Low-order stochastic model and “past-noise forecasting” of the Madden-Julian oscillation

D. Kondrashov, M.D. Chekroun, A. W. Robertson, M. Ghil

Corresponding author: D. Kondrashov, Department of Atmospheric and Oceanic Sciences, 405 Hilgard Ave, Box 951565, 7127 Math Sciences Bldg. University of California, Los Angeles, CA 90095-1565, U.S.A. (dkondras@atmos.ucla.edu)

1Department of Atmospheric and Oceanic Sciences and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, U.S.A

2Geosciences Department and Laboratoire de Météorologie Dynamique (CNRS and IPSL), Ecole Normale Supérieure, F-75231 Paris Cedex 05, FRANCE.

3International Research Institute for Climate and Society, Columbia University, Palisades, NY 10964-8000, U.S.A.
This paper presents a predictability study of the Madden-Julian Oscillation (MJO) that relies on combining empirical model reduction (EMR) with the “past-noise forecasting” (PNF) method. EMR is a data-driven methodology for constructing stochastic low-dimensional models that account for nonlinearity, seasonality, and serial correlation in the estimated noise, while PNF constructs an ensemble of forecasts that accounts for interactions between (i) high-frequency variability (“noise”), estimated here by EMR; and (ii) the low-frequency mode (LFM) of MJO, as captured by singular-spectrum analysis (SSA). A key result is that — compared to an EMR ensemble driven by generic white noise — PNF is able to considerably improve prediction of MJO phase. When forecasts are initiated from weak MJO conditions, the useful skill is of up to 30 days. PNF also significantly improves MJO prediction skill for forecasts that start over the Indian Ocean.
1. Introduction

The Madden–Julian oscillation (MJO) is the dominant mode of intraseasonal variability across the global tropics. The MJO is a natural component of the coupled atmosphere–ocean system and it affects other important climate processes, in both the tropics and extratropics; these processes include the seasonal evolution of temperature and precipitation, tropical cyclone frequency, and weather extremes.

In this paper, we built on the extensive studies to improve prediction of the daily indices of the Real-time Multivariate MJO Index, known as RMM1 and RMM2 [Wheeler and Hendon, 2004]. These indices are dominated by intraseasonal fluctuations due to MJO variability in the 40–50-day band. The predictive statistical models used so far can be grouped roughly into (i) multiple lagged-regression–based models [Jiang et al., 2008; Seo et al., 2009; Maharaj and Wheeler, 2005; Kang and Kim, 2010], and (ii) models based on advanced time series analysis, such as singular-spectrum analysis (SSA) and wavelets [Kang and Kim, 2010], neural networks [Love and Matthews, 2009], and analogues [Seo et al., 2009]. The useful predictive skill for the MJO of such empirical models is typically of about 15–20 days, and it is similar to the MJO forecast skill achieved by state-of-the-art dynamical models [Vitart and Molteni, 2010; Zhang and Van den Dool, 2012; Zhang, 2013].

We propose here to bring to bear on empirical MJO prediction a combination of novel tools (i) for the construction of stochastic low-dimensional models (LDMs); and (ii) for the application of the LDMs so constructed to prediction. The LDM construction is carried out by data-driven empirical model reduction (EMR), which accounts for nonlinearity,
seasonality, and serial correlation in the noise estimate [Kravtsov et al., 2005, 2009]. The
prediction method is the pathwise past-noise forecasting (PNF) method [Chekroun et al.,
2011a], which improves predictions by accounting for the modulation of high-frequency
variability, or “noise,” by the MJO’s low-frequency mode (LFM), as captured by SSA.

The understanding and reliable description of LFM s is of the essence for the prediction of
the high-amplitude but irregularly occurring events in the frequency bands of interest [Ghil
and Robertson, 2002]. Furthermore, it is crucial to understand the interaction between
LFMs and the higher-frequency variability [Chekroun et al., 2011a].

This interaction can be better understood in the framework of random dynamical sys-
tems (RDS) theory, in which the stochastic dynamics is studied pathwise — i.e., for
individual realizations of the noise — rather than being merely sampled by an ensemble
of forward trajectories that aim to approximate the system’s probability density function
(PDF) [Chekroun et al., 2011b]. This pathwise view of the dynamics helps in applying
inverse stochastic nonlinear models to prediction, since such an inverse model “lives”
naturally on the estimated path of the noise, as derived from the data.

So far, though, it was not clear how to derive such pathwise models. Most of the
inherent difficulties have been overcome by using the EMR methodology [Kravtsov et al.,
2005, 2009] to yield nonlinear stochastic inverse models that successfully capture the
LFMs of the climatic phenomena of interest [Kondrashov et al., 2011, 2005], while also
estimating the path of the noise.

Chekroun et al. [2011a] extended the results of Kondrashov et al. [2005] beyond the
sampled-PDF approach, and developed PNF as a general pathwise method for nonlin-
ear stochastic systems that exhibit LFMs. This method exploits information on the
estimated path on which the inverse stochastic model evolves, as well as on the interac-
tion between the model’s deterministic, nonlinear components and the stochastic ones.
PNF has been shown to outperform classical ensemble prediction techniques on synthetic,
model-generated data, as well as on observational data for ENSO [Chekroun et al., 2011a].

2. Methods

2.1. Empirical Model Reduction (EMR)

The EMR reduction approach represents a generalization of linear inverse modeling
(LIM; [Penland and Sardeshmukh, 1995]). As an operational prediction methodology,
EMR constructs a low-order nonlinear system of prognostic equations driven by stochastic
forcing; the method estimates the model’s deterministic part, as well as the properties of
the driving noise from the observations or from a more detailed model’s simulation.

EMR relies on multi-level polynomial inverse modeling: it thus allows (a) the model to
be nonlinear; and (b) the noise terms to be correlated in space, and serially correlated
in time [Kravtsov et al., 2005]. Like other model-reduction methodologies, it adopts a
stochastic approach to describe a system of coupled, slow-and-fast variables. Kravtsov
et al. [2009] reviewed EMR and its applications, and compared it with other method-
ologies. They showed EMR to be particularly effective (1) when variability of the faster
parts of the system is both modulated by and feeds back on the slower components of the
system; and (2) when there is no significant time scale separation between slow and fast
variables.
A multi-level EMR model can be formally written as the following set of $M + 1$ vector equations:

\[
\dot{x} = F - Ax + B(x, x) + r_t^0, \\
\dot{r}_{t}^{m-1} = L^{(m)}(x, r_t^0, ..., r_t^{m-1}) + r_t^m, \quad 1 \leq m \leq M,
\]

(1)

Here $x = (x_i; i = 1, ..., d)$ represents the resolved modes on the main level of the model, while $r_t^m$ at additional levels accounts for unresolved modes, which at the last level are approximated as spatially correlated white noise $r_t^M \approx \Sigma \hat{W}$.

For our MJO application, $d = 2$ and $x_1$ and $x_2$ are the RMM1 and RMM2 indices. The terms $-Ax$ and $B(x, x)$ represent the linear dissipation and the quadratic self-interactions, respectively, while $F$ accounts for the deterministic forcing [Kondrashov et al., 2011].

Memory effects of the interactions between the resolved variables $x$ and the unresolved $r_t^m$ arise as convolution terms by integrating recursively the “matrioshka” of levels from the lowest level $M$ to the main level [Kondrashov et al., 2013; Wouters and Lucarini, 2013]. The coupling between the variable $r_t^m$ and the variables $(x, r_t^0, ..., r_t^{m-1})$ from the previous levels is modeled by rectangular matrices $L^{(m)}$ of increasing size.

When deriving a time-discrete formulation of Eq. (1), instantaneous tendencies $\delta x/\delta t$ and $\delta r_t^m/\delta t$ are computed numerically by Euler time-differencing, and are used to estimate by recursive least-squares the coefficients $F$, $A$, $B$ and $L^{(m)}$, as well as computing the regression residuals $r_t^{m+1}$, to be modeled at the next level. The procedure is stopped when the estimated $\xi_t \equiv r_t^M$ has an autocorrelation that vanishes at unit lag and its spatial covariance matrix has converged to a constant matrix $\Sigma$. 

---

DRAFT September 24, 2013, 1:23pm DRAFT
Standard forecasting by EMR relies on forward integration of the model from a given initial state and driven at the last model level by a large ensemble of random realizations of the estimated noise $\Sigma \dot{W}$. ENSO real-time prediction by EMR [Kondrashov et al., 2005] is highly competitive among other dynamical and statistical forecasts in the ENSO multi-model prediction plume of the IRI [Barnston et al., 2012].

For this study, we tested an energy-conserving EMR formulation to ensure that $\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2} < \infty$, see auxiliary materials. We have found that, in the case of MJO, there was no noticeable difference in prediction skill between strictly energy-conserving EMR and its non-energy conserving implementation.

2.2. Past-Noise Forecasting (PNF) by EMR

We start by building a nonlinear EMR model using past observations and we estimate the path of the “weather” noise $\xi_t$ (see Sec. 2.1) that drove this model over previous finite-time windows. The PNF method is then applied in two steps. First, noise snippets — obtained as copies of this estimated path — are selected from past time intervals during which the LFM phase resembles that observed prior to a particular forecast. Second, these noise snippets — which have the same length as the forecast lead time — are used to drive the system into the future.

We rely on SSA to identify a nearly periodic LFM in the (RMM1, RMM2) data set. SSA is a data-adaptive, nonparametric method for spectral estimation that extends classic PCA into the time-lag domain [Vautard and Ghil, 1989]. In practice, the LFMs are described by the reconstructed components (RCs) given by SSA [Ghil et al., 2002]. These RCs are then used for the noise snippet selection here, by finding in the past record analogs of
the LFM that best match the LFM phase just preceding the start of the forecast. Such
a conditioning of the noise forcing on the LFM phase improves forecast skill, and so-
called linear pathwise response with respect to noise perturbations helps explain the PNF
method’s success, cf. Chekroun et al. [2011a].

3. Results and Discussion

We developed a quadratic EMR model with three levels (main plus two, i.e. \( M = 2 \)
in Section 2.1) to simulate and predict the (RMM1, RMM2) daily indices (RMM1-2
hereafter) for June 1974–January 2009\(^1\). Following Kondrashov et al. [2005], we included
the seasonal-cycle effects into the forcing and the linear part, i.e. \( \mathbf{F} \) and \( \mathbf{A} \) in Eq. (1), on
the model’s main level.

Figures 1a,b compare the autocorrelation functions (ACFs) of the RMM1-2 time series,
EMR-simulated vs. observed, when using a large ensemble of EMR stochastic realizations.
These two panels demonstrate that the EMR model captures very well the ACFs, and
hence the MJO power spectrum, as well as its 40–60 day LFM. We found that the use of
\( M > 1 \) improves the EMR-MJO model’s prediction skill at long lead times (not shown).

The EMR model was trained on the June 1974–July 2005 portion of the RMM1-2
record, and forecasts were validated for the independent time interval July 2005–January
2009. In this study, we modified the noise snippet selection in the PNF procedure of
Chekroun et al. [2011a] by adopting a suggestion of Feliks et al. [2010], namely to determine
an instantaneous phase \( 0 < \phi < 2\pi \) of narrow-band LFM components via the Hilbert
transform applied to the SSA reconstruction of the latter, i.e. to the leading RC pair of

\( \text{See } \text{http://cawcr.gov.au/staff/mwheeler/maproom/RMM/ } \text{for the data set.} \)
RMM1. This selection procedure results in a fixed size of 20 PNF ensemble members per
$t^\ast$, where $t^\ast$ is the forecast start time within the validation interval, compared to EMR
ensemble driven by 200 independent realizations of white noise; see the auxiliary materials
for details.

In real-time prediction, the PNF algorithm for choosing noise snippets has to deal with
“end effects” that bias the SSA reconstruction and the estimated value of the LFM phase
at the forecast starting time. For the predictability study presented here, we computed
the RC pair that captures the MJO’s LFM by using the entire data record available, in
order to present the optimal PNF skill as proof-of-concept. It is left for future research
to explore ways to minimize such end effects for applications to operational forecasting.

For the purpose of defining MJO forecast skill, we use the commonly adopted bivari-
ate correlation ($corr$) and the root-mean-square error ($rmse$) between the observed and
forecast RMM indices [Gottschalck et al., 2010]. The skill of the PNF method is notably
better than the standard EMR: see the blue vs. red curves in Fig. 2a. Moreover, PNF
compares favorably with existing statistical and dynamical models for MJO prediction,
cf. Sec. 1.

Even though PNF is applied here to RMM1 only, prediction of RMM2 is also improved,
since both indices are physically coupled. The potential increase in predictability by PNF
(blue) vs. EMR (red) in the time domain coincides mostly with energetic phases of MJO’s
LFM; these phases are well captured by the leading RC pair with an SSA window of 50
days (cyan), cf. Fig. 1d. This PNF feature is due to the method’s intrinsic conditioning
of the driving noise on the LFM phase. On the other hand, EMR and PNF predictions are similar when LFM is weak, i.e. during March-April 2006, as seen in Fig. 1d.

The time series of EMR forecasts are expected to have smaller variance than the PNF ones since — at long lead times and when averaged over a large ensemble of driving-noise realizations — they converge to the climatological mean. PNF, though, picks out certain members of that ensemble that enhance LFM prediction. Moreover, averaging over the PNF ensemble aims to follow the relatively smooth LFM trajectory and, therefore, PNF improvement for $\text{rmse}$ (amplitude) is less pronounced than in $\text{corr}$ (phase); see auxiliary materials.

Figures 2c–f compare prediction skill for the forecasts started for weak vs. strong MJO events, defined as $\{(RMM1)^2 + (RMM2)^2\}^{1/2} < 1$ or $\geq 1$, respectively. The $\text{rmse}$ barely depends on the MJO strength for either EMR or PNF predictions (Fig. 2d,f). Correlation skill $\text{corr}$ when using EMR alone, however, is much worse for weak vs. strong MJO variability (Fig. 2c), while PNF improves forecasts started during strong as well as weak MJO events. It does so much more dramatically in terms of $\text{corr}$ in the latter case, when it exceeds the overall mean skill (Fig. 2e). PNF is thus able to better capture the growth of strong MJO events, an observation that is also consistent with improvement in the time domain (Figs. 1c,d).

Figures 3a–d present prediction skill conditioned on the eight MJO phases (as defined by $\tan^{-1}(\text{RMM2}/\text{RMM1})$ in Wheeler and Hendon [2004], cf. Fig. 1c) at the start of the forecast. Physically, these phases describe different geographical locations of MJO convective activity.
MJO events started in phases 3 and 8, i.e. over the Indian Ocean and tropical Africa, are best predicted; they result in forecasts with the largest corr and smallest rmse, respectively. There is also general improvement in PNF for rmse around phases 2—5. The decrease in corr for PNF predictions started during phases 4–5 corresponds to the well-known difficulties of simulating [Vitart and Molteni, 2010] and predicting [Zhang and Van den Dool, 2012] MJO when it crosses the Maritime Continent. These difficulties are sometimes referred to as “the Maritime Continent prediction barrier,” reminiscent of the “spring barrier” for ENSO prediction [Barnston et al., 2012].

Figures 3e–h show strong seasonality in the corr and the rmse. The correlation gets higher in boreal summer and early fall, as well as during boreal winter, while the rmse is smallest during late summer and early fall. This result agrees with the perception that MJO is more active and better organized during certain seasons [Ghil and Mo, 1991; Wheeler and Hendon, 2004; Zhang and Van den Dool, 2012]. Our results here show that the EMR and PNF prediction methods can capture well the mostly eastward propagating MJO in boreal winter, as well as the additional, northward propagating component in boreal summer.

4. Conclusions

We presented results of a predictability study in which the EMR and PNF methodologies were applied to MJO, with the ultimate goal to use and test these methods in real-time forecasting. The bivariate forecast skill obtained by PNF, in both phase and amplitude, is useful up to 30 days and comparable with the skill demonstrated by a dynamical multi-model ensemble [Zhang, 2013]. The PNF improvement over EMR alone is most
pronounced for predictions started during weak MJO events, over Africa and the Indian Ocean, and during boreal summer and early fall.

**Acknowledgments.** It is a pleasure to thank Adam Sobel and Daehyun Kim for their valuable input, and an anonymous reviewer for thoughtful and stimulating comments.

This study was supported by grants NSF DMS-0934426 and NSF OCE-1243175 from the National Science Foundation, ONR-MURI N00014-12-1-0911 from the the Office of Naval Research, and DOE-DE-F0A-0000411 from the Department of Energy.

**References**


Figure 1. Comparison between the simulation-and-prediction capabilities of the EMR model with the PNF-modified version thereof. (a,b) Autocorrelation functions (ACFs) of the (a) RMM1 and (b) RMM2 daily indices: observations—red; ensemble mean of the EMR simulations—blue, standard deviation of the EMR ensemble—black. (c) Improved prediction of the observed RMM1-RMM2 indices (light black) by PNF-based driving-noise selection: PNF ensemble mean (blue) vs. EMR ensemble mean (red), while individual members of the PNF ensemble are shown in green; the predictions in this panel, out to 30 days, all start on August 26, 2007, namely the date that is marked by the black dashed line in panel (d). (d) PNF prediction in the time domain of RMM1 (light black) at a 25-day lead is consistent with MJO’s LFM, captured by SSA RCs 1-2 (heavy cyan); x-axis is time in calendar months within the validation interval of July 2005–January 2009.
Figure 2. Skill scores for prediction by EMR alone and by EMR driven by PNF snippets. (a) Bivariate correlation (corr) and (b) root-mean-square error (rmse) between the observed and forecast RMM indices showing increase in predictability of RMM1-2 beyond 15 days by PNF (blue) vs. EMR (red); the black curve shows damped persistence as a basis for comparison. (c–f) Prediction skill as a function of the MJO strength at the start of the forecast: PNF improvement is most pronounced for weak MJO, especially in correlation.
Figure 3. Prediction skill conditioned (a–d) on the phase of the MJO; and (e–h) on the calendar month at the start of the forecast. Left column (a, c, e, g) EMR only; right column (b, d, f, h) EMR driven by PNF snippets.
The purpose of these Auxiliary Materials is to provide details on the discrete-time formulation of our empirical model reduction (EMR) model for the Madden-Julian oscillation (MJO) and to help explain the success of past-noise forecasting (PNF) in improving the EMR predictions.

Discrete EMR model formulation

To derive a discrete-time formulation of any EMR model, as stated in Eq. (1) of the Main Text, we start from the main level and form time series of instantaneous tendencies \( \delta x_k / \delta t = (x_{k+1} - x_k) / \delta t \) from a given discrete time series \( x_{tk} = x_k \). Then, given a parametric form of the main level in Eq. (1) of the Main Text — in which a quadratic nonlinearity is typically assumed — the coefficients \( F, A, B \) are found by a least-squares estimation procedure.

The regression residual of the main level defines a time series of \( r_k^{(0)} \), and the least-squares estimation procedure is then repeated at the next level \( m = 0 \) to estimate \( \delta r_k^{(0)} / \delta t \); this yields \( L^{(1)} \) and a new time series of regression residuals \( r_k^{(1)} \). We continue adding levels until the estimated regression residual at the last level \( \xi_t \equiv r_k^{(M)} \) has an autocorrelation that vanishes at unit lag, and its spatial covariance matrix has converged to a constant matrix \( \Sigma = \langle \xi_i^T \xi_i \rangle \). Unless it is necessary for the application at hand to do otherwise, it is convenient to use a unit time step of \( \delta t = 1 \). For instance this is what we used for the MJO data set of this study, given here by the pair of daily indices of the Real-time Multivariate MJO Index (RMM1, RMM2).

To ensure that the deterministic part of the dynamics is stable, we require that \( B_{ijk} x_i x_j x_k \equiv 0 \) and \( A_{ij} x_i x_j > 0 \) should hold for any \( x \neq 0 \) [Kondrashov et al. (2013)]. The condition for the linear-part coefficients \( A \) is a combination of linear equality and inequality constraints. In particular, skew-symmetry for the off-diagonal terms \( A_{ij} + A_{ji} = 0 \) reflects the wave propagation nature of MJO in the phase space of the daily (RMM1, RMM2) indices, while the diagonal terms \( A_{ii} > 0 \) account for dissipation.
The condition for the quadratic-term coefficients $B$ is satisfied by introducing an appropriate number of linear constraints on the $B_{ijk}$. In the MJO case, with $d = 2$, we have $B_{ijk} \equiv 0$, and skew-symmetry constraints for two pairs of coefficients $B_{jkk} + B_{kjj} = 0$, and $B_{jkk} + B_{kjk} = 0$, $j \neq k$. The least-square minimization problem with linear constraints is solved using a preconditioned conjugate gradient method; see [Gould et al., 2001].

By applying this EMR energy-conserving formalism to daily time series of the RMM indices $(x_{1,n}, x_{2,n})$ for July 1974–June 2005, and including seasonal cycle dependence in $F$ and $A$, we obtain the following EMR model:

$$\begin{bmatrix} x_{1,n+1} - x_{1,n} \\ x_{2,n+1} - x_{2,n} \end{bmatrix} = F - A \begin{bmatrix} x_{1,n} \\ x_{2,n} \end{bmatrix} + B \begin{bmatrix} x_{1,n}^2 \\ x_{1,n}x_{2,n} \\ x_{2,n}^2 \end{bmatrix} + S \begin{bmatrix} \sin(\omega n) \\ x_{1,n}\sin(\omega n) \\ x_{2,n}\sin(\omega n) \\ \cos(\omega n) \\ x_{1,n}\cos(\omega n) \\ x_{2,n}\cos(\omega n) \end{bmatrix} + \begin{bmatrix} r_{1,n}^{(0)} \\ r_{2,n}^{(0)} \end{bmatrix},$$

(1)

$$\begin{bmatrix} r_{1,n+1}^{(0)} - r_{1,n}^{(0)} \\ r_{2,n+1}^{(0)} - r_{2,n}^{(0)} \end{bmatrix} = L^{(1)} \begin{bmatrix} x_{1,n} \\ x_{2,n} \\ r_{1,n}^{(0)} \\ r_{2,n}^{(0)} \end{bmatrix} + \begin{bmatrix} r_{1,n}^{(1)} \\ r_{2,n}^{(1)} \end{bmatrix},$$

$$\begin{bmatrix} r_{1,n+1}^{(1)} - r_{1,n}^{(1)} \\ r_{2,n+1}^{(1)} - r_{2,n}^{(1)} \end{bmatrix} = L^{(2)} \begin{bmatrix} x_{1,n} \\ x_{2,n} \\ r_{1,n}^{(0)} \\ r_{2,n}^{(0)} \\ r_{1,n}^{(1)} \\ r_{2,n}^{(1)} \end{bmatrix} + Q \begin{bmatrix} \eta_{1,n} \\ \eta_{2,n} \end{bmatrix}.$$

Here

$$F = \begin{bmatrix} -0.00003 \\ -0.00728 \end{bmatrix},$$

(2)

$$A = \begin{bmatrix} 0.0268 & 0.1150 \\ -0.1150 & 0.0278 \end{bmatrix},$$

(3)

$$B = \begin{bmatrix} 0 & -0.0064 & -0.0002 \\ 0.0064 & 0.0002 & 0 \end{bmatrix},$$

(4)

$$L^{(1)} = \begin{bmatrix} -0.0219 & -0.0025 & -0.4893 & 0.0987 \\ -0.0002 & -0.0176 & -0.0736 & -0.5323 \end{bmatrix},$$

(5)

$$L^{(2)} = \begin{bmatrix} 0.0022 & 0.0020 & -0.0474 & -0.0576 & -0.9560 & 0.0306 \\ -0.0017 & 0.0017 & 0.0463 & -0.0423 & -0.0292 & -0.9687 \end{bmatrix}. $$

(6)

The matrix $S$ in the main level of Eq. (1) here represents the linear interactions on the main level between the pair of RMM indices and the seasonal cycle with frequency
\( \omega = \frac{2\pi}{365} \), and \( n \) accounting for a time index in days. This matrix is given by:

\[
S = \begin{bmatrix}
-0.0028 & 0.0228 & 0.0101 & -0.0015 & -0.0011 & -0.0069 \\
0.0033 & -0.0242 & -0.0250 & 0.0063 & 0.0013 & -0.0048
\end{bmatrix}
\]  

(7)

The matrix \( Q \) in the equation governing the evolution of \( r_t^{(1)} \) accounts for the Cholesky decomposition \( Q^T Q \) of the spatial correlation matrix associated with the last-level regression residual \( \xi_t \equiv r_t^{(2)} \):

\[
Q = \begin{bmatrix}
1.0 & 0.0 \\
0.0412 & 0.9992
\end{bmatrix}
\]  

(8)

This matrix allows us to scale properly the magnitude of the stochastic forcing by two independent normally distributed random variables \( \eta_1 \) and \( \eta_2 \), with \( \eta_1 \sim \mathcal{N}(0, 0.1704) \) and \( \eta_2 \sim \mathcal{N}(0, 0.1688) \).

Note that the multi-level EMR model aims to reproduce the full power spectrum of the RMM indices, interpreted here as the resolved variables \( x \) in Eq. (1) of the Main Text, i.e. as the leading MJO modes. Alternatively, we can think of them as selected observables for these modes. The EMR model construction proceeds by estimating the dynamics of the residual \( r \)-variables, at the additional levels; the latter are aimed to model the dynamics of the unresolved or unobserved modes. The effective contribution of the \( r \)-variables to the dynamics of the RMM indices on the main level may be interpreted as red-noise type forcing with slow-decay time scales, and memory effects related to the past history of the RMM indices also present; [Kondrashov et al.(2013)].

**Practical insights into PNF prediction**

**Proper initialization for prediction.** When performing prediction by using time-discrete formulation of Eq. (1), unobserved \( r \)-variables of additional levels have to be properly initialized in cross-validation interval where only observed RMM indices of main level are available. It is easily achieved by explicitly using RHS of Eq. (1), i.e. time series of \( r \)-variables are estimated in a **backward** procedure by going sequentially from the main level to the last, in a same way as it has been described in the previous section for estimating regression residuals when fitting the model in training interval.

Moreover, forward integration for prediction purposes has to be initialized a certain number of time steps back into the past, equal to number of additional levels in Eq. (1). The last level variable \( r_t^{(1)} \) can be forced either by spatially correlated random variables \( \eta_t \) (standard EMR prediction) or by chosen PNF noise snippet from the regression residual \( \xi_t \) of the last level in the training interval. Equation (1) is then integrated **forward** from the last level to the main one over an interval of length equals to the lead-time, the latter corresponding also to the length of the snippet.

Such “backward-forward” procedure ensures initial conditions for the forecast that are consistent with observations in validation interval, and the random forcing introduced at the last level is then properly accounted at the main level beyond start of the forecast.
Snippet selection via Hilbert transform. To select PNF noise snippets associated with a particular \( t^* \) (the forecast start time in validation interval), we first perform a Hilbert transform \( H(s_t) = s_t + i h_t \), where \( s_t \) denotes the time series of MJO LFM as obtained by applying SSA to either RMM1 or RMM2 index. The instantaneous phase \( \phi_t \) is uniquely defined as \( \phi_t = \arctan(h_t/s_t) \) and can be thus determined for all data points up to the forecast start time \( t^* \). The noise snippets — whose length equals the forecast lead — are then selected from the history of \( \xi_t \) (as determined by EMR during the training interval) for LFM phase values \( \phi_t \) that fall within a small \( \epsilon_\phi \)-neighborhood of \( \phi^*_t \), the value at the forecast start time \( t^* \).

We used \( \epsilon_\phi \approx 0.1\pi \), resulting in a typical subset of \( \approx 500 \) ensemble members — out of a maximum possible number of \( \approx 11000 \), the total number of days in the training interval. In addition, we refined further this subset of noise snippets by choosing only those that correspond to initial states in the raw RMM data closest to those at forecast start \( t^* \). This final ensemble of noise snippets reduces then to a fixed size of 20 members per \( t^* \).

Behavior of individual members from PNF ensemble. The prediction by averaging over the PNF ensemble aims to follow the smooth behavior of the MJO’s LFM, as captured by SSA. Therefore, its phase speed is not expected to exhibit abrupt local changes.

This behavior, which is clearly due to averaging, is not shared by individual ensemble members. To illustrate this point, we included below separate plots for 20 individual members of the PNF ensemble used for the MJO case presented in Fig. 1c of the Main Text. These 20 members are ranked by increasing bivariate root-mean-square error with respect to the true evolution over 30 days. The 20 plots illustrate very rich dynamics for individual trajectories, including both slow-down and acceleration in RMM phase space. This rich dynamics can be attributed to the dynamical interaction between the multi-level variables of the low-order model and the PNF selection of the driving noise at the last level. Note that the leading PNF ensemble members (ranked 1 through 3 below) come very close to actual MJO evolution over 30 days.

References


Figure 1: Individual members (1–4) of the PNF ensemble, observed RMM1-RMM2 indices (light black), PNF ensemble mean (blue); EMR ensemble mean (red)
Figure 2: Individual members (5–8) of the PNF ensemble, observed RMM1-RMM2 indices (light black), PNF ensemble mean (blue); EMR ensemble mean (red)
Figure 3: Individual members (9–12) of the PNF ensemble, observed RMM1-RMM2 indices (light black), PNF ensemble mean (blue); EMR ensemble mean (red)
Figure 4: Individual members (13–16) of the PNF ensemble, observed RMM1-RMM2 indices (light black), PNF ensemble mean (blue); EMR ensemble mean (red)
Figure 5: Individual members (17–20) of the PNF ensemble, observed RMM1-RMM2 indices (light black), PNF ensemble mean (blue); EMR ensemble mean (red)