ahead predictions. The red arrows indicate the moving window of the predictors. Including the predictions from the previous step as predictors in the current step is indicated by the pink arrow. See Appendix B for more details.

We denote the values of RMM1 and RMM2 on the tth day by $z_t^{(1)}$ and $z_t^{(2)}$, respectively. As shown by the diagram in Fig. 1, the input to the GP model is a contiguous time series of RMM1 and RMM2 of length L (blue rectangles). L is referred to as lag in days and corresponds to $T_1 = T_2 = L$ in Appendix A. The goal of this work is to obtain the predictive distribution of the vector $[z_t^{(1)}, z_t^{(2)}]^\top$ (yellow rectangles) at the next τ times conditioned on the previous L days:

$$p\Big(\begin{bmatrix} z^{(1)}_{L+1:L+\tau} \\ z^{(2)}_{L+1:L+\tau} \end{bmatrix} \Big| \Big| \begin{bmatrix} z^{(1)}_{1:L} \\ z^{(2)}_{1:L} \end{bmatrix} ; \Theta \Big) , \tag{1}$$

where Θ is the parameter of the distribution. We will model $[z_t^{(1)}, z_t^{(2)}]^\top$ as a bivariate GP.

The model employs a classical regression algorithm based on one-step-ahead Gaussian process predictions. The one-step-ahead approach involves making predictions at step k using all available information up to step $k-1$. This information is assumed to be Gaussian (normal) distributed. A Gaussian process (GP) is a collection of random variables, such that any finite set of which has multivariate Gaussian distribution (Williams & Rasmussen, 2006). A GP is specified by two functions: the mean function $\mu(\cdot)$ and the covariance function $K(\cdot,\cdot)$. The mean function represents the expected value of the process at any given time. It provides a baseline prediction and captures the trend of the timeseries. The covariance function, also known as the kernel, describes how points in the time series are related to each other. It captures the periodicity and other patterns in the data as well as the uncertainties in the time series. Using the GP model, the time series of RMM1 and RMM2 can be modeled as:

$$f(Z) \sim \mathcal{N}(\mu(Z), K(Z, Z')), \tag{2}$$

where $Z = \begin{bmatrix} z_t^{(1)} \\ z_t^{(2)} \end{bmatrix} = \begin{bmatrix} \text{RMM1}(t) \\ \text{RMM2}(t) \end{bmatrix}$, and $Z' = \begin{bmatrix} z_{t'}^{(1)} \\ z_{t'}^{(2)} \end{bmatrix} = \begin{bmatrix} \text{RMM1}(t') \\ \text{RMM2}(t') \end{bmatrix}$, $f(Z)$ is the bivariate time series of RMMs, where t, t' represent all the time indexes in the series.

During the training, the model takes as input n overlapping batches of RMM1 and RMM2 indices, each of length L. The training data is then divided into an input subset $\mathbf{X}^{(1:2)} = [\mathbf{X}^{(1)}; \mathbf{X}^{(2)}]$ and an output subset $\mathbf{y}^{(1:2)} = [\mathbf{y}^{(1)}; \mathbf{y}^{(2)}]$, each of length $2L$. These subsets are used to estimate an empirical mean by the average of the corresponding subsets. The empirical covariance function is estimated by partitioning the training data into four blocks that represent the covariance between all inputs $\text{Cov}[\mathbf{X}^{(1:2)}, \mathbf{X}^{(1:2)}]$, covariance between all outputs $\text{Cov}[\mathbf{y}^{(1:2)}, \mathbf{y}^{(1:2)}]$, cross-covariance between inputs and outputs $\text{Cov}[\mathbf{X}^{(1:2)}, \mathbf{y}^{(1:2)}]$, and cross-covariance between outputs and inputs $\text{Cov}[\mathbf{y}^{(1:2)}, \mathbf{X}^{(1:2)}]$. The cross- and auto-covariance of the RMMs is modeled using a cubic spline interpolation of the cross- and auto-correlations of the indices, shown in Figure 2.

During the validation, the empirical mean and covariance are used to predict the posterior mean $\boldsymbol{\mu}_{t^*}$ and covariance $\mathbf{K}_{t^*}$ at time t^*. The details of these calculations are provided in the Appendix B1. As the one-step-ahead prediction is iterated forward, the last prediction becomes input for the next prediction (the red dashed rectangle). Therefore, when predictions are carried out into the future, "observations" are replaced by the predictions. As the prediction window moves farther ahead of the start time, more and more components of the input vectors are replaced by GP predictions. This process introduces systematic uncertainties because the covariance is related only to the lag value L and not to the lead time τ of the prediction or the predictor values. At leads beyond L the predictive variance should increase because of the uncertainties introduced by replacing observations with predicted values. The covariance function must be corrected to account for the additional uncertainty. We design the correction by computing the average variance bias between the posterior mean and true observations. This bias is then added to the covariance function at each forecast lead time to obtain the modified posterior covariance $\tilde{\mathbf{K}}_{t^*}(\tau)$. The details of these calculations are provided in the Appendix B2.

One important element of the GP model is the confidence interval of the forecast, which is the confidence region of the normal distribution characterized by the posterior mean and corrected covariance function. Johnson et al. (2002) have shown that $(1-\alpha)$ confidence region of the p-variate (or multivariate) normal distribution is a hyperellipsoid bounded by chi-square distribution with p degrees of freedom at the level α. Since RMMs are bivariate time series, here $p = 2$ in our GP model. Therefore, the ellipsoid of the $(1-\alpha)$ confidence

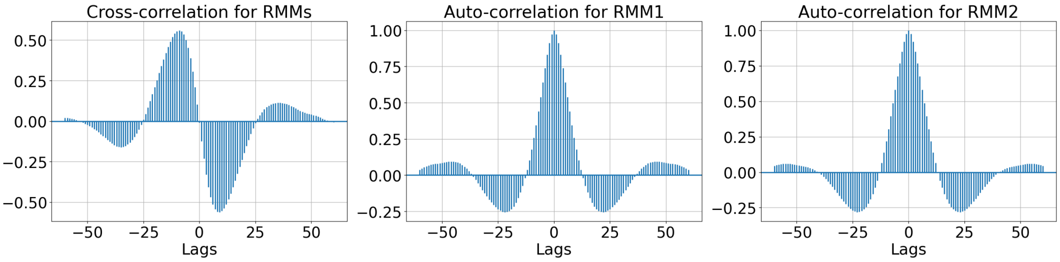

Figure 2. Cross-correlations and auto-correlations of RMMs with maximum lag = 60 days.

region for the GP model is centered on the posterior mean with the axes $\pm\chi_2(\alpha)\sqrt{\lambda_i}\mathbf{e}_i$, $i = 1, 2$, where $\{\lambda_i\}_{i=1}^2$ and $\{\mathbf{e}_i\}_{i=1}^2$ are the eigenvalues and eigenvectors of the corrected covariance $\tilde{\mathbf{K}}_{t^*}(\tau)$.

A limitation of this confidence interval estimation is that it relies on normality assumptions; nevertheless, due to its relatively smooth behavior, it is a reasonable assumption, which is also supported by our numerical results.

3 Metrics

We will use two different types of quantitative metrics to analyze the performance of our models.

3.1 Deterministic prediction skill

For the deterministic prediction skill, we use the predictive mean of the GP model, obtained from equation (B3), as the RMM predictions, denoted by $(\hat{z}_t^{(1)}, \hat{z}_t^{(2)})$ in the subsequent equations. The performance of the model is measured by the bivariate correlation coefficient (COR) and root mean squared error (RMSE) defined as follows:

$$\mathrm{COR}(\tau) = \frac{\sum_{t=1}^{n_p}\left(z_t^{(1)}\hat{z}_t^{(1)}(\tau) + z_t^{(2)}\hat{z}_t^{(2)}(\tau)\right)}{\sqrt{\sum_{t=1}^{n_p}\left(\left(z_t^{(1)}\right)^2 + \left(z_t^{(2)}\right)^2\right)}\sqrt{\sum_{t=1}^{n_p}\left(\left(\hat{z}_t^{(1)}(\tau)\right)^2 + \left(\hat{z}_t^{(2)}(\tau)\right)^2\right)}}, \tag{3}$$

$$\mathrm{RMSE}(\tau) = \sqrt{\frac{1}{n_p}\sum_{t=1}^{n_p}\left(\left(z_t^{(1)} - \hat{z}_t^{(1)}(\tau)\right)^2 + \left(z_t^{(2)} - \hat{z}_t^{(2)}(\tau)\right)^2\right)}, \tag{4}$$

where $z_t^{(1)}$ and $z_t^{(2)}$ are the observations of RMM1 and RMM2 on the tth day in the test set, $\hat{z}_t^{(1)}(\tau)$ and $\hat{z}_t^{(2)}(\tau)$ are the predictions of RMM1 and RMM2 on the tth day in the test set for the lead time of τ days, and n_p is the number of the predictions.

We also analyze the phase error E_ϕ and the amplitude error E_A of RMMs defined as

$$E_\phi(\tau) = \frac{1}{n_p}\sum_{t=1}^{n_p}\left(\hat{P}_t(\tau) - P_t\right), \tag{5}$$

$$E_A(\tau) = \frac{1}{n_p}\sum_{t=1}^{n_p}\left(\hat{A}_t(\tau) - A_t\right), \tag{6}$$

where P_t is the angle in degrees $(0^\circ - 360^\circ)$ of the observation of RMMs $(z_t^{(1)}, z_t^{(2)})$ on the tth day in the test set, $\hat{P}_t(\tau)$ is the angle in degrees $(0^\circ - 360^\circ)$ of the predictions of RMMs $(\hat{z}_t^{(1)}(\tau), \hat{z}_t^{(2)}(\tau))$ on the tth day in the test set for the lead time of τ days. A_t is the observation of RMM amplitude on the tth day in the test set, and $\hat{A}_t(\tau) :=$

$\sqrt{\big(\hat{z}_t^{(1)}(\tau)\big)^2 + \big(\hat{z}_t^{(2)}(\tau)\big)^2}$ is the predicted amplitude on the tth day in the test set for the lead time of τ days. The evaluation is conducted for two values of the lag, $L = 40, 60$, size of the training set $n = 10000$, size of the validation set $n_v = 2000$, number of predictions for computing the errors $n_p = 528$, and forecast lead time $\tau = 1, 2, \ldots, 60$.

To better visualize the skill of the model for the MJO phase, we also assess the model's skill by the Heidke skill score (HSS) (Heidke, 1926) defined in equation (13).

HSS is a measure of how well a forecast is relative to a randomly selected forecast. HSS plays a crucial role in evaluating the accuracy of deterministic forecasts. The definition of HSS (Hyvärinen, 2014) is given by

$$\mathrm{HSS} = \frac{\mathrm{PC} - E}{1 - E} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}, \tag{7}$$

where a, b, c, d are different numbers of cases observed to occur in each category in the contingency table (see Table 1); PC is the proportion correct defined as

$$\mathrm{PC} = \frac{a+d}{a+b+c+d}; \tag{8}$$

E is the expectation of the probability of the correct forecasts defined as

$$E = p(\{z_t \in \mathcal{A}, \hat{z}_t \in \mathcal{A}\} \cup \{z_t \notin \mathcal{A}, \hat{z}_t \notin \mathcal{A}\}) = p(z_t \in \mathcal{A})p(\hat{z}_t \in \mathcal{A}) + p(z_t \notin \mathcal{A})p(\hat{z}_t \notin \mathcal{A}); \tag{9}$$

and its maximum-likelihood estimate is given by

$$E = \Big(\frac{a+c}{a+b+c+d}\Big)\Big(\frac{a+b}{a+b+c+d}\Big) + \Big(\frac{b+d}{a+b+c+d}\Big)\Big(\frac{c+d}{a+b+c+d}\Big). \tag{10}$$

To combine the strong/weak MJO and 8 phases, we divide the plane into 9 parts and introduce phase 0 (inactive MJO) by defining $\{\mathcal{A}_i\}_{i=0}^{8}$ as follows:

# of cases		Observation $z_t \in \mathcal{A}$	
		True	False
Forecast $\hat{z}_t \in \mathcal{A}$	True	a (true positive/hit)	b (false positive/false alarm)
	False	c (false negative/miss)	d (true negative/correct rejection)

Table 1. Contingency table

$$(z_t^{(1)}, z_t^{(2)}) \in \mathcal{A}_0 \iff \sqrt{(z_t^{(1)})^2 + (z_t^{(2)})^2} < 1, \tag{11}$$

$$(z_t^{(1)}, z_t^{(2)}) \in \mathcal{A}_i \ (i = 1, \ldots, 8) \iff \begin{cases} \operatorname{atan2}(z_t^{(2)}, z_t^{(1)}) \in (-\pi, -\frac{3}{4}\pi] + \frac{\pi}{4}(i-1) \\ \text{and} \ \sqrt{(z_t^{(1)})^2 + (z_t^{(2)})^2} \geq 1, \end{cases} \tag{12}$$

where $(z_t^{(1)},\ z_t^{(2)})$ are the observations of (RMM1, RMM2) at time t and atan2 is the 2-argument arctangent function whose range is $(-\pi, \pi]$. For the strong/weak MJO ($i = 0$) and each MJO phase i ($i = 1, \ldots, 8$), we can calculate the corresponding HSS(i) by setting $\mathcal{A} \colon = \mathcal{A}_i$ in equations (11) and (12) and applying them to $\mathcal{A}$ in Table 1. Hence,

$$\mathrm{HSS}(i) = \frac{2(a_i d_i - b_i c_i)}{(a_i + b_i)(b_i + d_i) + (a_i + c_i)(c_i + d_i)}, \tag{13}$$

where $a_i = \mathbf{card}\Big(t\Big|\,(z_t^{(1)}, z_t^{(2)}) \in \mathcal{A}_i \text{ and } (\hat{z}_t^{(1)}, \hat{z}_t^{(2)}) \in \mathcal{A}_i\Big)$; $b_i = \mathbf{card}\Big(t\Big|\,(z_t^{(1)}, z_t^{(2)}) \notin \mathcal{A}_i \text{ and } (\hat{z}_t^{(1)}, \hat{z}_t^{(2)}) \in \mathcal{A}_i\Big)$; $c_i = \mathbf{card}\Big(t\Big|\,(z_t^{(1)}, z_t^{(2)}) \in \mathcal{A}_i \text{ and } (\hat{z}_t^{(1)}, \hat{z}_t^{(2)}) \notin \mathcal{A}_i\Big)$; $d_i =$

$\textbf{card}\left(t\middle|\,(z_t^{(1)}, z_t^{(2)}) \notin \mathcal{A}_i \text{ and } (\hat{z}_t^{(1)}, \hat{z}_t^{(2)}) \notin \mathcal{A}_i\right)$, $i = 0, 1, \ldots, 8$; $(z_t^{(1)},\ z_t^{(2)})$ are the observations of (RMM1, RMM2) at time t; and $(\hat{z}_t^{(1)},\ \hat{z}_t^{(2)})$ are the predictions of (RMM1, RMM2) at time t. Note that $\textbf{card}\,(\cdot)$ denotes the cardinality of the set, which is the number of elements in the set. In our case, it represents the number of days t where the corresponding condition is met.

3.2 Probabilistic prediction skill

The probabilistic nature of the GP model allows a natural evaluation of the probabilistic skill of the MJO prediction. We assess the model using two probabilistic scores: continuous ranked probability score (CRPS) (Hersbach, 2000) and the ignorance score (Roulston & Smith, 2002).

CRPS is a scoring rule that compares a single ground truth value to a cumulative distribution function, first introduced in (Matheson & Winkler, 1976) and widely used in weather forecasts. It is defined as

$$\text{CRPS}(F_D, y) = \int_{\mathbb{R}} \Big(F_D(x) - H(x \geq y)\Big)^2 dx, \tag{14}$$

where F_D is the cumulative distribution function of the forecasted distribution D, H is the Heaviside step function and $y \in \mathbb{R}$ is the observation. We assume the forecasted distribution D is Gaussian distribution, then the CRPS formula is given by

$$\text{CRPS}(\mathcal{N}(\mu, \sigma^2), y) = \sigma \left(\omega(2\Phi(\omega) - 1) + 2\phi(\omega) - \frac{1}{\sqrt{\pi}}\right), \quad \omega = \frac{y - \mu}{\sigma}, \tag{15}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are cumulative distribution function and probability density functions of the standard normal distribution $\mathcal{N}(0, 1)$. The CRPS for MJO is then computed as the sum of the CRPS for RMM1 and RMM2 following (Marshall et al., 2016).

The log-likelihood of the normal distribution is used to compute the ignorance score, which is given as follows:

$$\mathcal{L}(\tau) = \frac{1}{n_p} \sum_{t=1}^{n_p} -\frac{1}{2} \left(\log(2\pi) + \log|\mathbf{\Sigma}_t(\tau)| + \begin{bmatrix} z_t^{(1)} - \hat{z}_t^{(1)}(\tau) \\ z_t^{(2)} - \hat{z}_t^{(2)}(\tau) \end{bmatrix}^\top \mathbf{\Sigma}_t(\tau)^{-1} \begin{bmatrix} z_t^{(1)} - \hat{z}_t^{(1)}(\tau) \\ z_t^{(2)} - \hat{z}_t^{(2)}(\tau) \end{bmatrix} \right), \tag{16}$$

where $\mathbf{\Sigma}_t(\tau) \in \mathbb{R}^{2\times 2}$ is the covariance matrix of the predictions of RMM1 and RMM2 on the tth day for the lead time of τ days, and $|\mathbf{\Sigma}_t(\tau)|$ is the determinant of the covariance matrix $\mathbf{\Sigma}_t(\tau)$.

4 Results

In this section we present the results of the prediction skill of our model in Section 4.1, the results of HSS for each MJO phase over the forecast lead time in Section 4.2, and the visualizations of the uncertainty quantification with the GP model in Section 4.3.

4.1 Prediction skill

Figure 3 presents the results of the prediction skill and errors of the GP model compared to the sub-seasonal to seasonal prediction project (S2S) models, including the European Center for Medium-Range Weather Forecasts (ECMWF) with 35 forecast lead days, Bureau of Meteorology (BOM) with 62 lead days, and Centre National de Recherche Météorologiques (CNRM) with 60 lead days. The metrics are calculated from predictions made on different days for each model, as the S2S models are initialized on different dates. We calculated the metrics for the GP model and ECMWF over the same period from January 3, 2012, to January 10, 2017, and for the BOM and CNRM models over the same period from January

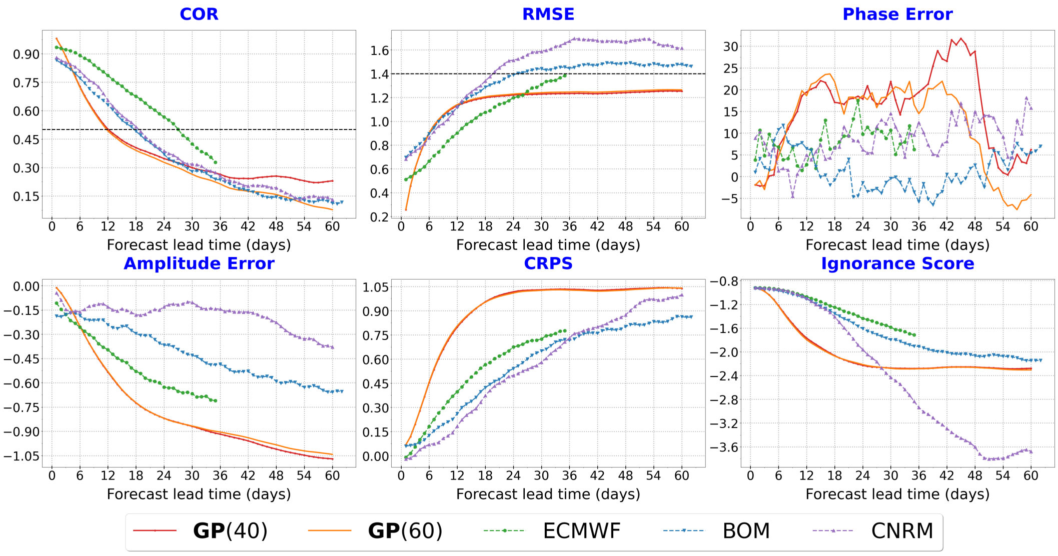

Figure 3. Prediction skill quantifiers and errors of the GP model with lag $L = 40$, 60, respectively, compared to three models in the sub-seasonal to seasonal prediction project (S2S). *Top:* COR, RMSE, and phase error (degress) over 528 predictions. *Bottom:* Amplitude error, CRPS, and ignorance score (log-likelihood) over 528 predictions. Red lines and orange lines represent the GP model with lag $L = 40$ and $L = 60$ respectively, green lines represent the European Center for Medium-Range Weather Forecasts (ECMWF), blue lines represent the Bureau of Meteorology (BOM), purple lines represent the Centre National de Recherche Météorologiques (CNRM).

1, 1993, to December 15, 2014. The values COR $= 0.5$ and RMSE $= 1.4$ are the commonly used skill thresholds for a climatological forecast (Rashid et al., 2011). In this figure we see that our model has a prediction skill of 12 days for both lag $L = 40$ and $L = 60$ with threshold COR $= 0.5$. The ECMWF model demonstrates the best overall performance for COR. While the GP model performs best during the first three forecast lead days, it declines rapidly and eventually reaches similar COR values as the BOM and CNRM models. Regarding the RMSE, the prediction skill is longer than 60 days for $L = 40$ and $L = 60$ with threshold RMSE $= 1.4$. The GP model has a much lower RMSE than S2S models during the first three forecast days, then RMSE increases to values larger than in ECMWF over the next 20 lead days. It eventually stabilizes around RMSE $= 1.25$, outperforming BOM and CNRM across the full 60 forecast lead days. The fast decline of COR for the GP model is due to the fact that we use the empirical correlations from historical RMMs of large size in our model. Specifically, when the forecast lead time increases, the predicted RMMs will become smaller and smoother because of the empirical correlations over a long period of time, giving rise to the smaller variations of RMMs than the true observations and therefore a lower COR. The small value of the predicted RMMs also accounts for the tiny changes in RMSE after day 24 of the forecast lead time. As for the phase error (the angle of RMMs in degrees), we observe that most phase errors for the GP model are positive and larger than ECMWF and CNRM, indicating a faster propagation relative to the observations. For the amplitude errors, we note that all of them are negative. Because of the smaller values of the predicted RMMs of the GP model with forecast lead time increasing, the amplitude is underestimated, resulting in negative and worse amplitude errors than S2S models. The GP model performs worse than the S2S models in terms of probabilistic skill, as measured by CRPS and the ignorance score (log-likelihood). This is due to the larger variances in the GP model, causing its probability distribution to diverge significantly from the observations.

We also note that the results with lags $L = 40$ and $L = 60$ are similar; therefore, for the rest of the paper we will show only results with lags $L = 40$.

4.2 HSS

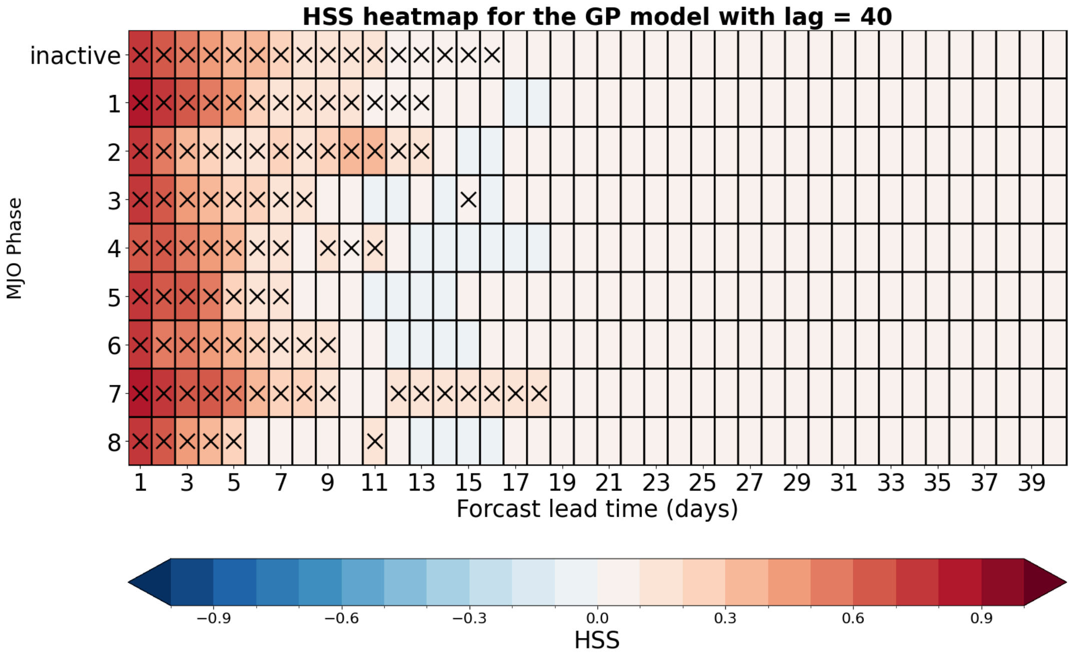

Figure 4. HSS heatmap for the GP model over 528 predictions with lag $L = 40$. The cells with black cross marker "X" represent the significant samples from *Fisher's exact test* with the critical value $\alpha = 0.05$.

Figure 4 shows the HSS heatmap for the combination of phases (1–8) and inactive (weak) MJO for the forecast lead times (1–40 days) over 528 predictions. From this figure we can see that our model has a positive skill for most phases and forecast lead times and has high skill scores for the first 10 forecast lead days for all 8 phases and inactive MJO. We also use Fisher's exact test (Fisher, 1922) with critical value $\alpha = 0.05$ to determine the significant samples for HSS. The cells with the black cross marker in Figure 1 indicate the statistically significant associations between observations and forecasts, which is consistent with the results of Section 4.1 indicating that the model has a good prediction skill within the first 12 days of the forecast lead time. The results reported above provide better skill than the ANN model results reported by (Kim et al., 2021) for the first five forecast lead days in terms of correlation coefficient and overall in terms of root mean square error.

4.3 Uncertainty quantification

Here we pick two samples (Nov–03–2012 to Jan–01–2013, Jan–14–2013 to Mar–14–2013) out of $n_p = 528$ predictions with $\tau = 60$ forecast lead days to present the uncertainty quantification of the predicted MJO. We compare the GP model with the ECMWF ensemble means, including standard deviations from 11 members, which performs best among the S2S models, as well as with observations from BOM. Figure 6 shows an example in which the MJO is mostly inactive within 60 days, and Figure 7 shows an example of an active MJO event. These two examples show that predictions of the GP model capture the general trend seen in observations and outperforms ECMWF during the first 5 lead days. The $\pm\sigma$

confidence intervals (CI) grow as the forecast lead time increases and cover a larger portion of the observation range compared to the ECMWF model's CI. To obtain the complete picture of MJO prediction, we summarize results in Figure 5, which shows the MJO phase diagram for Nov–03–2012 to Jan–01–2013 and Jan–14–2013 to Mar–14–2013 of our model with 68.0% confidence region. The figure clearly shows that almost all observations (black lines) mostly lie within the confidence region (colorful shadings), which demonstrates the quality of the uncertainty quantification of our model. Animated phase diagrams can also be found on the project website `https://gp-mjo.github.io/`, which show how the elliptical confidence region enlarges with time.

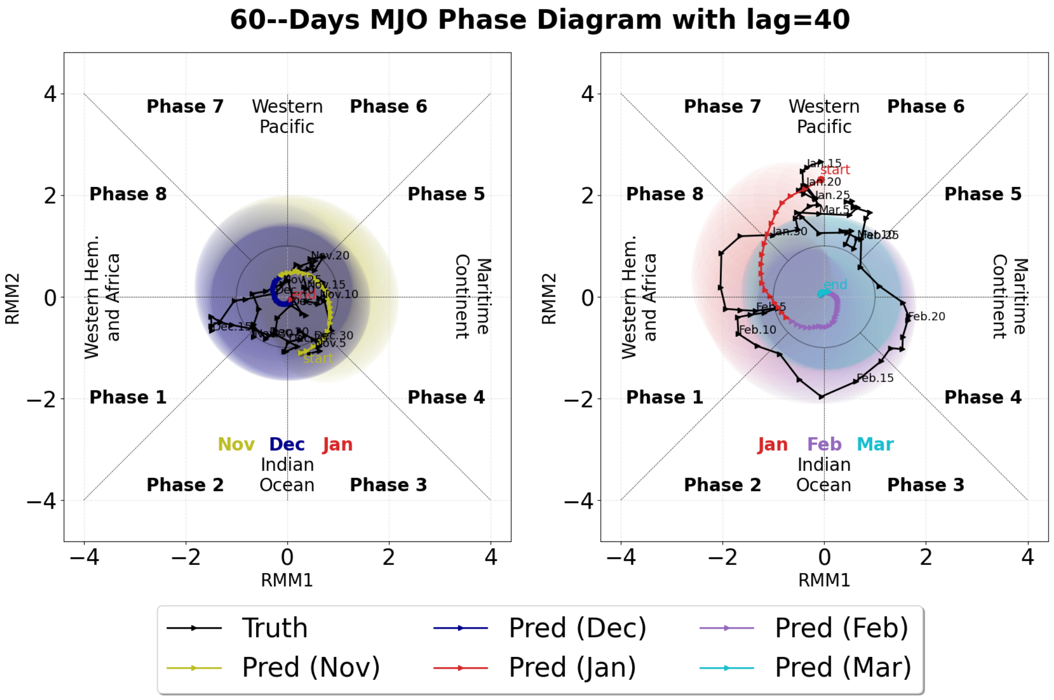

Figure 5. *Left:* 60-days MJO phase diagram for Nov–03–2012 to Jan–01–2013 with lag $L = 40$. Black lines are observations (truth). Olive lines are predictions in November, and olive shadings are 68% confidence regions (CR) in November. Dark blue lines are predictions in December, and dark blue shadings are CR in December. Red lines are predictions in January, and red shadings are CR in January. *Right:* 60-days MJO phase diagram for Jan–14–2013 to Mar–14–2013 with lag $L = 40$. Black lines are observations (truth). Red lines are predictions in January, and red shadings are CR in January. Purple lines are predictions in February, and purple shadings are CI in February. Cyan lines are predictions in March, and cyan shadings are CR in March.

5 Conclusions

In this study we have developed a robust, probabilistic, data-driven model to predict the MJO with high accuracy and quantify prediction uncertainty using GPs with empirical correlations. Our methodology primarily focused on employing the daily RMM index dataset from January 1, 1979, to December 31, 2023, to train, test, and validate the model. We have successfully demonstrated that our model's mean prediction of the daily RMM index remains accurate within a 12-day forecast window, as evidenced by our evaluations using

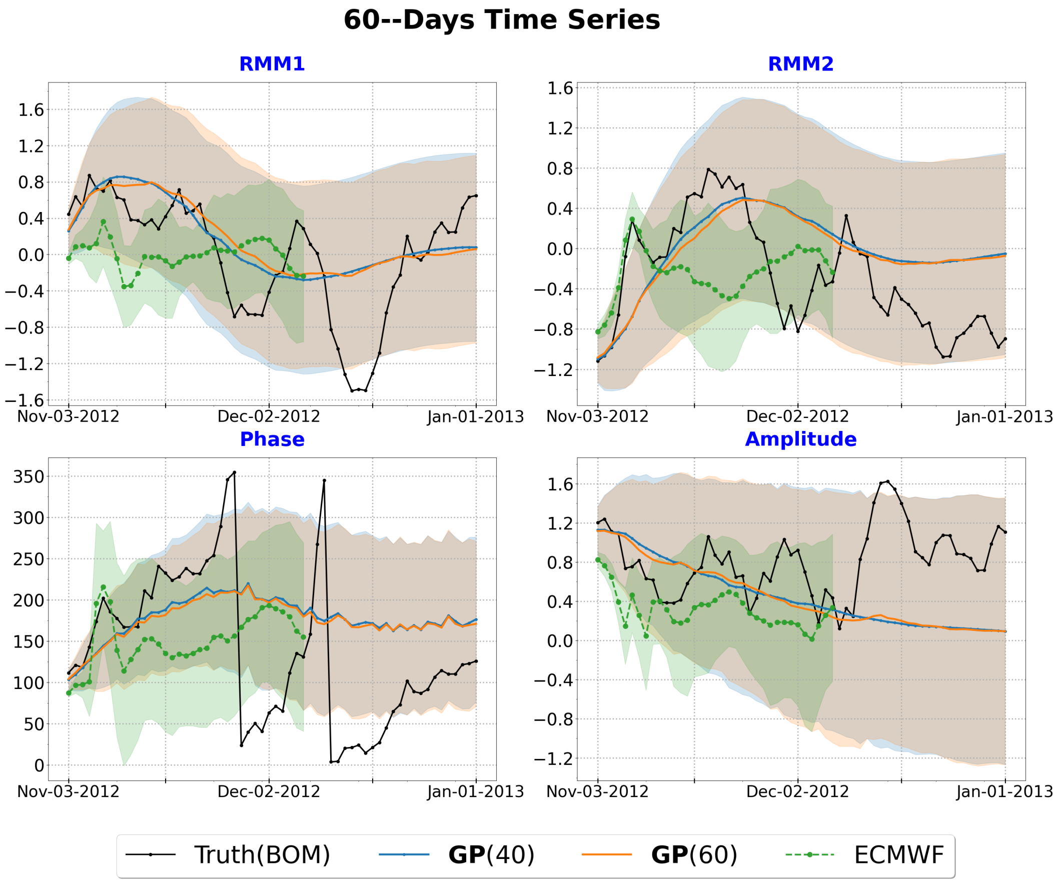

Figure 6. 60-days time series of MJO for Nov–03–2012 to Jan–01–2013 for lag $L = 40, 60$. We denote observations (truth) from the BOM by black dots; predictions of the GP model for lag $L = 40$ and $L = 60$ by blue cross and orange cross, respectively; $\pm\sigma$ CI of the GP model for lag $L = 40$ and $L = 60$ by blue shading and orange shading, respectively; predictions of the ECMWF model by green dots, $\pm\sigma$ CI of the ECMWF model by green shading. *Top left*: Time series of RMM1. *Top right*: Time series of RMM2. *Bottom left*: Time series of phase (angle in the degrees). *Bottom right*: Time series of amplitude.

metrics including the correlation, RMSE, phase errors, amplitude errors, CRPS, ignorance score, and the HSS.

The specific aspect that provides the model's efficacy lies in the approach used to handle GPs for time series prediction and uncertainty quantification. We avoid the typical need for optimizing hyperparameters, thus streamlining the process and enhancing the model's efficiency and stability. This approach is driven by using training data to empirically determine covariance, which is then fitted to a continuous function. The advantage of this method is twofold. It offsets the need for external hyperparameters and ensures stability, especially for long-term predictions, where the model reverts to the mean or prior. Furthermore, our model is robust to the lags of predictors, maintaining accuracy and reliability in predictions without being significantly impacted by lag beyond a certain threshold. This characteristic is especially notable in the context of long-term forecasting and in scenarios where data input may be subject to variable delays.

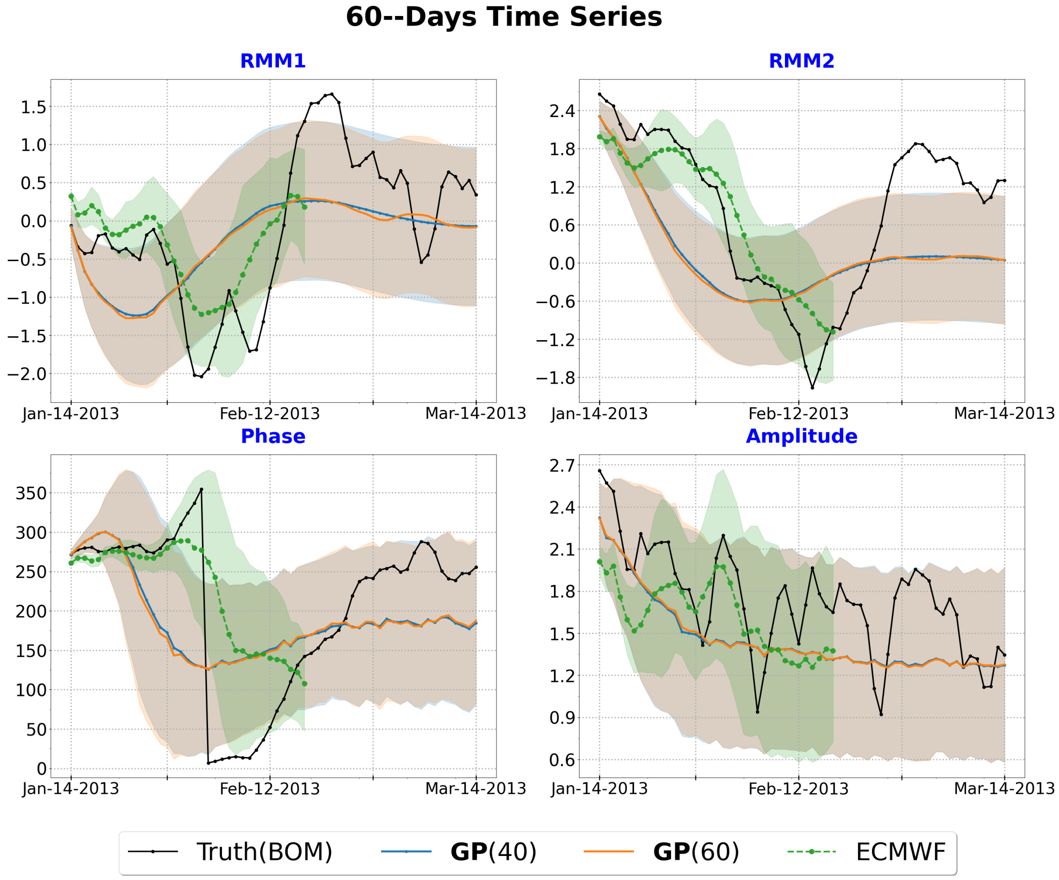

Figure 7. 60-days time series of MJO for Jan–14–2013 to Mar–14–2013 for lag $L = 40, 60$. We denote observations (truth) from the BOM by black dots; predictions of the GP model for lag $L = 40$ and $L = 60$ by blue cross and orange cross, respectively; $\pm\sigma$ CI of the GP model for lag $L = 40$ and $L = 60$ by blue shading and orange shading, respectively; predictions of the ECMWF model by green dots, $\pm\sigma$ CI of the ECMWF model by green shading. *Top left*: Time series of RMM1. *Top right*: Time series of RMM2. *Bottom left*: Time series of phase (angle in the degrees). *Bottom right*: Time series of amplitude.

Moreover, our prediction also provides uncertainty bounds. The uncertainty in our method is state-independent, meaning it is unrelated to the initialized MJO event and depends solely on lead time. The probabilistic model's confidence region covers the observations well, maintaining an average coverage of close to 60 days. This aspect is crucial for reliable forecasting in dynamic and uncertain climatic conditions governed by the MJO. Assuming that the dynamic model fit through a Gaussian process is optimal, this study indeed suggests that the limit of predictability of RMM1 and RMM2 based on their history alone is constrained to the results presented in this paper. Furthermore, it indicates that the memory of the dynamical system, based on these inputs, is limited to about 40 to 60 days in the past.

The approach proposed in this study can be improved by including aspects of seasonal variability and adding additional predictors. In our future work we aim to mitigate these limitations by incorporating seasonal factors into the model and expanding the range of physical variables in the inputs. These aspects are expected to improve our GP model performance significantly. Additionally, while effective, our current empirical approach to constructing GPs could be further advanced by exploring parametric methods in modeling GPs. This future direction could potentially offer more nuanced insights and greater precision in our predictions.

In summary, this study introduces a new data-driven method for predicting the MJO, providing a reliable, efficient, and robust model that provides competitive accuracy and offers extensive insight into prediction uncertainties. As we move forward, our focus will be on refining and enhancing this model to address its current limitations and adapt it to the challenges in climatic forecasting.

Appendix A Background

In this section we review the probabilistic forecasting and the iterative method for the time series forecasting in Section A1 and GP models in Section A2.

A1 Probabilistic forecasting with an iterative method

In the general probabilistic forecasting problem (Rangapuram et al., 2018; Wang et al., 2019), we usually denote M univariate time series by $\{z^{(j)}_{1:T_j}\}_{j=1}^{M}$, where $z^{(j)}_{1:T_j} :=(z^{(j)}_1, z^{(j)}_2, \ldots, z^{(j)}_{T_j})$ is the jth time series and $z^{(j)}_t$ is the value of the jth time series at time t, $1 \leq t \leq T_j$. Our goal is to model the distribution of $z^{(j)}_{T_j+1:T_j+\tau}$ at the next τ time conditioned on the past:

$$p(z^{(j)}_{T_j+1:T_j+\tau} \Big| z^{(j)}_{1:T_j}; \Theta), \quad j = 1, \ldots, M, \tag{A1}$$

where Θ is the set of the learnable parameters shared by all M time series.

The objective of multistep time series forecasting (Weigend, 2018; Cheng et al., 2006; Sorjamaa et al., 2007) is to predict M-variate time series at the next τ time $\{z^{(j)}_{T_j+1:T_j+\tau}\}_{j=1}^{M}$ given $\{z^{(j)}_{1:T_j}\}_{j=1}^{M}$, where $\tau > 1$. A multistep prediction is typically carried out using the iterative method. In this technique, the values computed for each step ahead are sent to the next step as inputs. The iterative method can be written in the autoregressive model as follows:

$$\begin{bmatrix} z^{(1)}_t \\ \vdots \\ z^{(M)}_t \end{bmatrix} = \begin{bmatrix} f_1(z^{(1)}_{t-T_1:t-1}) \\ \vdots \\ f_M(z^{(M)}_{t-T_M:t-1}) \end{bmatrix}, \tag{A2}$$

where $f_1, \ldots, f_M$ are random functions. After the learning process, the predicted values at the next τ time are given by

$$\hat{z}^{(j)}_{t+\tau-1} = \begin{cases} f_j(z^{(j)}_{t-T_j:t-1}) & \text{if } \tau = 1 \\ f_j(z^{(j)}_{t-T_j-1+\tau:t-1}, \hat{z}^{(j)}_{t:t-2+\tau}) & \text{if } \tau = 2, \ldots, T_j \\ f_j(\hat{z}^{(j)}_{t-T_j-1+\tau:t-2+\tau}) & \text{if } \tau = T_j + 1, \ldots, \end{cases} \tag{A3}$$

where $j = 1, \ldots, M$, $\hat{z}^{(j)}_t$ is the predicted value of the jth sequence of time series at time t. The lower diagram in Figure 1 illustrates the case where $M = 2$, $T_1 = T_2 = L$ for the iterated method. The iterated method has also been applied to many classical machine learning models such as *recurrent neural networks* (Medsker & Jain, 2001; Galván & Isasi, 2001; Yunpeng et al., 2017) and *hidden Markov models* (Rabiner & Juang, 1986; Rossi & Gallo, 2006; Horelu et al., 2015).

A2 Gaussian processes

A Gaussian process (Williams & Rasmussen, 2006) is a collection of random variables such that every finite number of which has a multivariate normal distribution. A GP is defined by a mean function $\mu(\cdot)$ and a covariance function $K(\cdot,\cdot)$ and is denoted by $\mathcal{GP}(\mu(\cdot), K(\cdot,\cdot))$.

Given a dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ comprising the inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ (where $\mathbf{x}_i \in \mathbb{R}^d$) and the corresponding observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ (where $y_i \in \mathbb{R}$), suppose $y_i = f(\mathbf{x}_i)$, where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a random function. Gaussian process regression assumes that the unknown function is a prior GP, denoted as $f(\cdot) \sim \mathcal{GP}(\mu(\cdot), K(\cdot,\cdot))$. Then the posterior distribution at a set of test points $\mathbf{X}^* = \{\mathbf{x}_i^*\}_{i=1}^m$ (where $\mathbf{x}_i^* \in \mathbb{R}^d$) has the following form:

$$p(\, f(\mathbf{X}^*)|\, \mathcal{D}) = \mathcal{N}(\mathbb{E}[f(\mathbf{X}^*)|\mathcal{D}], \mathrm{Cov}[f(\mathbf{X}^*)|\mathcal{D}]), \tag{A4}$$

with the posterior mean and covariance as follows:

$$\mathbb{E}[f(\mathbf{X}^*)|\mathcal{D}] = \mu(\mathbf{X}^*) + K(\mathbf{X}^*, \mathbf{X})\big[K(\mathbf{X}, \mathbf{X})\big]^{-1}(\mathbf{y} - \mu(\mathbf{X}))\,, \tag{A5a}$$

$$\mathrm{Cov}[f(\mathbf{X}^*)|\mathcal{D}] = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})\big[K(\mathbf{X}, \mathbf{X})\big]^{-1}K(\mathbf{X}, \mathbf{X}^*). \tag{A5b}$$

Appendix B Algorithm

B1 Empirical GPs for the bivariate time series

Here we denote the bivariate time series of RMMs by $\{z_t^{(j)}\}_{t=1}^T$, $j = 1, 2, \cdots, T$, where T is the length of the entire time series. As before we assume that we model the two time series by a joint GP:

$$\begin{bmatrix} z_t^{(1)} \\ z_t^{(2)} \end{bmatrix} \sim \mathcal{GP}\Big(\mu\big(\begin{bmatrix} z_t^{(1)} \\ z_t^{(2)} \end{bmatrix}\big), K\big(\begin{bmatrix} z_t^{(1)} \\ z_t^{(2)} \end{bmatrix}, \begin{bmatrix} z_{t'}^{(1)} \\ z_{t'}^{(2)} \end{bmatrix}\big)\Big). \tag{B1}$$

We seek to calculate the distribution of the two components at the next time step conditioned on the previous L values. In other words, we need to calculate the predictive distribution of $[z_t^{(1)}, z_t^{(2)}]^\top$ at time t^* for the lag L, which is expressed as

$$p\Big(\begin{bmatrix} z_{t^*}^{(1)} \\ z_{t^*}^{(2)} \end{bmatrix} \Big| \begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix}\Big) = \mathcal{N}(\boldsymbol{\mu}_{t^*}, \mathbf{K}_{t^*}), . \tag{B2}$$

The predictive mean and covariance, $\boldsymbol{\mu}_{t^*} \in \mathbb{R}^{2\times 1}$, $\mathbf{K}_{t^*} \in \mathbb{R}^{2\times 2}$, are estimated by following (B3) and (B4):

$$\begin{aligned}
\boldsymbol{\mu}_{t^*} &= \mathbb{E}\Big[\begin{bmatrix} z_{t^*}^{(1)} \\ z_{t^*}^{(2)} \end{bmatrix} \Big| \begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix}\Big] \\
&= \mathbb{E}\Big[\begin{bmatrix} z_{t^*}^{(1)} \\ z_{t^*}^{(2)} \end{bmatrix}\Big] + \mathrm{Cov}\Big[\begin{bmatrix} z_{t^*}^{(1)} \\ z_{t^*}^{(2)} \end{bmatrix}, \begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix}\Big] \\
&\quad \mathrm{Cov}\Big[\begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix}, \begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix}\Big]^{-1} \left(\begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix} - \mathbb{E}\Big[\begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix}\Big]\right) \\
&\approx \mathbb{E}\Big[\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix}\Big] + \mathrm{Cov}\Big[\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix}, \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}\Big] \\
&\quad \mathrm{Cov}\Big[\begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}, \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}\Big]^{-1} \left(\begin{bmatrix} z_{t^*-L:t^*-1}^{(1)} \\ z_{t^*-L:t^*-1}^{(2)} \end{bmatrix} - \mathbb{E}\Big[\begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}\Big]\right),
\end{aligned} \tag{B3}$$

$$
\begin{aligned}
\mathbf{K}_{t^*} &= \mathrm{Cov}\Big[\begin{bmatrix} z^{(1)}_{t^*} \\ z^{(2)}_{t^*} \end{bmatrix} \Big| \begin{bmatrix} z^{(1)}_{t^*-L:t^*-1} \\ z^{(2)}_{t^*-L:t^*-1} \end{bmatrix} \Big] \\
&= \mathrm{Cov}\Big[\begin{bmatrix} z^{(1)}_{t^*} \\ z^{(2)}_{t^*} \end{bmatrix}, \begin{bmatrix} z^{(1)}_{t^*} \\ z^{(2)}_{t^*} \end{bmatrix} \Big] - \mathrm{Cov}\Big[\begin{bmatrix} z^{(1)}_{t^*} \\ z^{(2)}_{t^*} \end{bmatrix}, \begin{bmatrix} z^{(1)}_{t^*-L:t^*-1} \\ z^{(2)}_{t^*-L:t^*-1} \end{bmatrix} \Big] \\
&\quad \mathrm{Cov}\Big[\begin{bmatrix} z^{(1)}_{t^*-L:t^*-1} \\ z^{(2)}_{t^*-L:t^*-1} \end{bmatrix}, \begin{bmatrix} z^{(1)}_{t^*-L:t^*-1} \\ z^{(2)}_{t^*-L:t^*-1} \end{bmatrix} \Big]^{-1} \mathrm{Cov}\Big[\begin{bmatrix} z^{(1)}_{t^*-L:t^*-1} \\ z^{(2)}_{t^*-L:t^*-1} \end{bmatrix}, \begin{bmatrix} z^{(1)}_{t^*} \\ z^{(2)}_{t^*} \end{bmatrix} \Big] \\
&\approx \mathrm{Cov}\Big[\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix}, \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} \Big] - \mathrm{Cov}\Big[\begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix}, \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} \Big] \\
&\quad \mathrm{Cov}\Big[\begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}, \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} \Big]^{-1} \mathrm{Cov}\Big[\begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}, \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} \Big],
\end{aligned}
\tag{B4}
$$

where

$$
\mathbf{X}^{(j)} = \begin{bmatrix} z^{(j)}_{1:L} \\ z^{(j)}_{2:L+1} \\ \vdots \\ z^{(j)}_{n:L+n-1} \end{bmatrix}^\top \in \mathbb{R}^{L\times n}, \quad \mathbf{y}^{(j)} = \begin{bmatrix} z^{(j)}_{L+1} \\ z^{(j)}_{L+2} \\ \vdots \\ z^{(j)}_{L+n} \end{bmatrix}^\top \in \mathbb{R}^{1\times n}, \quad j = 1, 2, \tag{B5}
$$

and

$$
\mathbf{X}^{(1:2)} := \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} \in \mathbb{R}^{2L\times n}, \quad \mathbf{y}^{(1:2)} := \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \end{bmatrix} \in \mathbb{R}^{2\times n}. \tag{B6}
$$

In equations (B3) and (B4) we use the empirical mean and covariance of n batches of training data with lag L to approximate the expectation of the target and the covariance of the target and predictors.

B2 Covariance update

The forecast lead time is reached by repeated one-step predictions. Therefore, the covariance $\mathbf{K}_{t^*}$ in equation (B4) is related only to the value of lag L, which is 40 or 60 in our study and is unrelated to the lead time τ or the predictor values. However, as we predict for longer lead times, the predictive variance should increase because of the uncertainties introduced by replacing observations by predicted values. To account for this additional uncertainty, we propose the following covariance correction. For each lead time we use a validation set of size $n_v(L)$ with lag L to compute the averaged variance bias between the posterior mean and true observations. Hence, the corrected variance $\tilde{V}^{(j)}_*(\tau)$ is given by

$$
\begin{aligned}
\tilde{V}^{(j)}_*(\tau) := \mathrm{Var}[z^{(j)}_{t^*}(\tau)] &\approx \mathrm{Var}[\hat{z}^{(j)}_{t^*}(\tau)] + \mathrm{Bias}\Big(\hat{z}^{(j)}_{t^*}(\tau), z^{(j)}_{t^*}(\tau)\Big)^2, \\
&\approx \mathbf{K}_{t^*}[j,j] + \frac{1}{n_v}\sum_{t=1}^{n_v}\big(\hat{z}^{(j)}_t(\tau) - z^{(j)}_t(\tau)\big)^2,
\end{aligned}
\tag{B7}
$$

where $\hat{z}^{(j)}_{t^*}(\tau)$ is the predicted value for lead time τ obtained by the above iteration, $z^{(j)}_{t^*}(\tau)$ is the corresponding true observation, and $\mathbf{K}_{t^*}[j,j]$ is the $[j,j]$th entry of the covariance matrix $\mathbf{K}_{t^*}$, $j = 1, 2$. Then we scale the $\mathbf{K}_{t^*}$ to the corrected covariance $\tilde{\mathbf{K}}_{t^*}(\tau)$ for lead time τ in (B8) by using the variances $\{\tilde{V}^{(j)}_*(\tau)\}^2_{j=1}$. Therefore, the corrected covariance $\tilde{\mathbf{K}}_{t^*}(\tau)$ corresponds to the lead time τ and can be scaled via the following transformation:

$$
\mathbf{K}_{t^*} = \begin{bmatrix} \mathbf{K}_{t^*}[1,1] & \mathbf{K}_{t^*}[1,2] \\ \mathbf{K}_{t^*}[2,1] & \mathbf{K}_{t^*}[2,2] \end{bmatrix} \longrightarrow \tilde{\mathbf{K}}_{t^*}(\tau) = \begin{bmatrix} \tilde{V}^{(1)}_*(\tau) & \frac{\mathbf{K}_{t^*}[1,2]\sqrt{\tilde{V}^{(1)}_*(\tau)}\sqrt{\tilde{V}^{(2)}_*(\tau)}}{\sqrt{\mathbf{K}_{t^*}[1,1]}\sqrt{\mathbf{K}_{t^*}[2,2]}} \\ \frac{\mathbf{K}_{t^*}[2,1]\sqrt{\tilde{V}^{(1)}_*(\tau)}\sqrt{\tilde{V}^{(2)}_*(\tau)}}{\sqrt{\mathbf{K}_{t^*}[1,1]}\sqrt{\mathbf{K}_{t^*}[2,2]}} & \tilde{V}^{(2)}_*(\tau), \end{bmatrix} \tag{B8}
$$

where $\tilde{V}^{(1)}_*(\tau)$ and $\tilde{V}^{(2)}_*(\tau)$ are defined in equation (B7). This corrected covariance is ultimately used to estimate the confidence region described below.

Workflow	Parameters
Input	n : number of samples in the training dataset L : number of lags t_v : start index for the predictions in validation dataset t_0 : start index for the predictions in testing dataset τ : forecast lead time $\{[z_t^{(1)}, z_t^{(2)}]\}_{t=1}^{L+n}$: training dataset $\{[z_t^{(1)}, z_t^{(2)}]\}_{t=t_v}^{t_v+L+\tau+n_v-2}$: validation set $\{[z_t^{(1)}, z_t^{(2)}]\}_{t=t_0-L}^{t_0-1}$: starting predictors in test set$\}$
Computation steps	1. Construct the training dataset $\mathcal{D}^{(1:2)} = \{\mathbf{X}^{(1:2)}, \mathbf{y}^{(1:2)}\}$ by equations (B6) and (B5), $\mathbf{X}^{(1:2)} \in \mathbb{R}^{2L\times n}$, $\mathbf{y}^{(1:2)} \in \mathbb{R}^{2\times n}$ 2. Compute $\mathbb{E}[\mathbf{y}^{(1:2)}]$ 3. Obtain $\mathrm{Cov}\Big[\begin{bmatrix}\mathbf{X}^{(1:2)} \\ \mathbf{y}^{(1:2)}\end{bmatrix}, \begin{bmatrix}\mathbf{X}^{(1:2)} \\ \mathbf{y}^{(1:2)}\end{bmatrix} \Big] = \begin{bmatrix}\mathrm{Cov}[\mathbf{X}^{(1:2)}, \mathbf{X}^{(1:2)}] & \mathrm{Cov}[\mathbf{X}^{(1:2)}, \mathbf{y}^{(1:2)}] \\ \mathrm{Cov}[\mathbf{y}^{(1:2)}, \mathbf{X}^{(1:2)}] & \mathrm{Cov}[\mathbf{y}^{(1:2)}, \mathbf{y}^{(1:2)}]\end{bmatrix}$ by cubic spline interpolation 4. In the validation set, obtain the $\{\boldsymbol{\mu}_t, \mathbf{K}_t\}_{t=t_v+L+i-1}^{t_v+L+\tau+i-2}$ condition on $\{[z_t^{(1)}, z_t^{(2)}]\}_{t=t_v+t-1}^{t_v+L+i-2}$ for $i = 1, \ldots, n_v$ by equations (B3) and (B4); here $\mathbf{K}_t$ is equivalent for all t 5. In the validation set, obtain modified covariances as a function of lead time $\{\tilde{\mathbf{K}}_{t_v}(t - t_v + 1)\}_{t=t_v}^{t_v+\tau-1}$ by (B7) and (B8) 6. In the test set, obtain $\{\boldsymbol{\mu}_t\}_{t=t_0}^{t_0+\tau-1}$ by equation (B3) 7. In the test set, apply the covariances obtained in the validation set to the covariances in the test set according to the corresponding lead time, $\tilde{\mathbf{K}}_{t_0}(l) \leftarrow \tilde{\mathbf{K}}_{t_v}(l)$, $l = 1, \ldots, \tau$ 8. Return $\boldsymbol{\mu}_t$, $\tilde{\mathbf{K}}_{t_0}(t - t_0 + 1)$, $t = t_0, \ldots, t_0 + \tau - 1$
Output	$\{\boldsymbol{\mu}_t\}_{t=t_0}^{t_0+\tau-1}$: predicted mean of $\{[\hat{z}_t^{(1)}, \hat{z}_t^{(2)}]\}_{t=t_0}^{t_0+\tau-1}$ $\{\tilde{\mathbf{K}}_{t_0}(t - t_0 + 1)\}_{t=t_0}^{t_0+\tau-1}$: predicted covariance of $\{[\hat{z}_t^{(1)}, \hat{z}_t^{(2)}]\}_{t=t_0}^{t_0+\tau-1}$

Table B1. GP model for the MJO forecast

B3 Estimation of the confidence region

To obtain the confidence region of the distribution $\mathcal{N}(\boldsymbol{\mu}_{t^*}, \tilde{\mathbf{K}}_{t^*}(\tau))$, we first introduce Lemmas Appendix B.1 and Appendix B.2 as follows.

Lemma Appendix B.1. (Result 4.7 in Section 4.2 in (Johnson et al., 2002)) *Let $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a p-variate normal distribution with location $\boldsymbol{\mu}$ and known covariance $\boldsymbol{\Sigma}$. Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then*

(a) $(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is distributed as χ_p^2, where χ_p^2 denotes the chi-square distribution with p degrees of freedom.

(b) The $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution assigns probability $1 - \alpha$ to the solid hyperellipsoid $\{\mathbf{x} \colon (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\}$, where $\chi_p^2(\alpha)$ denotes the upper (100α)th percentile of the χ_p^2 distribution.

Proof. See proof of Result 4.7 in Section 4.2 in (Johnson et al., 2002). □

Lemma Appendix B.2. ((4-7) in Section 4.2 in (Johnson et al., 2002)) *The hyperellipsoids $\{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2\}$ are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\,\mathbf{e}_i$, where λ_i's, $\mathbf{e}_i$'s are the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$, namely, $\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, $i = 1, 2, \ldots, p$.*

Proof. From Result 4.1 in Section 4.2 in (Johnson et al., 2002) we know that if $\mathbf{\Sigma}$ is positive definite and $\mathbf{\Sigma}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, then $\lambda_i > 0$ and $\mathbf{\Sigma}^{-1}\mathbf{e}_i = \frac{1}{\lambda_i}\mathbf{e}_i$. That is, $(\frac{1}{\lambda_i}, \mathbf{e}_i)$ is an eigenvalue-eigenvector pair for $\mathbf{\Sigma}^{-1}$. According to the definition of the hyperellipsoid in quadratic form, we can conclude that the hyperellipsoids $\{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2\}$ are centered at $\boldsymbol{\mu}$ and have axes $\pm c\sqrt{\lambda_i}\,\mathbf{e}_i$. □

According to the above lemmas, the $(1 - \alpha)$ confidence region of the p-variate normal distribution is a hyperellipsoid bounded by $\chi_p^2(\alpha)$, the chi-square distribution with p degrees of freedom at the level α (Johnson et al., 2002). Therefore, we can construct a confidence region for the prediction $[\hat{z}_{t^*}^{(1)}(\tau), \hat{z}_{t^*}^{(2)}(\tau)]^\top$ at lead time τ, where $[\hat{z}_{t^*}^{(1)}(\tau), \hat{z}_{t^*}^{(2)}(\tau)]^\top \sim \mathcal{N}(\boldsymbol{\mu}_{t^*}, \tilde{\mathbf{K}}_{t^*}(\tau))$ after updating the covariance.

Data Availability Statement

The daily MJO RMM index dataset is available through the Bureau of Meteorology (`http://www.bom.gov.au/`) and can be accessed at `http://www.bom.gov.au/climate/mjo/`. The codes for the numerical experiments in this work can be found at `https://doi.org/10.5281/zenodo.13654353` (H. Chen et al., 2024).

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) and the SciDAC FASTMath Institute programs under Contract No. DE-AC02-06CH11357 and Argonne National Laboratory Directed Research and Development (LDRD) program. We would also like to thank the reviewers for their constructive criticism and Hannah Christensen for her valuable help with using and interpretation of the ECMWF MJO dataset.

References

Bergman, J. W., Hendon, H. H., & Weickmann, K. M. (2001). Intraseasonal air–sea interactions at the onset of El Niño. *Journal of Climate*, *14*(8), 1702–1719.

Camargo, S. J., Wheeler, M. C., & Sobel, A. H. (2009). Diagnosis of the MJO modulation of tropical cyclogenesis using an empirical index. *Journal of the Atmospheric Sciences*, *66*(10), 3061–3074.

Chen, G., Ling, J., Zhang, R., Xiao, Z., & Li, C. (2022). The MJO from CMIP5 to CMIP6: Perspectives from tracking MJO precipitation. *Geophysical Research Letters*, *49*(1), e2021GL095241.

Chen, H., Constantinescu, E., Rao, V., & Stan, C. (2024, September). *Reproduce the results in the paper "Improving the Predictability of the Madden-Julian Oscillation at Subseasonal Scales with Gaussian Process Models"*. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.13654353` doi: 10.5281/zenodo.13654353

Cheng, H., Tan, P.-N., Gao, J., & Scripps, J. (2006). Multistep-ahead time series prediction. In *Advances in knowledge discovery and data mining: 10th pacific-asia conference, pakdd 2006, singapore, april 9-12, 2006. proceedings 10* (pp. 765–774).

CLIVAR, M.-J. O. W. G. (2009). MJO Simulation Diagnostics. *Journal of Climate*, *22*(11), 3006–3030.

Delaunay, A., & Christensen, H. M. (2022). Interpretable deep learning for probabilistic MJO prediction. *Geophysical Research Letters*, *49*(16), e2022GL098566.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94.

Galván, I. M., & Isasi, P. (2001). Multi-step learning rule for recurrent neural models: an application to time series forecasting. *Neural Processing Letters*, *13*, 115–133.

Hagos, S., Leung, L. R., Zhang, C., & Balaguru, K. (2022). An observationally trained Markov model for MJO propagation. *Geophysical Research Letters*, *49*(2), e2021GL095663.

He, S., Li, X., DelSole, T., Ravikumar, P., & Banerjee, A. (2021). Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 169–177).

Heidke, P. (1926). Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, *8*(4), 301–349.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Horelu, A., Leordeanu, C., Apostol, E., Huru, D., Mocanu, M., & Cristea, V. (2015). Forecasting techniques for time series from sensor data. In *2015 17th international symposium on symbolic and numeric algorithms for scientific computing (synasc)* (pp. 261–264).

Hyvärinen, O. (2014). A probabilistic derivation of Heidke skill score. *Weather and Forecasting*, *29*(1), 177–181.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis.* Prentice hall Upper Saddle River, NJ.

Kim, H., Ham, Y., Joo, Y., & Son, S. (2021). Deep learning for bias correction of MJO prediction. *Nature Communications*, *12*(1), 3087.

Kim, H., Janiga, M. A., & Pegion, K. (2019). MJO propagation processes and mean biases in the SubX and S2S reforecasts1. *Journal of Geophysical Research: Atmospheres*, *124*(16), 9314–9331.

Kim, H., Vitart, F., & Waliser, D. E. (2018). Prediction of the Madden–Julian oscillation: A review. *Journal of Climate*, *31*(23), 9425–9443.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Lim, Y., Son, S.-W., & Kim, D. (2018). MJO prediction skill of the subseasonal-to-seasonal prediction models. *Journal of Climate*, *31*(10), 4075–4094.

Love, B. S., & Matthews, A. J. (2009). Real-time localised forecasting of the Madden-Julian Oscillation using neural network models. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *135*(643), 1471–1483.

Lybarger, N. D., & Stan, C. (2019). Revisiting MJO, Kelvin waves, and El Niño relationships using a simple ocean model. *Climate Dynamics*, *53*(9-10), 6363–6377.

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *Journal of Atmospheric Sciences*, *28*(5), 702–708.

Madden, R. A., & Julian, P. R. (1972). Description of global-scale circulation cells in the tropics with a 40–50 day period. *Journal of Atmospheric Sciences*, *29*(6), 1109–1123.

Maloney, E. D., & Hartmann, D. L. (2000). Modulation of eastern North Pacific hurricanes by the Madden–Julian oscillation. *Journal of Climate*, *13*(9), 1451–1460.

Marshall, A. G., Hendon, H. H., & Hudson, D. (2016). Visualizing and verifying probabilistic forecasts of the Madden-Julian Oscillation. *Geophysical Research Letters*, *43*(23), 12–278.

Martin, Z. K., Barnes, E. A., & Maloney, E. (2022). Using simple, explainable neural networks to predict the madden–Julian oscillation. *Journal of Advances in Modeling Earth Systems*, *14*(5), e2021MS002774.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, *22*(10), 1087–1096.

Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. *Design and Applications*, *5*(64-67), 2.

Molina, M. J., O'Brien, T. A., Anderson, G., Ashfaq, M., Bennett, K. E., Collins, W. D., . . . Ullrich, P. A. (2023). A review of recent and emerging machine learning applications for climate variability and Wweather phenomena. *Artificial Intelligence for the Earth Systems*, 1–46.

Myren, S., & Lawrence, E. (2021). A comparison of Gaussian processes and neural networks for computer model emulation and calibration. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *14*(6), 606–623.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... others (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., ... Willson, M. (2023). GenCast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*.

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16.

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., & Januschowski, T. (2018). Deep state space models for time series forecasting. *Advances in Neural Information Processing Systems*, *31*.

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

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibbson, N., & Aigrain, S. (2013). Gaussian processes for time–series modeling. *Phylosophical Transactions of the Royal Society A*, 371.

Rossi, A., & Gallo, G. M. (2006). Volatility estimation via hidden Markov models. *Journal of Empirical Finance*, *13*(2), 203–230.

Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*(6), 1653–1660.

Silini, R., Barreiro, M., & Masoller, C. (2021). Machine learning prediction of the Madden–Julian Oscillation. *npj Climate and Atmospheric Science*, *4*(1), 57.

Silini, R., Lerch, S., Mastrantonas, N., Kantz, H., Barreiro, M., & Masoller, C. (2022). Improving the prediction of the Madden–Julian oscillation of the ECMWF model by post-processing. *Earth System Dynamics*, *13*(3), 1157–1165.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., & Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing*, *70*(16-18), 2861–2869.

Stan, C., Straus, D. M., Frederiksen, J. S., Lin, H., Maloney, E. D., & Schumacher, C. (2017). Review of tropical-extratropical teleconnections on intraseasonal time scales. *Reviews of Geophysics*, *55*(4), 902–937.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging.* Springer Science & Business Media.

Suematsu, T., Nakai, K., Yoneda, T., Takasuka, D., Jinno, T., Saiki, Y., & Miura, H. (2022). Machine learning prediction of the MJO extends beyond one month. *arXiv preprint arXiv:2301.01254*.

Toms, B. A., Kashinath, K., Yang, D., et al. (2019). Testing the reliability of interpretable neural networks in geoscience using the Madden–Julian Oscillation. *arXiv preprint arXiv:1902.04621*.

Vaughan, A., Markou, S., Tebbutt, W., Requeima, J., Bruinsma, W. P., Andersson, T. R., ... Turner, R. E. (2024). Aardvark Weather: end-to-end data-driven weather forecasting. *arXiv preprint arXiv:2404.00411*.

Verdoja, F., & Kyrki, V. (2020). Notes on the behavior of MC dropout. *arXiv preprint arXiv:2008.02627*.

Wang, Y., Smola, A., Maddix, D., Gasthaus, J., Foster, D., & Januschowski, T. (2019). Deep factors for forecasting. In *International conference on machine learning* (pp. 6607–6617).

Weigend, A. S. (2018). *Time series prediction: forecasting the future and understanding the past.* Routledge.

Wheeler, M. C., & Hendon, H. H. (2004). An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Monthly Weather Review*, *132*(8), 1917–1932.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2) (No. 3). MIT Press Cambridge, MA.

Yunpeng, L., Di, H., Junpeng, B., & Yong, Q. (2017). Multi-step ahead time series forecasting for different data patterns based on LSTM recurrent neural network. In *2017 14th web information systems and applications conference (wisa)* (pp. 305–310).

Zhang, C. (2013). Madden–Julian oscillation: Bridging weather and climate. *Bulletin of the American Meteorological Society*, *94*(12), 1849–1870.