Motivation

• Much of the observed spatio-temporal climate variability can be represented by a few number of "resolved" (predictable) modes of “large” (space) and “slow” (time) scales

• Low-order models (LOM):
  – linear and nonlinear interactions between the resolved modes
  – estimation of “fast&small” scales - unresolved modes (“noise”)
  – proper description of interactions between the few resolved and the large number of unresolved modes

• Empirical Model Reduction (EMR) - data-driven LOM methodology - no prior assumptions on scale separation between resolved and unresolved modes.

- low-order nonlinear equations driven by stochastic forcing
- estimates dynamical operator and driving noise from the data
Outline

• Under certain conditions, EMR can be understood as a data-driven discrete formulation of nonlinear Multi-Layer Stochastic Model (MSM) in the spirit of Mori-Zwanzig formalism of statistical mechanics that estimates unresolved dynamics as (time-domain) orthogonal, and represents resolved-unresolved interaction as non-Markovian and stochastic. *Data-Driven Model Reduction by a Multilayered Stochastic Approach with Energy-Preserving Nonlinearities*, (2013) D. Kondrashov, M. Chekroun, M. Ghil, Physica D. imminent submission.

• **Orthogonal** property of unresolved dynamics facilitates energy-conserving EMR formulation

• **Past-Noise Forecasting (PNF)**

• **EMR&PNF Matlab Toolbox**

• Climate application of EMR&PNF: Madden-Julian Oscillation
The data series \( x(t) \) is given time series (observations, or high-end, complex model simulation) of \textbf{resolved modes} modeled at main level.

- **Unresolved modes** \( r^{(m)} \) are estimated and modeled by “\textit{matrioshka}” of addt’l levels. Model coefficients \( F, A, B \) and \( L^{(m)} \) are computed by multiple linear regression recursively from main level down to the last, using estimated tendencies \( \Delta x, \Delta r^{(m)} \) as predictants.

- \( r^{(m)} \) are estimated as \textit{regression residuals} at each level and are orthogonal in time domain to the variables from the previous levels \([x, r^{(0)}, ..., r^{(m-1)}]\), i.e

  \[
  \langle x, r^{(m-1)} \rangle = 0; \\
  \langle r^{(j)}, r^{(m)} \rangle = 0, \forall j, m \in \{1, ..., M - 1\}, \text{ with } j \leq m - 1
  \]

- The data series \( x(t) \) is reconstructed exactly when LOM is driven by last level regression residual \( r^{(M)} \)
EMR as Multi-layered Stochastic Model

For a special case of resolved-unresolved interactions when

\[ L^{(m)} [(x)^T, (r^{(0)})^T, \ldots, (r^{(m-1)})^T]^T = C_m x - D_m r^{(m-1)} \]

additional levels can be integrated and expressed formally as

\[ r^{(m-1)}(t) = \int_0^t e^{-(t-s)D_m} C_m x(s) ds + \int_0^t e^{-(t-s)D_m} r^{(m)}(s) ds \]

\[ \kappa_m(t) = e^{tD_m} C_m; \mu_m(t) = e^{tD_m} \]

\[ r^{(m-1)}(t) = (\kappa_m * x)(t) + (\mu_m * r^{(m)})(t) \]

It allows to formally integrate *matrioshkas* from the last level to the top and obtain

\[
\frac{dx}{dt} = -(A x + B(x, x)) + \sum_{l=0}^{M-1} \sum_{m=0}^{l-1} (\mu_0 * \cdots * \mu_m * \kappa_{l+1} * x)(t) + (\mu_1 * \cdots * \mu_M * r^{(M)})(t) .
\]

**Interpretation within Mori-Zwanzig formulation of statistical mechanics:** The evolution of \( x(t) \) relies on stochastic integro-differential system of equations with the deterministic terms in (a) that rule the nonlinear self-interactions of \( x \); it is the Markovian contribution. The terms in (b) convey memory effects through repeated convolutions involving the past history of \( x \) and constitute thus the non-Markovian contribution; (c) represents last level random forcing orthogonal to \( x \).
Due to orthogonality of unresolved dynamics, on average there is zero net energy flux between resolved and unresolved modes. The following conditions then should hold for EMR to be considered formally as stable forced-dissipative system:

\[
\begin{align*}
\| \mathbf{x} \| &= \sum_{i=1}^{M} x_i^2 < \infty; \\
\mathbf{x} \cdot |\dot{\mathbf{x}}| &= \mathbf{F} - \mathbf{A}\mathbf{x} + \mathbf{B}(\mathbf{x}, \mathbf{x}) + \mathbf{r}_i^0 \\
\int \mathbf{x} \cdot \mathbf{r}_t^0 dt &\approx 0
\end{align*}
\]

The RHS of scalar EMR model (for univariate time series) has to be a gradient of the potential function, and the lowest possible value of nonlinearity K that results in a stable model is equal to K=3"! [Kravtsov, Kondrashov and Ghil, 2005]

The multivariate case is more interesting one to consider: with enough data record and proper regularization even non-constrained EMR models are well behaving [Lorenz 63, Kravtsov, Kondrashov and Ghil, 2005]
Energy-Conserving EMR formulation -II

\[ B_{ijk} x_i x_j x_k \equiv 0; A_{ij} x_i x_j > 0, \mathbf{x} \neq 0 \]

- Following conditions should hold for multivariate quadratic EMR to be considered formally as stable nonlinear forced-dissipative system:

\[ B_{iii} = 0 \]
\[ B_{jkk} + B_{kjj} = 0 \]
\[ B_{jkk} + B_{kjk} = 0; j \neq k \]
\[ B_{ijk} + B_{jik} + B_{kij} = 0, i \neq j \neq k \]
\[ A_{ij} + A_{ji} = 0, i \neq j, \]
\[ A_{ii} > 0. \]

- Linear equality and inequality constraints in MLR estimation of A and B coefficients solved by standard minimization methods.

- Effectively reduces number of independent EMR coefficients estimate
Given the fact that weather (small and fast scales) does affect climate (large and slow scales) it might be useful to try to use this for prediction, and in particular to improve long-term climate prediction.

We propose two circumstances when it can be done:

– substantial low-frequency variability (LFV) in climate modes

– linear response of the LOMs w.r.t. noise perturbation

The crucial idea is to composite ensemble of noise “snippets” according to the current phase of the LFV, and use this ensemble to drive the LOM into the future - hence the acronym Past-Noise Forecasting (PNF).

Singular Spectrum Analysis (SSA) and Hilbert Transform to estimate current phase of LFV.

Predictability study for MJO.
To make a forecast from time $t = t_0$, we determine the LFV phase at that time, and consider a subset $S_{t_0}$ of “snippets” of the residual noise $\xi_{t_0}$ of length $\delta$, which are associated with past occurrences of a similar LFV phase. These snippets are then used to drive the forecast using the EMR ENSO model from time $t_0$ on. The PNF prediction for $\delta = 16$ months ahead is given below.


Red–PC1; cyan–EMR driven by individual snippets (selected by the PNF method); heavy blue–PNF predictor (mean over cyan ensemble); dashed blue–EMR predictor (mean over large ensemble, no selection); and magenta–traditional analog method.

Singular Spectrum Analysis (SSA)

SSA decomposes (geophysical & other) time series into
Temporal EOFs (T-EOFs) and Temporal Principal Components (T-PCs), based on the series’ lag-covariance matrix.

Selected parts of the series can be reconstructed, via Reconstructed Components (RCs).

- SSA is good at isolating oscillatory behavior via paired eigenelements.
- SSA tends to lump signals that are longer-term than the window into one or two trend components.

Selected References:
Vautard & Ghil (1989, Physica D);
Ghil et al. (2002, Rev. Geophys.)
**EMR & PNF Matlab Toolbox**

- General fitting, stochastic simulation and prediction EMR routines
- Multivariate data as input
- Flexible options: regularization, order of nonlinearity, estimating optimal number of model levels from the output diagnostics, optional external periodic forcing (seasonal cycle)
- Fitting and simulation examples: Lorenz 63’ model with additive and multiplicative noise
- PNF and EMR prediction demo in a partially observed case (*Chekroun, Kondrashov, Ghil PNAS 2011*)
- Available at project’s website at Columbia to team members
EMR & PNF Toolbox - I

Written by Dmitri Kondrashov, UCLA, Oct 24, 2012

Email: dkondras@atmos.ucla.edu

Exclusive use for ONR-N00014-12-1-0911 project

Revision History

February 22, 2013: normalization option added to the input arguments
May 28, 2013: pls added as input argument for partial least-squares (PLS) on main level;
May 28, 2013: model integration is renormalized, if necessary
June 2, 2013: additive or multiplicative periodic forcing is added for main level

[xx,modstr,xt_res,vars]=fitemrplaxext(length,per,data,nelin,nlevel,niter,innor,pls)

CONSTRUCT MULTILEVEL-NONLINEAR EMR MODEL USING PARTIAL-LEAST SQUARES, AND PERFORM STOCHASTIC SIMULATION BY EMR

INPUT:
- data - multivariate array [N,L], N - length of the time series, L - number of channels,
  each channel is centered prior to EMR fitting;
- nelin - order of EMR model: nelin=1 - is linear; nelin=2 - is quadratic, nelin<=4;
- nlevel - total number of levels in the EMR model including main level;
- niter - total number of simulations (stochastic realizations) by EMR model;
- tlength - length of the simulated time series by EMR
  per - sets optional external periodic forcing, array of periodicities.
  inorm - controls data normalization:
    inorm = 0 - default, normalizes all channels by std(data(:,1));
    inorm = 1 - normalizes all channels by its own std.dev
    inorm = 2 - fit non-normalized data
- pls - controls partial least squares (PLS) for main level
  pls = 0 - no PLS
  pls = -1 - perform PLS

INTERNAL:
- ires - controls stochastic forcing at last level of EMR,
  ires = 0 - spatially correlated white noise
  ires = 1 - residual noise (xt_res)
  ires = 2 - no forcing

  nout - number of leading channels to store in output of EMR-simulation
  lim - threshold abs value, if exceeded in EMR simulation, it is restarted
  ietmax - maximum number of EMR simulations to try

  iext - specifies interaction with external periodic forcing
  iext = 0 - additive (a*sin(w.t) + b*cos(w.t))
  iext = 1 - default, adds multiplicative interactions (a*x*sin(w.t) + b*x*cos(w.t))

OUTPUT:
- xx - multivariate array [tlength,nout,niter], successfully simulated data by EMR
- xt_res - array of regression residuals at the last level, size [N,L];
- modstr - structure with model information, see comments at the end of the program
- varr - EMR diagnostics, array [nlevel,L];
  if varr(nlevel,:)=0.5, then optimum number of levels is nlevel-1!
Empirical Model Reduction Example 1: Orthogonal Dynamics

Dmitri Kondrashov, UCLA
for exclusive use by ONR-N00014-12-1-0911 project

June 4, 2013

We have the following 2-variable \((r, y)\) linear stochastic model:

\[
\begin{pmatrix}
\dot{x} \\
\dot{y}
\end{pmatrix} = \begin{pmatrix}
a & 1 \\
q & A
\end{pmatrix}\begin{pmatrix}
x \\
y
\end{pmatrix} + \begin{pmatrix}
0 \\
\sigma W_t
\end{pmatrix}
\] (1)

For model coefficients \(a = -2, q = 1, A = -1\), the eigenvalues of the linear part of the model Eq. (1) are \(\lambda_1 = -0.382\) and \(\lambda_2 = -2.618\). By using Euler scheme for time integration, the time-discrete formulation of the model becomes:

\[
\begin{pmatrix}
x_{i+1} - x_i \\
y_{i+1} - y_i
\end{pmatrix} = \begin{pmatrix}
a & 1 \\
q & A
\end{pmatrix}\begin{pmatrix}
x_i \\
y_i
\end{pmatrix} dt + \begin{pmatrix}
0 \\
\sigma dW_t
\end{pmatrix}
\] (2)

We are going to apply EMR to the time series \((x_i, \ldots, N)\) time series alone. For the main level of the linear EMR model we should regress time increment \(dx_i = x_{i+1} - x_i\) on \(x_i\) to find \(\hat{a} = \hat{a} x dt\). Using back-of-envelope regression formulas and in the limit \(N \to \infty, dt \to 0\), one readily obtains \(\hat{a} \to 0\). The regression residual \(r_idt = dx_i - \hat{a}x_i dt\) \(\to dx_i\) and by construction is orthogonal to \(dx_i\).

By Eqs. (1) and (2), the similarity transformation \(S\) of variables \((x, y) \to (r, r)\) is thus obtained:

\[
\begin{pmatrix}
x \\
y
\end{pmatrix} = S\begin{pmatrix}
x \\
r
\end{pmatrix}; S = \begin{pmatrix}
1 & 0 \\
-a & 1
\end{pmatrix}; S^{-1} = \begin{pmatrix}
1 & 0 \\
a & 1
\end{pmatrix}
\] (3)

and the EMR-recovered full model is then

\[
\begin{pmatrix}
\dot{x} \\
\dot{r}
\end{pmatrix} = S^{-1}\begin{pmatrix}
a & 1 \\
q & A
\end{pmatrix} S\begin{pmatrix}
x \\
r
\end{pmatrix} + \begin{pmatrix}
0 \\
\sigma W_t
\end{pmatrix},
\] (4)

or equivalently

\[
\begin{pmatrix}
x_{i+1} - x_i \\
r_{i+1} - r_i
\end{pmatrix} = \begin{pmatrix}
0 \\
q - Aa & 1
\end{pmatrix}\begin{pmatrix}
x_i \\
r_i
\end{pmatrix} dt + \begin{pmatrix}
0 \\
\sigma dW_t
\end{pmatrix}
\] (5)

Linear part of Eq. (5) has the same characteristic polynomial as Eq. (2), and so same eigenvalues: \(\lambda^2 - (a + A)q - q + Aa = 0\). The random forcing in Eq. (5) is of the same amplitude as in Eq. (2), so the underlying statistics of simulated variables is identical, but subject to the similarity transformation in Eq. (3). The only difference between two models is that EMR-estimated hidden \(r\) variable is orthogonal to \(x\), while the original \(y\) variable is not. This numerical example also illustrates EMR application to a partially observed system - by using only variable \(x\) obtained from the full two-variable original \((x, y)\) model of Eq. (2).

The following scripts use EMR to obtain Eq. (5) by explicitly accounting for value of \(dt\) in Eq. (2):

\text{lintest.dt.m} - main script
\text{findemr.dt.m} - fitting 2-level (main level + one addl.) linear EMR model,
\text{intemr.dt.m} - stochastic integration of EMR model

One can also proceed with EMR by simply assuming \(dt = 1\) and recover statistics of the variable \(x\):

\text{lintest.m} - main script
\text{findemr.m} - fitting optimal 2-level linear EMR model;
\text{intemr.m} - stochastic integration of EMR model

One can also use general purpose EMR routine (\text{fitemrplbsext.m}):

\text{Step 1} - use EMR diagnostics to find that optimum number of levels is 2: \text{nivel} = 3 - consider 3-level model;
\text{nelin} = 1 - consider linear model;
\text{iform} = 2 - fit non-normalized data;
\text{iter}=0 - don’t integrate the model (just do the fit);
\text{ipls} = 0 - no regularization;
\{xxx\} = \text{fitemrplbsext([],[],data(:,1),nelin,nivel,iter,inorm,ipls)}:

EMR Diagnostics per level and component of the state vector:
Level 1:
Component i=1: 2.1308e-07;
Empirical Model Reduction Example 2: Stochastically Forced Lorenz ’63 Model by Additive or Multiplicative Noise

Dmitri Kondrashov, UCLA
for exclusive use by ONR-N00014-12-1-0911 project
June 6, 2013

We use the Lorenz’63 model forced by multiplicative noise:

\[ dx = s(y - x)dt + \sigma x dW_{1t} \]
\[ dy = (rx - y - zx)dt + \sigma y dW_{2t} \]
\[ dz = (-bz + xy)dt + \sigma z dW_{3t} , \]

or additive noise:

\[ dx = s(y - x)dt + \sigma dW_{1t} \]
\[ dy = (rx - y - zx)dt + \sigma dW_{2t} \]
\[ dz = (-bz + xy)dt + \sigma dW_{3t} , \]

where \( dW_{1t}, dW_{2t}, \text{and } dW_{3t} \) can be independent or identical Wiener processes, \( \sigma \) is magnitude of noise, and \( s, r, b \) are classical Lorenz’63 model parameters.

The multi-level EMR model has the following form:

\[ dx = F dt - A x dt + B(x, x) dt + r^0 dt \]
\[ dr^l = M_l(x, r^0, ..., r^{l-1}) dt + r^{l+1} dt, \quad 0 \leq l \leq L - 1, \]
\[ r^l dt \approx \Sigma \tau \]

Instantaneous tendencies \( dx \) and \( dr^l dt \) are estimated by Euler time-differencing from the time series \( x \) and \( r^l \) respectively, and serve as predictands to compute \( F, A, B \) and \( M_l \) by recursive least-square procedure, whereas predictors are \( x \) variables of the main level augmented by the modeled regression residuals \( r^l \) at the additional levels. The addition of extra levels to EMR is
EMR & PNF Toolbox - IV

```
[fcast,true,rms,anc,modstr,xt_res,varr] = fcstemprl;setext(data0,per,neln,nlevel,niter,inorm,pls,lead,lstartt,lstart,lend)

INPUT:
data0 - multivariate array [N,L], N - length of the time series, L - number of channels,
each channel is centered prior to EMR fitting;
neln - order of EMR model: neln=1 is linear; neln=2 is quadratic, neln<=4;
nlevel - total number of levels in the EMR model including main level;
niter - total size of ensemble of random simulations (stochastic realizations) by EMR model to make prediction;
per - sets optional external periodic forcing, array of periodicities.
inorm - controls data normalization:
inorm = 0 - default, normalizes all channels by std(data(:,1));
inorm = 1 - normalizes all channels by its own std.dev
inorm = 2 - fit non-normalized data
pls - controls partial least squares (PLS) for main level
pls = 0 - no PLS
pls = 1 - perform PLS
lead - maximum prediction lead
lstartt - end point of model training interval [l,lstart] [lstart lend] - sets validation interval where at every point forecasts are made;
lead is assumed to be <=N-lead;

INTERNAL:
ires - controls stochastic forcing at last level of EMR,
ires = 0 - spatially correlated white noise
ires = 1 - residual noise (xt_res)
ires = 2 - no forcing
nout - number of leading channels to store in output of EMR-simulation
lim - threshold abs value, if exceeded in EMR simulation, it is restarted
itermax - maximum number of EMR simulations to try
iext - specifies interaction with external periodic forcing
iext = 0 - additive (a*sin(w.t) + b*cos(w.t))
iext = 1 - default, adds multiplicative interactions [a*x*sin(w.t) + b*x*cos(w.t)]

OUTPUT:
xt_res - array of regression residuals at the last level, size [N,L];
modstr - structure with model information, see comments at the end of the program
varr - EMR diagnostics, array [nlevel,L];
if vari(nlevel,:)>=0.5, then optimum number of levels is nlevel-1;
fcast - multivariate array of size [lend-lstart+1,lead,nout] and contains EMR ensemble mean forecast in validation interval;
ttrue - multivariate array of size [lend-lstart+1,lead,nout] with data for verification of EMR forecasts in validation interval;
rms - array of size [lead,nout] and contains normalized RMSE computed from FCST and TRUE arrays
anc - array of size [lead,nout] and contains anomaly correlation computed from FCST and TRUE arrays
```
The Past Noise Forecasting Demo

Dmitri Kondrashov and Mickael Chekroun, UCLA
For exclusive use for the ONR-N00014-12-1-0011 MURI project
June 6, 2013

1 A simple model for testing PNF

Largely based on Supplementary Information (SI) of Chekroun, Kondrashov and Ghil (2011) PNAS paper [1]. In this SI, we have considered the periodically and stochastically forced version (in the Itô sense) of a Holling population dynamics model [3]:

\[
\begin{align*}
\frac{dx_1}{dt} &= \left( (r + c \sigma W_t) x_1 (a + x_1)(1 - x_1) - c r_1 x_2 + a \sin(2 \pi f t) \right) dt \\
\frac{dx_2}{dt} &= \left( -m x_2 + (c - m) x_1 x_2 \right) dt.
\end{align*}
\] (1.1)

The parameters of this model are described hereafter.

Three key features of this model are of special interest for testing the PNF method: (i) it is nonlinear and stochastic, but contrarily to the model used in the Main Text of [1], it does not belong to the class of EMR models; (ii) it is forced by white noise, thus emphasizing that the PNF can work without memory effects in the stochastic forcing; and (iii) it can be subjected to an easily reproducible battery of numerical tests. The skill in the PNF forecasts is mainly due to the model’s low-frequency variability (LFV) manifested by low-frequency oscillatory mode (LFM), and to its pathwise linear response; see [1, Main Text] for more details.

Variables \( x_2 \) and \( x_1 \) here represent predator and prey density, respectively. The stochastic term in the system (1.1) represents a random perturbation of the parameter \( r \) by white noise. The variable \( x_1 \) is also assumed to exhibit a seasonal variation — due, for instance, to migration effects — as modeled by the presence of the deterministic, additive forcing \( a \sin(2 \pi f t) \).

We integrate the system (1.1) from \( t = 0 \) to \( t = T_f = 2000 \) (in dimensionless units) by using a classical stochastic Euler-Maruyama scheme with step size \( \Delta t = 0.1 \), in the context of Itô calculus. The values of the parameters are \( \sigma = 0.3, \ m = r = 1, \ c = 1.5, \ a = 0.3, \ a = 0.05, \) and \( f = 0.25 \), while the initial state is \( (x_2(0), x_1(0)) = (0.5, 0.5) \). For these parameter values, when both periodic forcing and noise are absent, the model has only one globally stable equilibrium.

When turning on the periodic forcing, with amplitude \( a = 0.05 \) and frequency \( f = 0.25 \), the system exhibits only one periodic orbit of period 4, which is globally stable. In the presence of noise \( (\sigma = 0.3) \), a low-frequency mode of period equal to approximately 25 units — i.e., a frequency \( f' = 0.04 \) — becomes dominant. In [1, Fig. S4A] it was noticed that — when the noise is turned off — this mode is rapidly damped and is visible only during the transient that leads up to the unique attracting periodic orbit. The model’s LFV in the case \( \sigma = 0.3 \) can be therefore attributed to a damped non-normal mode that is sustained by the noise.

Figure 2: Skill comparison between ensemble predictions (ENS) and PNF predictions for modified Holling model in Eq. (1.1); the ensemble has \( N = 100 \) members: (a) corr, and (b) rms. The validation interval is \( T_f = 500 \) units long, from 1400 to 1990, with 5900 values of \( t' \) issued in steps of \( \Delta t' = 0.1 \) throughout the interval, to make a 10-unit prediction out to the respective \( t' + T \). The snippets are selected in the interval that extends from \( t = 0 \) to \( t = 1400 \). The variables predicted are the anomalies \( x'_1, \ x'_2 \) of the prey and predator densities, \( x_1 \) and \( x_2 \), respectively.
3 Empirical Model Reduction (EMR) and Prediction Demo

Matlab script `toy_emrfit.m` demonstrates how by applying general purpose Matlab routine `fitemrplsext.m` solely on a time series \( \{x_2 : i = 1, ..., N\} \), a one-variable EMR model \[5\] in \( x_2 \) alone can be derived to replace the original two-variable model of Eq. (1.1). If multiplicative periodic forcing is included into two-level cubic EMR model on its main level, almost perfect match of autocorrelation function of \( x_2 \) is obtained vs. the EMR model without periodic forcing, compare Figures 4 and 5. Note that parametric form of univariate EMR model for \( x_2 \) is very different from its true form in Eq. (1.1) as it doesn’t explicitly account for interactions with periodically and stochastically forced “unresolved” variable \( x_1 \). The latter are parameterized by the multi-level EMR-estimated interactions.

Script `toy_emrpredict.m` demonstrates how to apply prediction and cross-validation EMR routine `f2temrplsext.m`. In addition to EMR fitting parameters such as in `f2temrplsext.m`, input parameters of `f2temrplsext.m` include specific predictive parameters such as maximum prediction lead time, end of the model training interval, the start and the end of the validation interval which may not necessarily fully overlap with the EMR training interval, i.e. out-of-sample. Output of `f2temrplsext.m` includes EMR ensemble mean forecast, validation time series, as well as prediction skill in terms of anomaly correlation and normalized root-mean-squared error. See more about available options by help `f2temrplsext`.

Figure 6 shows that best 2-level cubic EMR model handily beats in prediction linear single level (linear inverse or LIM) model, and that EMR prediction skill is actually almost the same as for the true ideal model Eq. (1.1) (compare it with blue dashed lines in Fig. 2a, b).

- **NB1:** When PLS is necessary for multivariate data input, some experimenting is required to find optimal data normalization `inorm` to obtain best predictive EMR model.
- **NB2:** Part of the output diagnostics of `f2temrplsext.m` is the total number of successful EMR simulations attempted to reach the necessary ensemble size as specified in the input parameter. Typically these numbers should be equal and this is another indication of quality of EMR model. Because predictions exceeding very large (by absolute magnitude) nonphysical values are simply discarded, the optimal EMR model will typically have none.
- **NB3:** For multi-level EMR model, predictions are initialized automatically back in the past so that random effect of stochastic forcing at the last level is reflected correctly in the forecast.

The following scripts are used to obtain figures for this demo:
- `toy_emrfit.m` – main script to run EMR fit example,
- `autocorrmr.m` – compute and compare autocorrelations
- `toy_emrpredict.m` – script to run example of out-of-sample prediction and cross-validation
- `f2temrplsext.m` – general purpose predictive and cross-validation EMR routine
- `model.proj.m` – project the out-of-sample data onto multi-level EMR model to estimate the latent variables of addl levels and to properly initialize forecasts.

![Figure 6: Prediction skill for reduced models obtained solely from \( x_2 \) time series; the best cubic 2-level EMR model handily beats LIM both in terms of anomaly correlation and RMS.](image)

**References**


MJO Basic Facts

- MJO is the dominant intraseasonal (30–90 days) variability in the tropical atmosphere and represents large-scale coupling between atmospheric circulation and tropical deep convection.

- Unlike ENSO which is mostly a standing pattern, MJO is a traveling eastward pattern.

- Real-time Multivariate MJO (RMM1,2) represent EOF-based prefiltering of 850 hPa zonal wind, 200 hPa zonal wind, and outgoing long-wave radiation (OLR) data.
A quadratic, energy-conserving three-level EMR model with seasonal cycle was developed to model and predict the leading pair (RMM1, RMM2) of real-time, daily MJO indices.

Shorter time-scale of MJO LFV (30-50) day compared to ENSO (2-4yr) allows robust testing of EMR/PNF in ~35yr of data.

The EMR model was fitted, and the PNF method was trained, on the first 30 yr of data (1974-2004) and validated on the last 4yr (2005-2008).

Apply the Hilbert Transform to the SSA reconstruction (Feliks et al., 2010) over the whole time series of the RMM1,2 to find instantaneous LFV phase.

Refine subset of noise samples by choosing only those that correspond to initial states in the raw (RMM1, RMM2) data closest to those at the start time of the forecast.
Energy-conserving EMR-MJO Model

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

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

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

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

\[
B = \begin{bmatrix}
  0 & -0.0064 & -0.0002 \\
  0.0064 & 0.0002 & 0
\end{bmatrix}
\]

\[B_{ijk} x_i x_j x_k = 0; A_{ij} x_i x_j > 0, \mathbf{x} \neq 0\]

\[Q = \begin{bmatrix}
  1.0 & 0.0 \\
  0.0412 & 0.9992
\end{bmatrix}; \quad \eta_1 \sim \mathcal{N}(0, 0.1704), \eta_2 \sim \mathcal{N}(0, 0.1688)\]
A quadratic, three-level EMR model with seasonal cycle for daily (RMM1, RMM2)
PNF and EMR for MJO - II

(a) Bivariate correlation (corr) and (b) RMSE 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.

The PNF skill is useful up to 30 days and comparable with the skill demonstrated by a state-of-the-art dynamical multi-model ensemble [Zhang, 2013].

Similar skill for non-energy conserving formulation (no constraints)
Prediction skill conditioned (a–d) on the phase of the MJO; and (e–h) on the calendar month at the start of the forecast.

MJO events started in phases 1-3 and 8, i.e. over tropical Africa and the Indian Ocean (IO), characterized by the strongest MJO signal, are best predicted. The ability of good predictions over IO are quite striking given difficulty of most dynamical models (reason for DYNAMO project).

The drop in skill during phase 4–5 corresponds to the well-known difficulty of predicting MJO when it crosses the Maritime Continent.

Seasonally, skill gets higher in boreal summer and early fall, as well as during boreal winter.