A Conditional Gaussian Framework for Uncertainty Quantification, Data Assimilation and Prediction of Complex Turbulent Dynamical Systems

Nan Chen

joint work with Andrew J. Majda, Dimitris Giannakis and Xin Tong

Center for Atmosphere Ocean Science (CAOS)
Courant Institute of Mathematical Sciences
New York University

Nonlinear & Stochastic Problems in Atmospheric and Oceanic Prediction
Banff International Research Station, November 19 – 24, 2017

Introduction

Turbulent dynamical systems

- ubiquitous in geoscience, engineering, neural and material sciences
- characterized by a large dimensional phase space and a large dimensional space of strong instabilities, which transfer energy throughout the system

Introduction

Turbulent dynamical systems

- ubiquitous in geoscience, engineering, neural and material sciences
- characterized by a large dimensional phase space and a large dimensional space of strong instabilities, which transfer energy throughout the system

Central math/science issues

- accurate descriptions of turbulent phenomena
- state estimation with partial and incomplete information from noisy observations
- effective predictions with improved initializations using filtering/data assimilation
- quantifying uncertainty and model error

Conditional Gaussian Nonlinear Systems

Many turbulent dynamical systems belong to conditional Gaussian framework.

The conditional Gaussian systems have the following abstract form,

$$d\mathbf{u}_\mathbf{I} = [\mathbf{A}_0(t,\mathbf{u}_\mathbf{I}) + \mathbf{A}_1(t,\mathbf{u}_\mathbf{I})\mathbf{u}_\mathbf{II}]dt + \mathbf{\Sigma}_\mathbf{I}(t,\mathbf{u}_\mathbf{I})d\mathbf{W}_\mathbf{I}(t), \tag{1a}$$

$$d\mathbf{u}_\mathbf{II} = [\mathbf{a}_0(t,\mathbf{u}_\mathbf{I}) + \mathbf{a}_1(t,\mathbf{u}_\mathbf{I})\mathbf{u}_\mathbf{II}]dt + \mathbf{\Sigma}_\mathbf{II}(t,\mathbf{u}_\mathbf{I})d\mathbf{W}_\mathbf{II}(t), \tag{1b}$$

Once $\mathbf{u}_\mathbf{I}(s)$ for $s \le t$ is given, $\mathbf{u}_\mathbf{II}(t)$ conditioned on $\mathbf{u}_\mathbf{I}(s)$ becomes a Gaussian process,

$$p\big(\mathbf{u}_\mathbf{II}(t)|\mathbf{u}_\mathbf{I}(s \le t)\big) \sim \mathcal{N}(\bar{\mathbf{u}}_\mathbf{II}(t), \mathbf{R}_\mathbf{II}(t)). \tag{2}$$

- Despite the conditional Gaussianity, the coupled system (1) remains highly nonlinear and is able to capture the non-Gaussian features as in nature.
- The conditional distribution in (2) has closed analytic form (Liptser & Shiryaev 2001).

$$d\bar{\mathbf{u}}_\mathbf{II}(t) = [\mathbf{a}_0(t,\mathbf{u}_\mathbf{I}) + \mathbf{a}_1(t,\mathbf{u}_\mathbf{I})\bar{\mathbf{u}}_\mathbf{II}]dt + (\mathbf{R}_\mathbf{II}\mathbf{A}_1^*(t,\mathbf{u}_\mathbf{I}))(\mathbf{\Sigma}_\mathbf{I}\mathbf{\Sigma}_\mathbf{I}^*)^{-1}(t,\mathbf{u}_\mathbf{I})\left[d\mathbf{u}_\mathbf{I} - (\mathbf{A}_0(t,\mathbf{u}_\mathbf{I}) + \mathbf{A}_1(t,\mathbf{u}_\mathbf{I})\bar{\mathbf{u}}_\mathbf{II})dt\right],$$

$$d\mathbf{R}_\mathbf{II}(t) = \Big\{\mathbf{a}_1(t,\mathbf{u}_\mathbf{I})\mathbf{R}_\mathbf{II} + \mathbf{R}_\mathbf{II}\mathbf{a}_1^*(t,\mathbf{u}_\mathbf{I}) + (\mathbf{\Sigma}_\mathbf{II}\mathbf{\Sigma}_\mathbf{II}^*)(t,\mathbf{u}_\mathbf{I}) - (\mathbf{R}_\mathbf{II}\mathbf{A}_1^*(t,\mathbf{u}_\mathbf{I}))(\mathbf{\Sigma}_\mathbf{I}\mathbf{\Sigma}_\mathbf{I}^*)^{-1}(t,\mathbf{u}_\mathbf{I})(\mathbf{R}_\mathbf{II}\mathbf{A}_1^*(t,\mathbf{u}_\mathbf{I}))^*\Big\}dt.$$

Examples of conditional Gaussian systems.

$$d\mathbf{u_I} = [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),$$

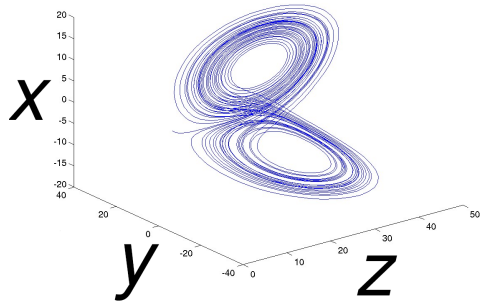
Noisy Lorenz 63 (L-63) model

$$dx = \sigma(y - x)dt + \sigma_x dW_x,$$
$$dy = (x(\rho - z) - y)dt + \sigma_y dW_y,$$
$$dz = (xy - \beta z)dt + \sigma_z dW_z.$$

$\rho = 28$

$\sigma = 10$

$\beta = 8/3$



Boussinesq equation

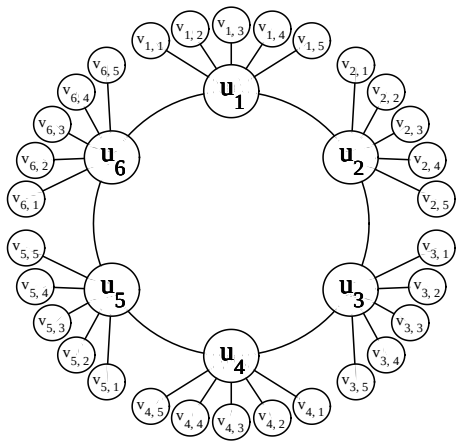
$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{1}{\rho_0}\nabla p + \nu \nabla^2 \mathbf{u} - g\alpha T,$$
$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \kappa \nabla^2 T + F.$$

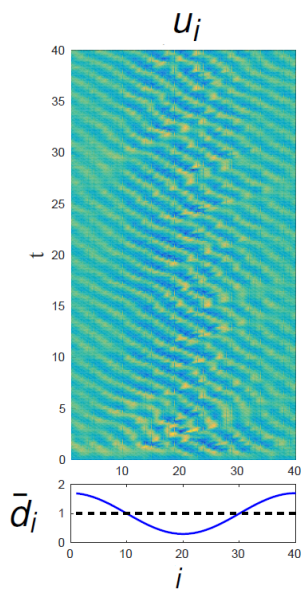
Finite Fourier-series expansion + stochastic noise.

A two-layer Lorenz 96 model

$$\frac{du_i}{dt} = u_{i-1}(u_{i+1} - u_{i-2}) + \sum_{j=1}^{J} \gamma_{i,j} u_i v_{i,j} - \bar{d}_i u_i + F + \sigma_u \dot{W}_{u_i}, \qquad i = 1, \ldots, I,$$

$$\frac{dv_{i,j}}{dt} = -d_{v_{i,j}} v_{i,j} - \gamma_j u_i^2 + \sigma_{i,j} \dot{W}_{v_{i,j}}, \quad j = 1, \ldots, J.$$





Outline

1. Predicting the large-scale Madden-Julian Oscillation (MJO) via a physics-constrained low-order nonlinear stochastic model.

2. Understanding the information barrier and data assimilation skill of recovering ocean flows with noisy Lagrangian tracers.

3. An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions with strongly non-Gaussian features.

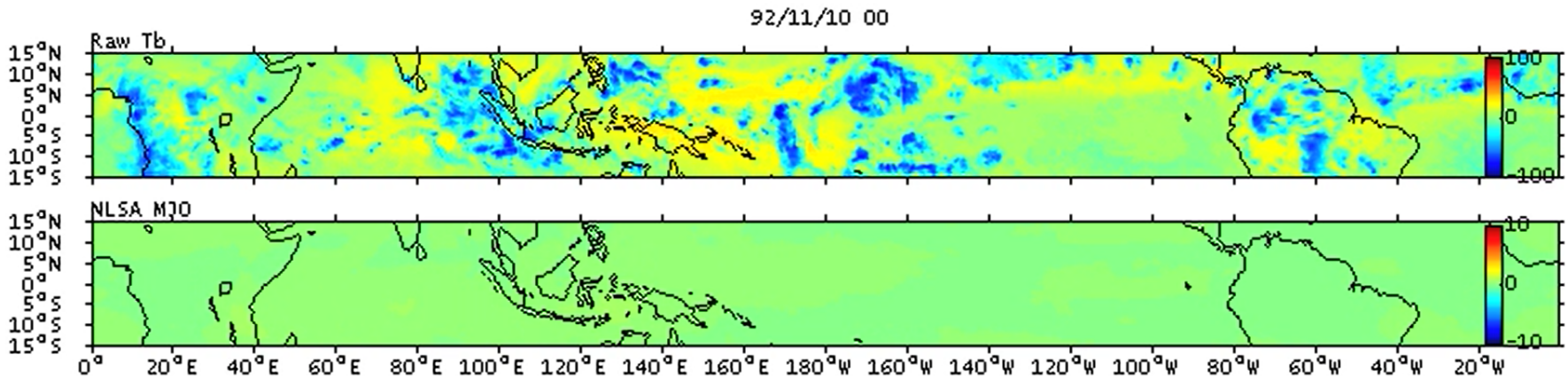
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

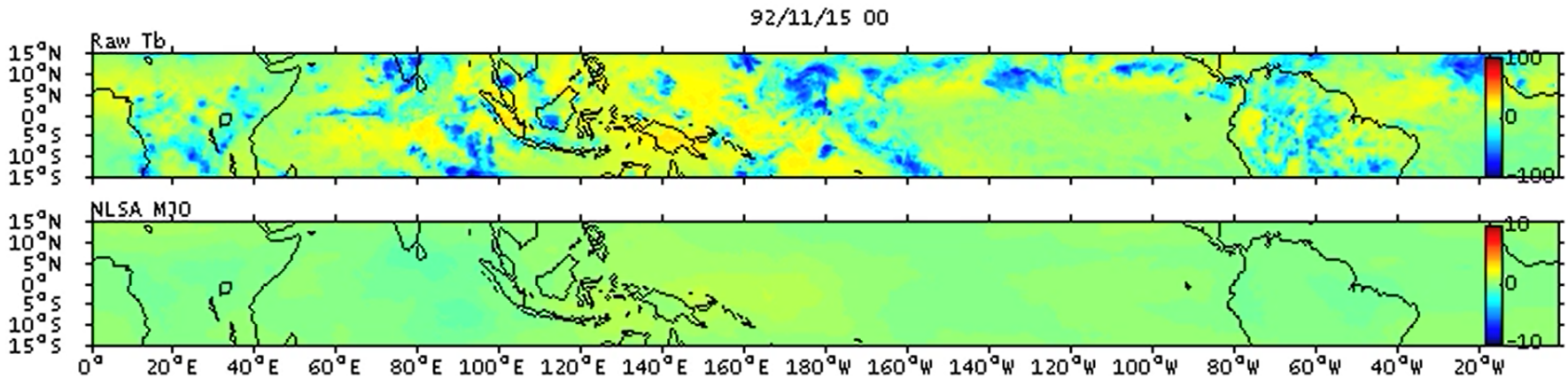
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

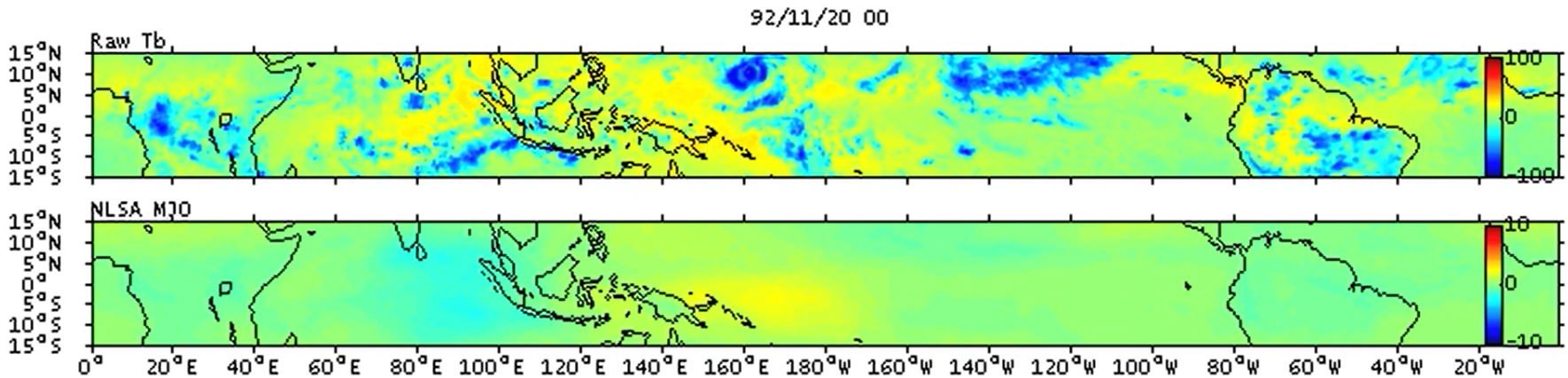
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

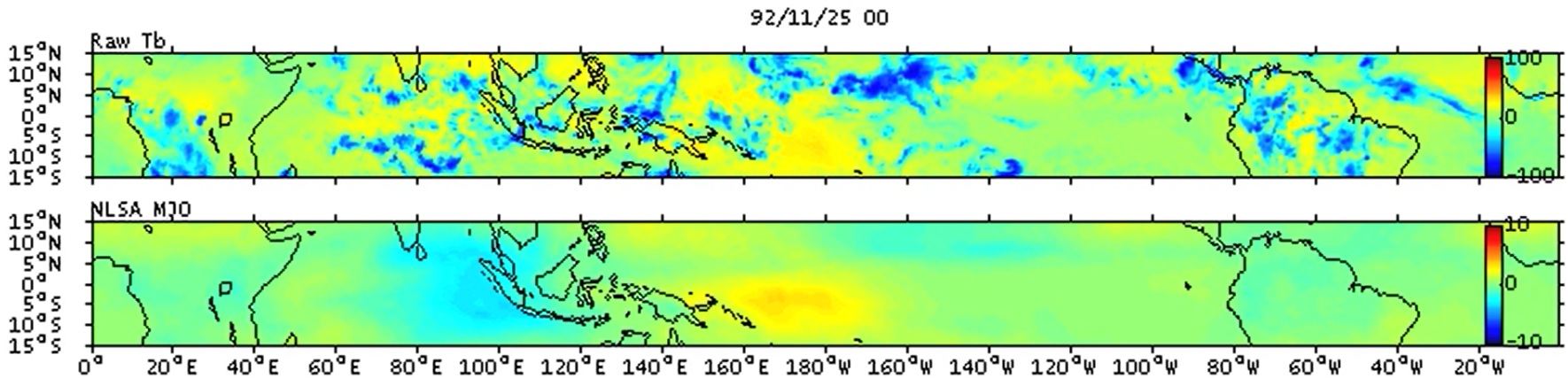
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

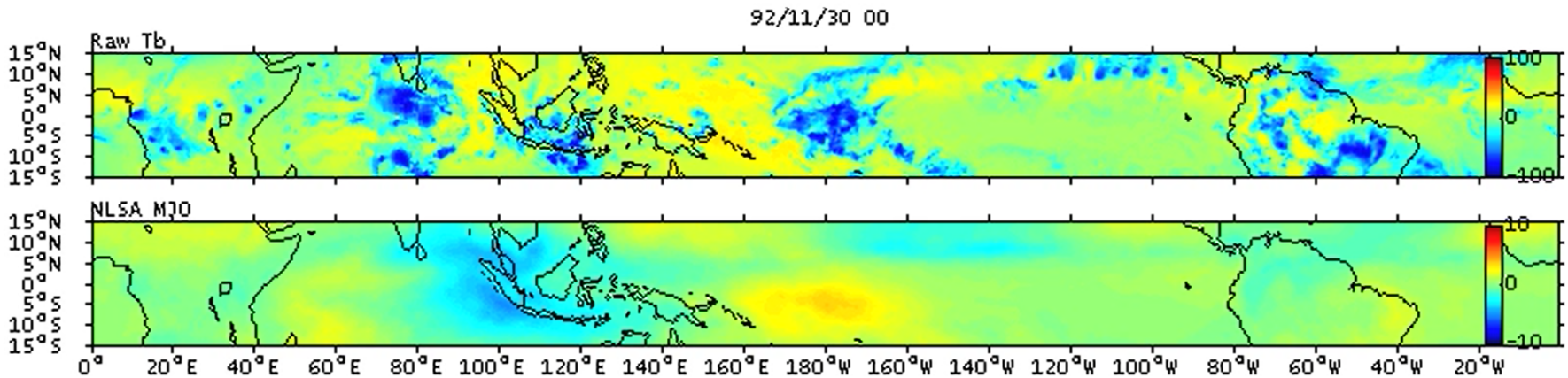
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

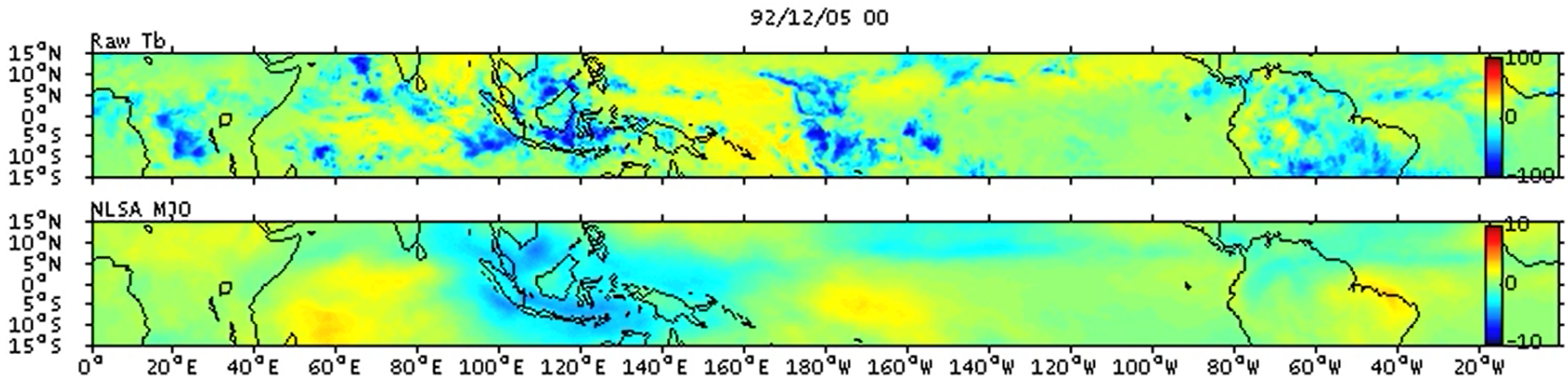
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

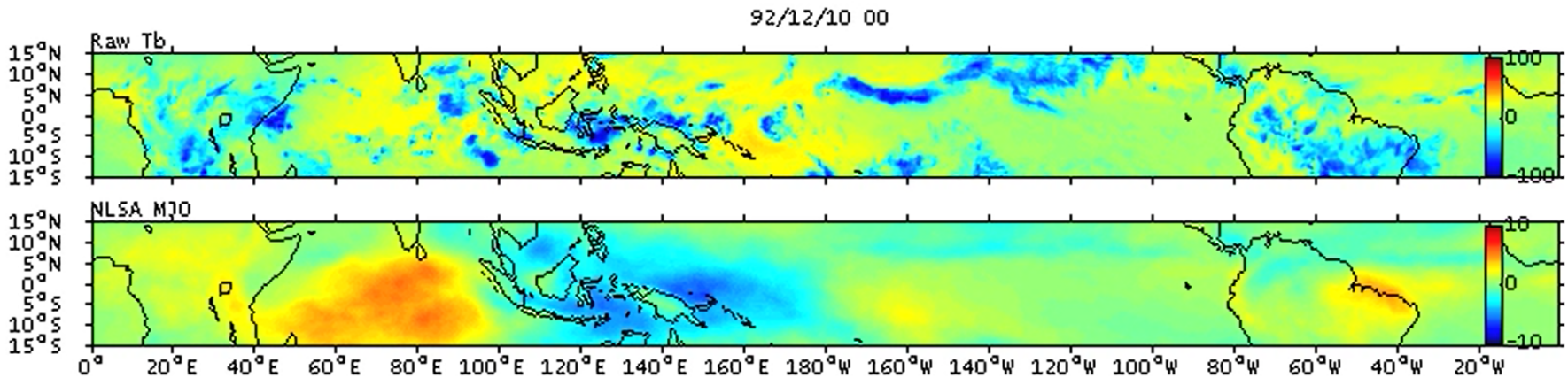
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

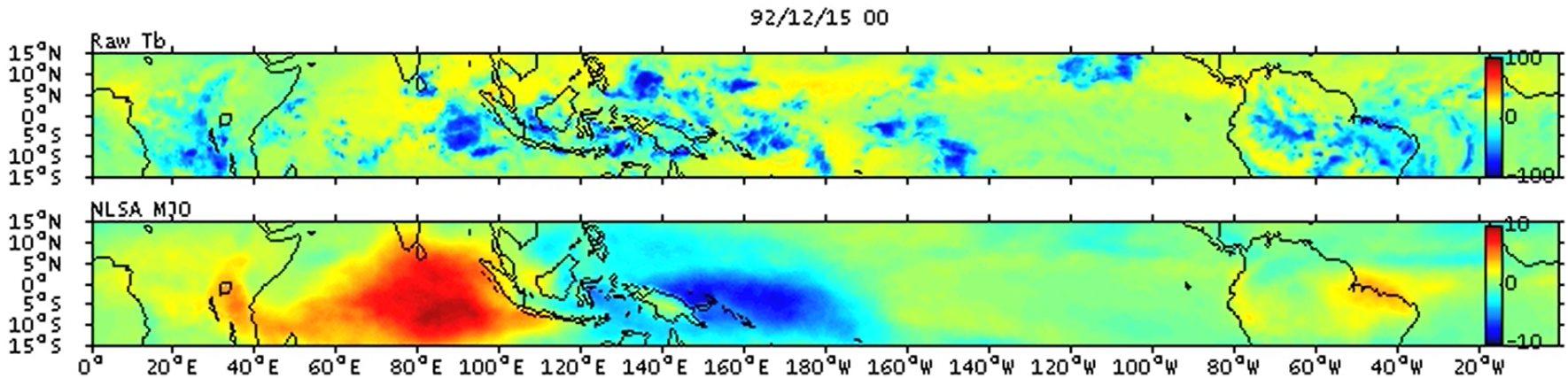
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

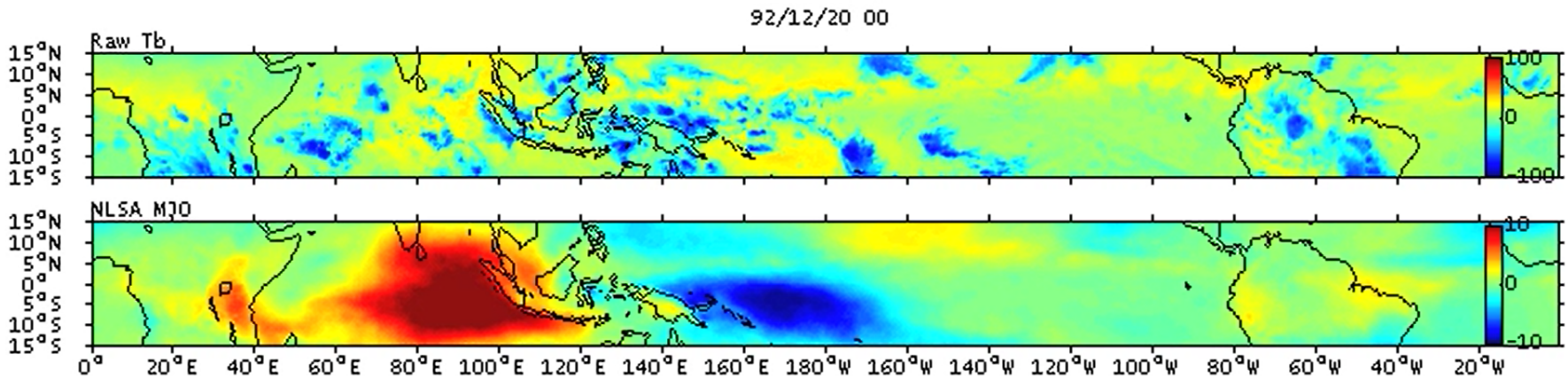
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

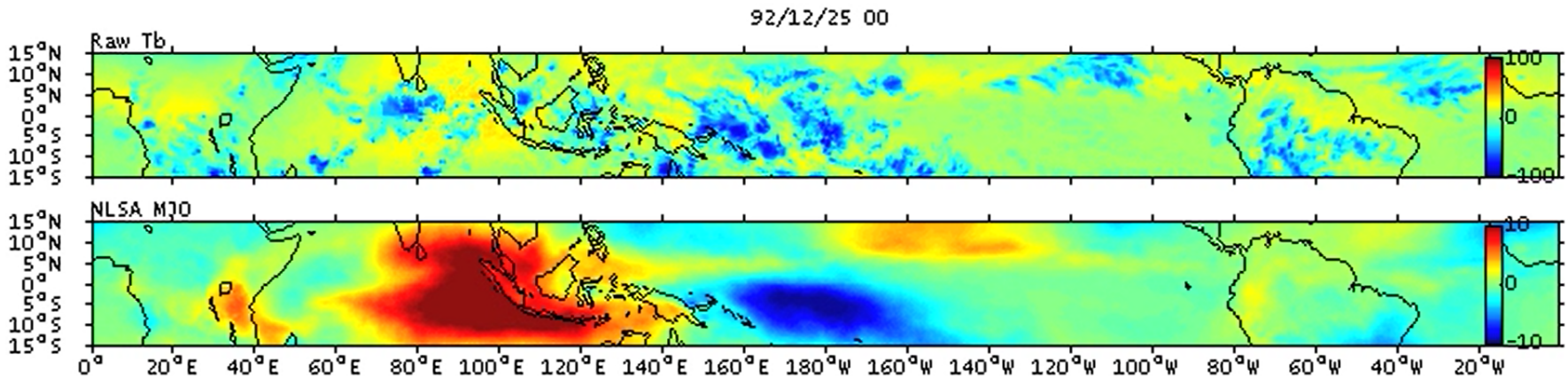
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

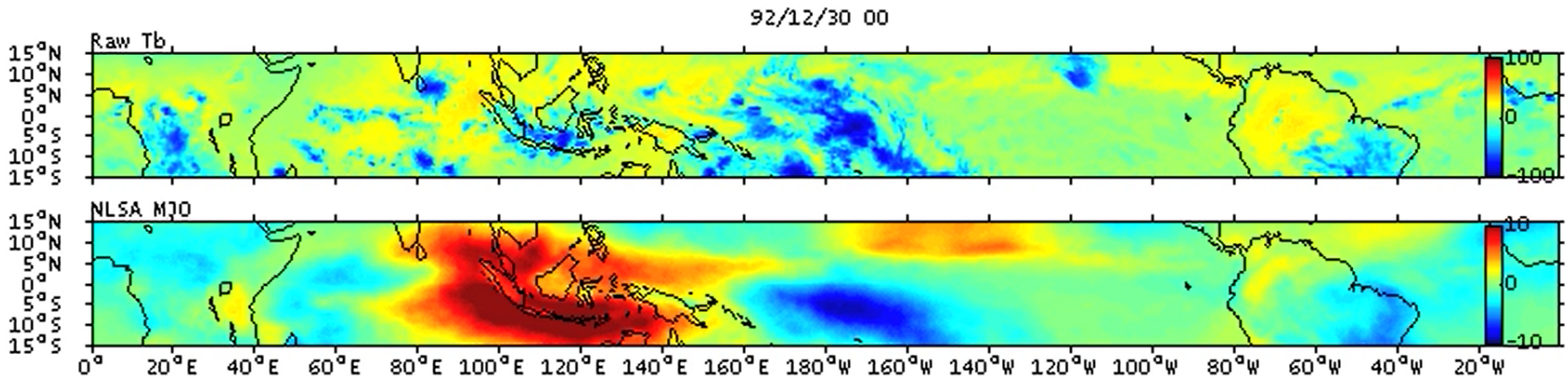
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

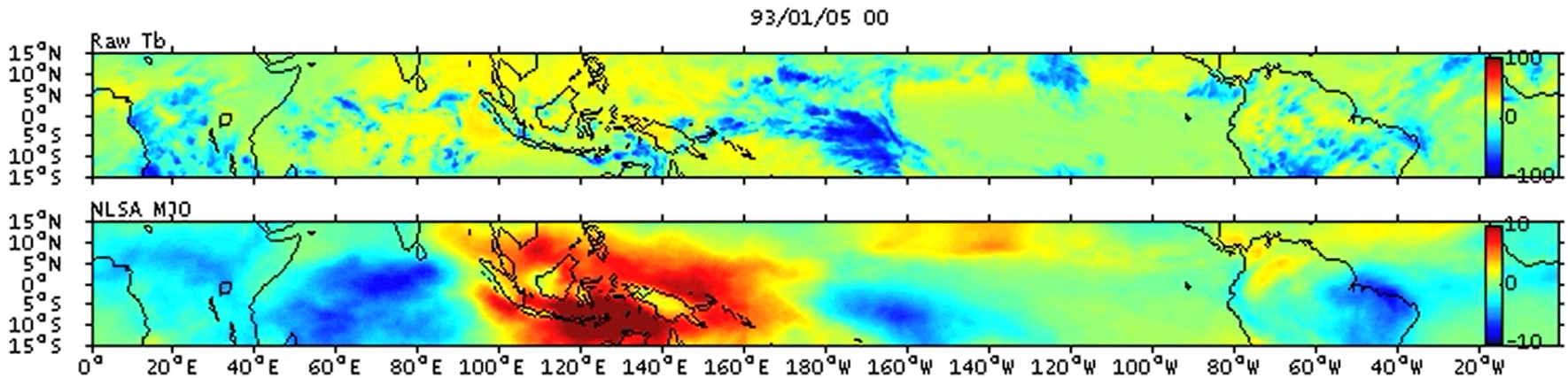
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

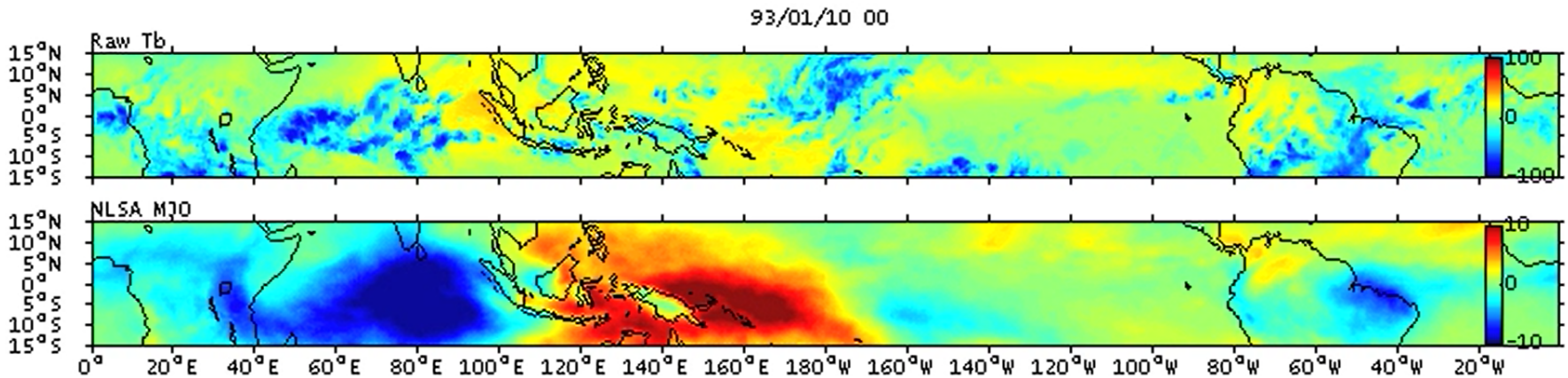
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

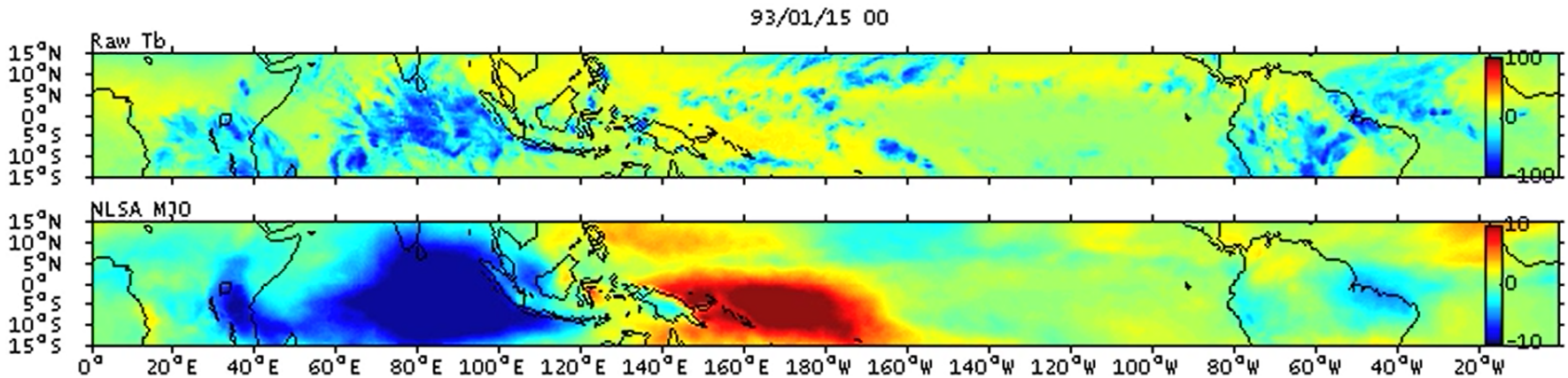
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

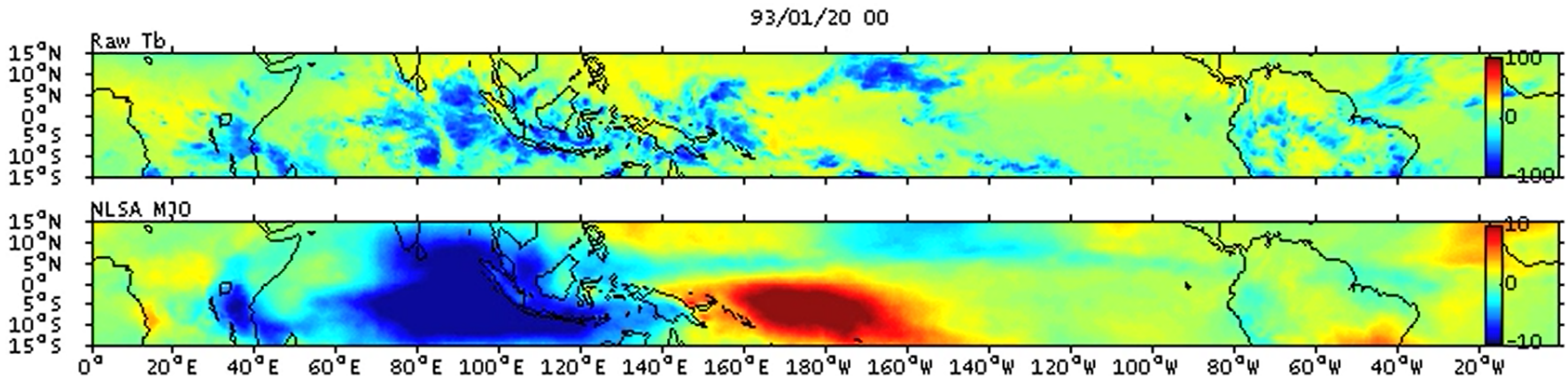
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

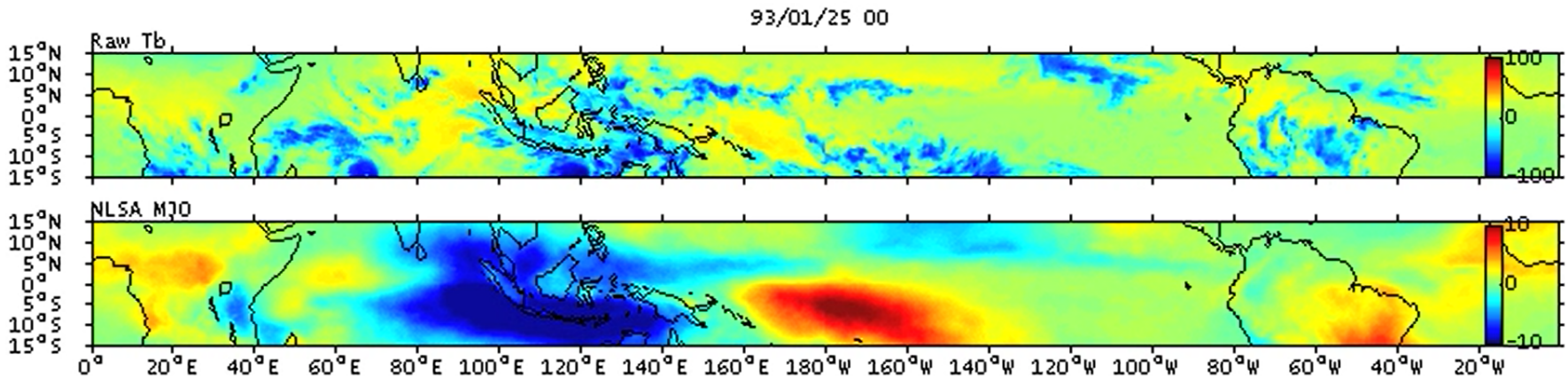
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

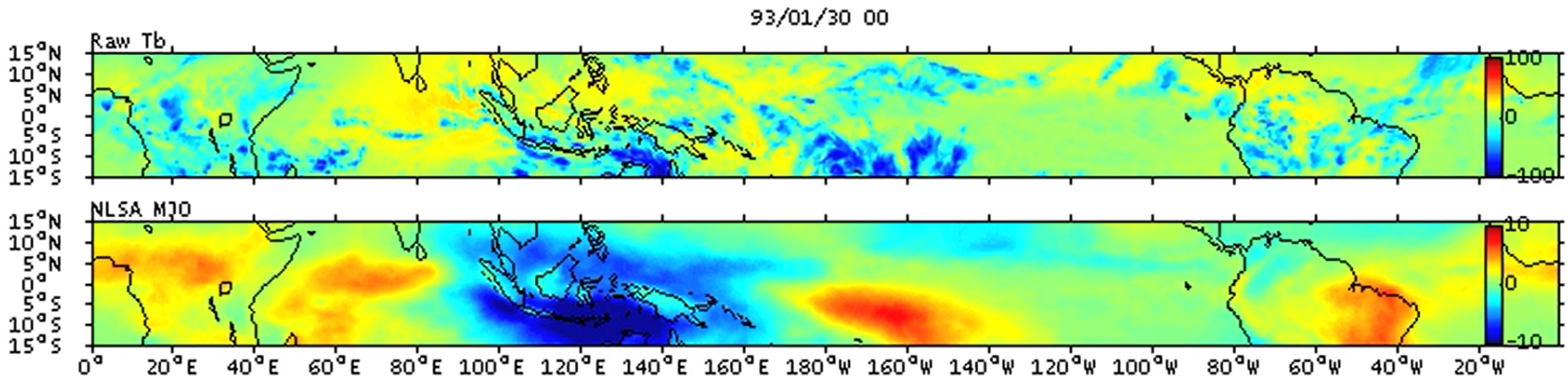
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

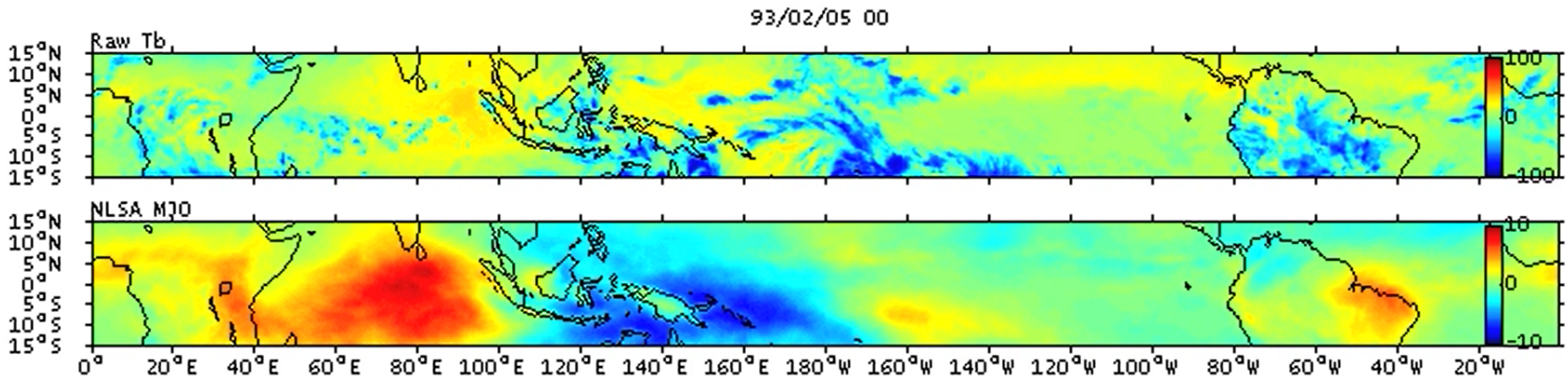
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

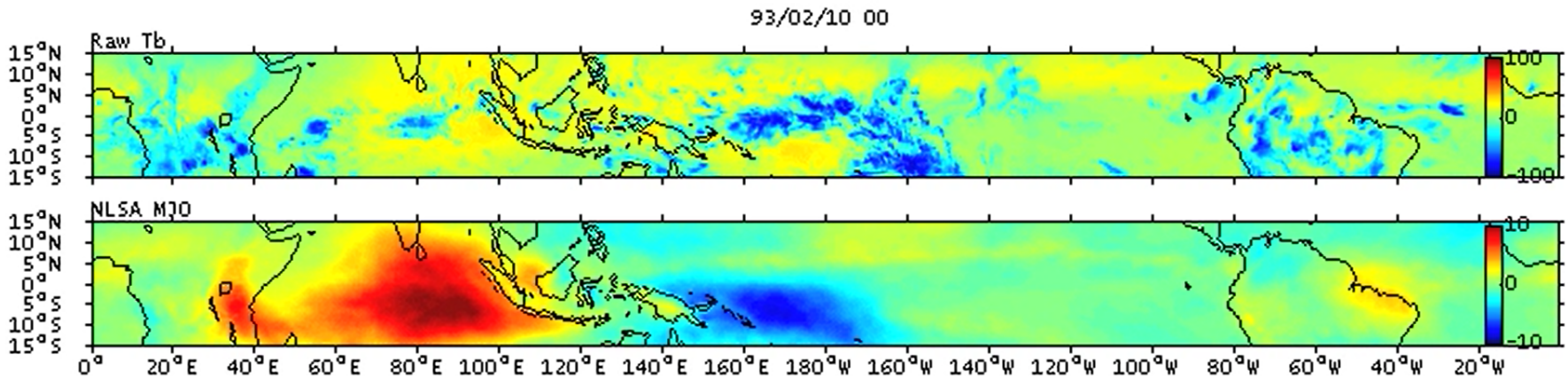
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

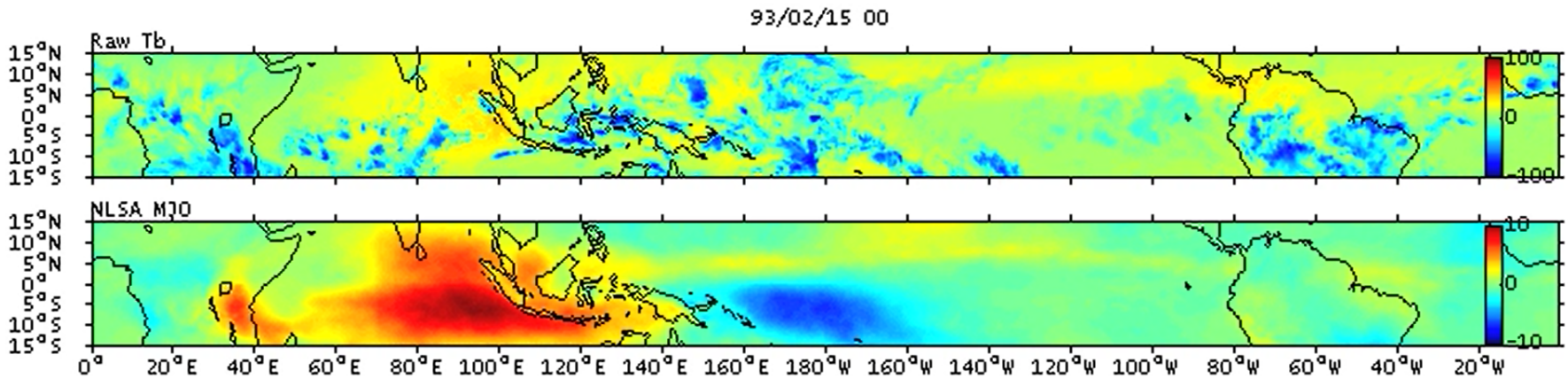
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

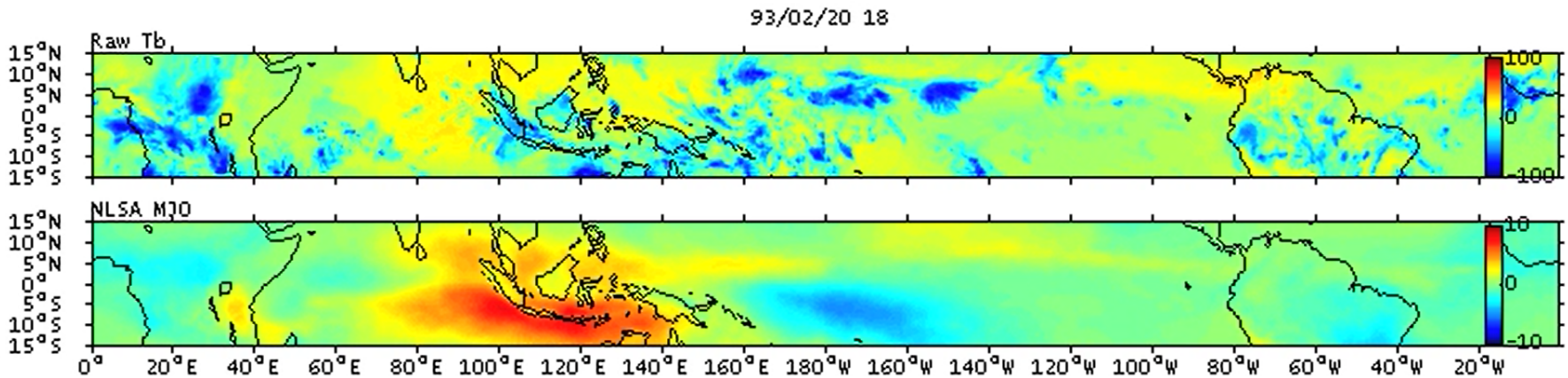
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

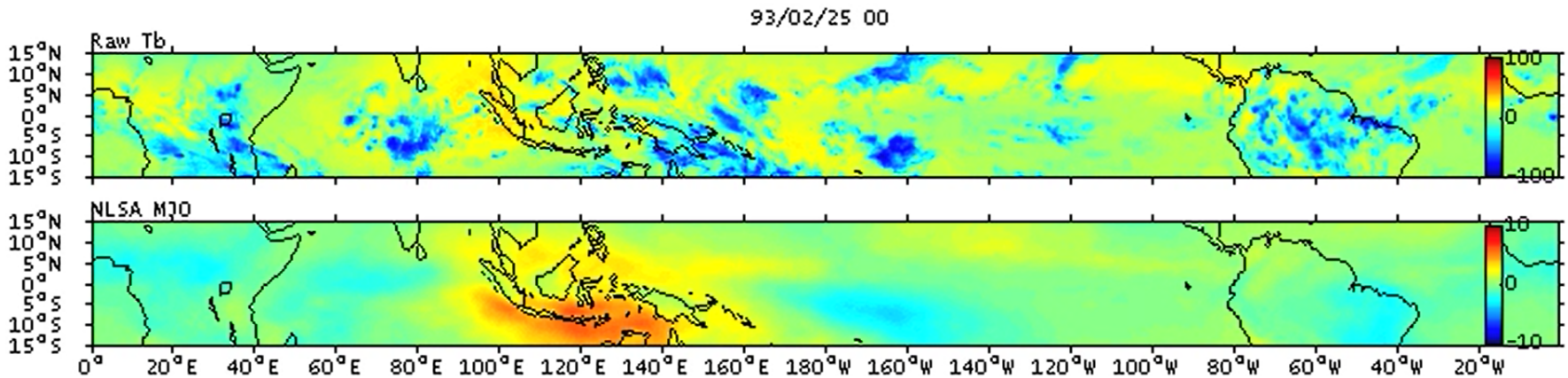
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

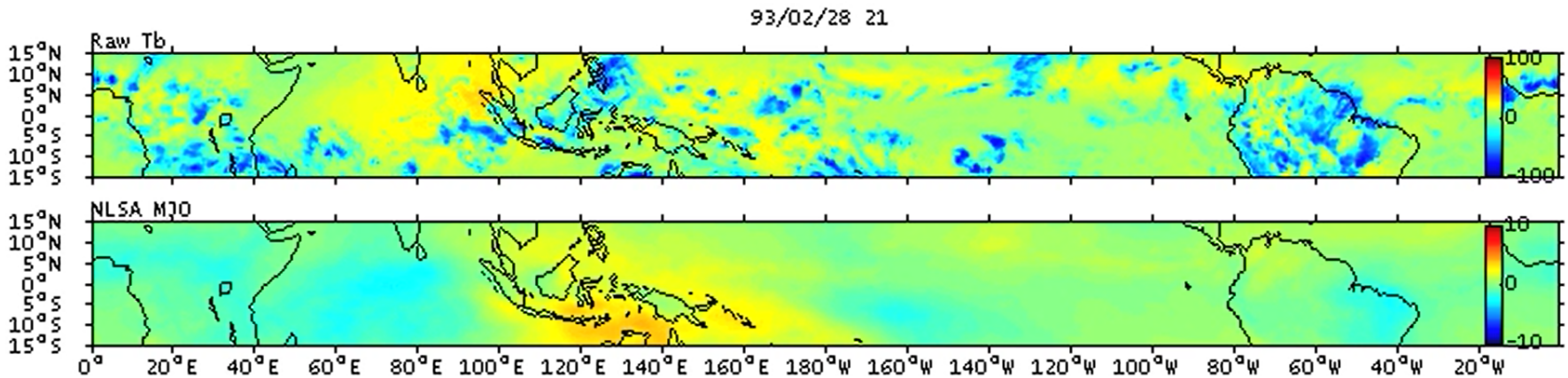
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

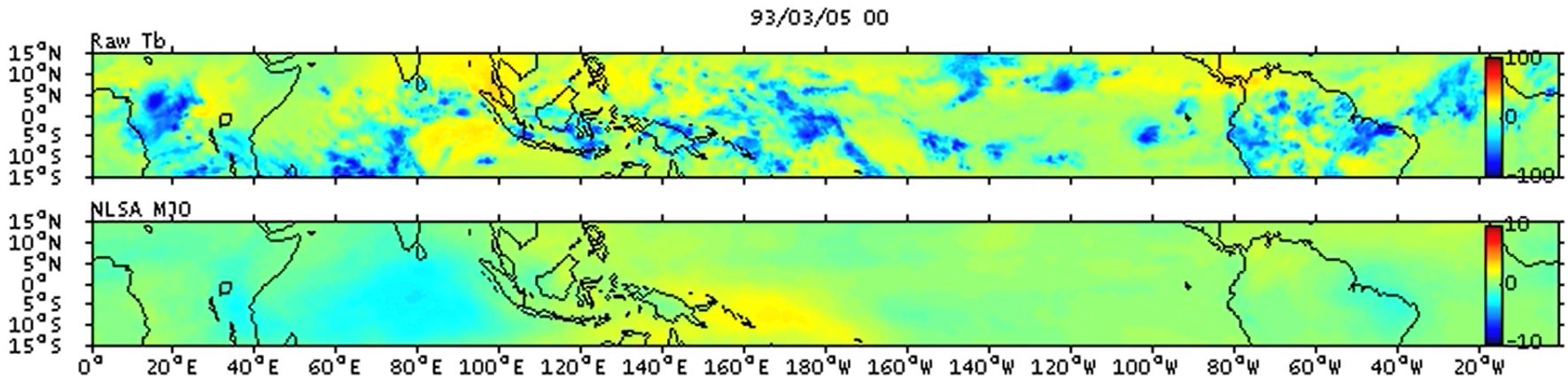
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

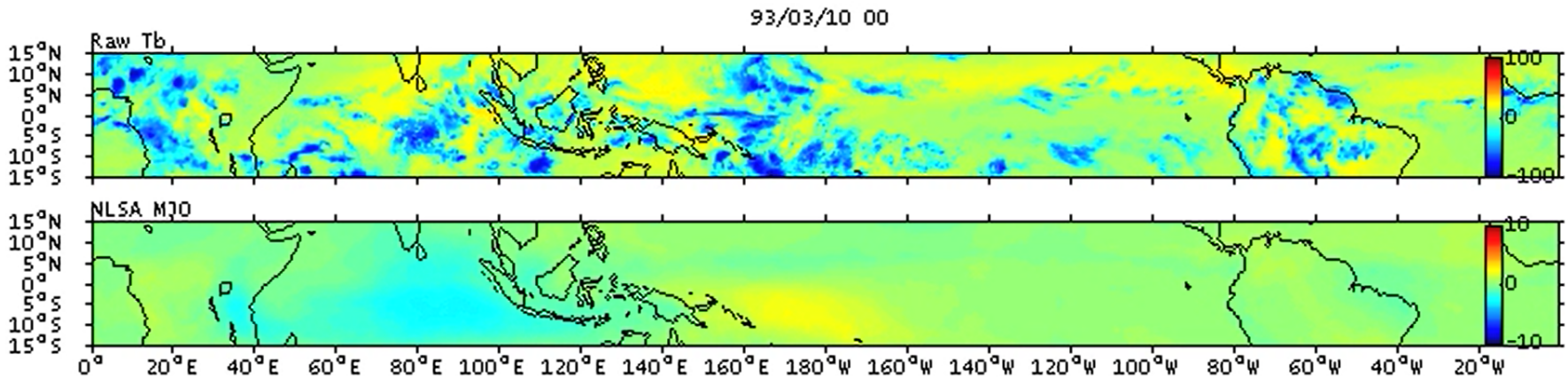
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

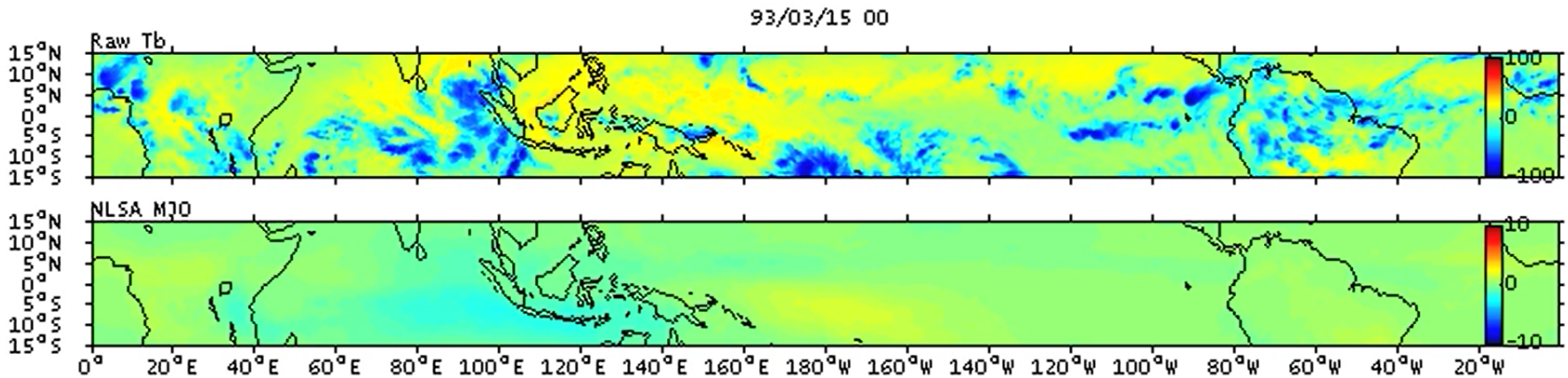
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

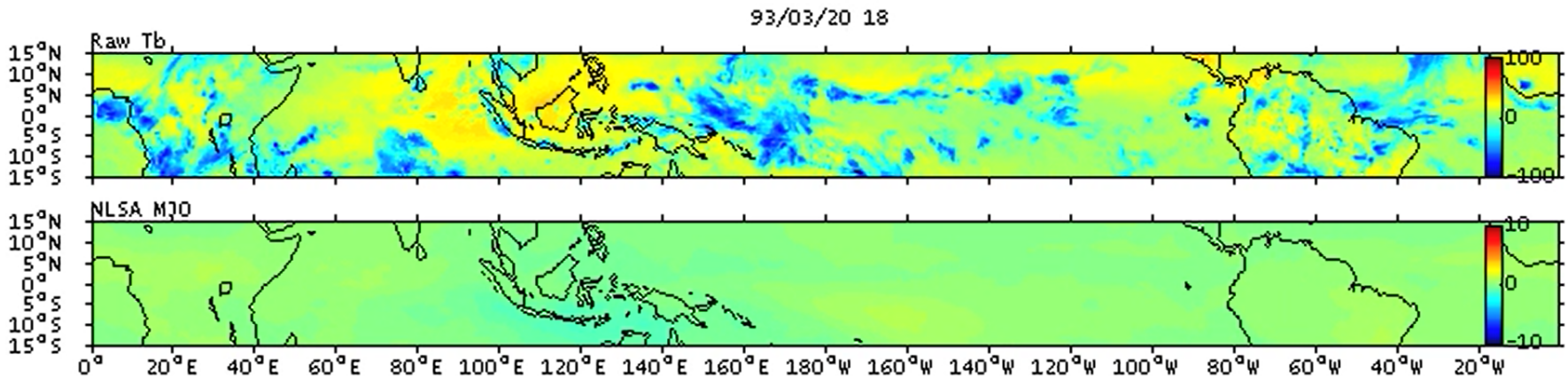
I. Predicting the Large-Scale Madden-Julian Oscillation

The Madden-Julian Oscillation (MJO) (Lau & Waliser 2011):

- the dominant mode of tropical intraseasonal (30-90 days) variability in boreal winter
- a slow eastward moving large-scale envelope of convection
- affecting tropical and global weather patterns, important triggering factor of the El Niño

Extracting the large-scale MJO from the noisy and turbulent raw data:

- Linear methods (e.g. EOFs/PCAs) may not be able to capture the nonlinear features.
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.

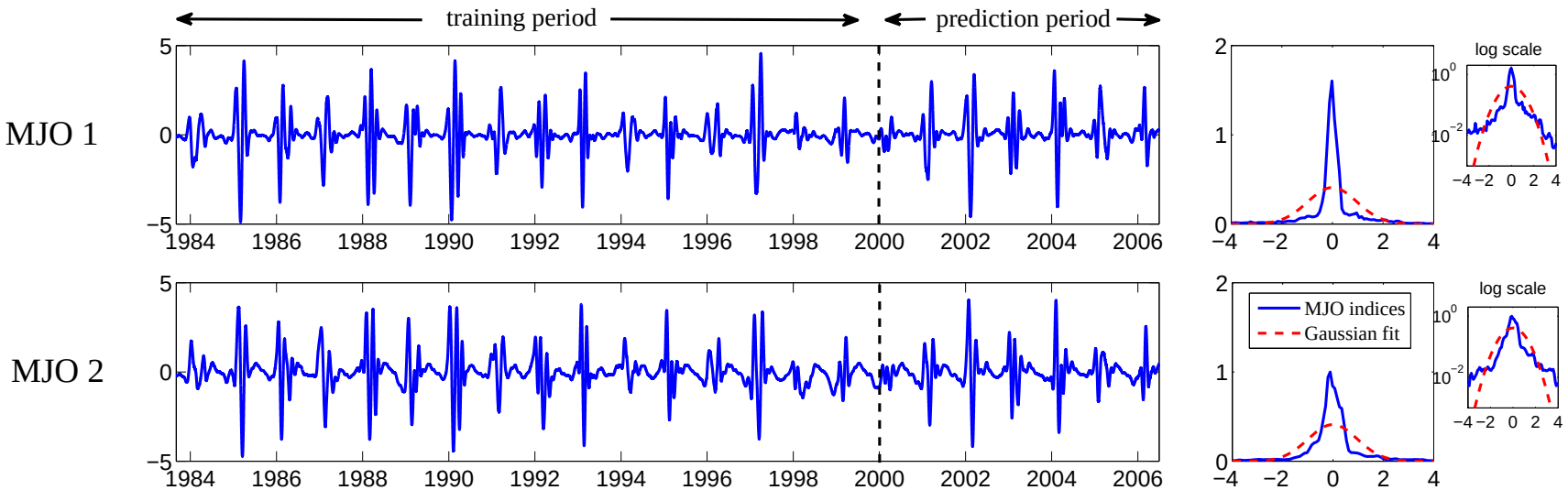


T_b: brightness temperature. (Movie source: Chen, Majda & Giannakis, *Geophys. Res. Lett.*, 2014.)

red: weak convection (clear sky). blue: strong convection (heavy rainfall).

Consistent with the MJOs observed during the TOGA-COARE of 1992-1993 (Webster & Lukas).

NLSA Time-Series Techniques $\Longrightarrow$ 2 components of MJO Cloud Patterns



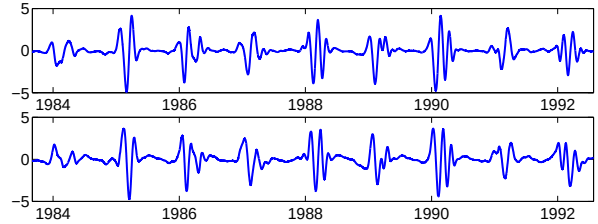
Intermittent bursts of MJO activity

Physics-Constrained Low-Order Nonlinear Stochastic Model for Predicting MJO Cloud Patterns (MJO1, MJO2)

(Chen, Majda, Giannakis, *Geophys. Res. Lett.*, 2014)

Physics-Constrained Low-Order Stochastic Model

$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 \qquad\quad -\hat{\omega}\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 \qquad\quad +\hat{\omega}\, u_1)\, dt + \sigma_u\, dW_{u_2},
\end{aligned}
$$



,

with

$$d_u(t) = d_{u0} + d_{u1} \sin(\omega_f t + \phi).$$

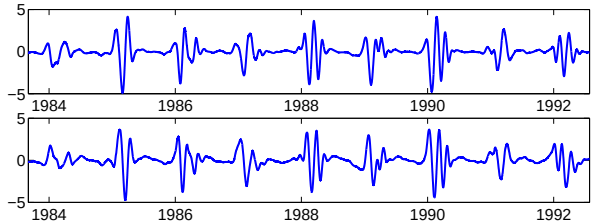
- Observed variables u_1, u_2: MJO 1 and MJO 2 indices from NLSA.
- **Standard regression model, insufficient in capturing the key features.**

Physics-Constrained Low-Order Stochastic Model

$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v\qquad\qquad\qquad\quad)\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\, dt + \sigma_\omega\, dW_\omega,
\end{aligned}
$$

with

$$
d_u(t) = d_{u0} + d_{u1} \sin(\omega_f t + \phi).
$$



Conditional Gaussian framework

$$
\begin{aligned}
d\mathbf{u_I} &= [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t), \\
d\mathbf{u_{II}} &= [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),
\end{aligned}
$$

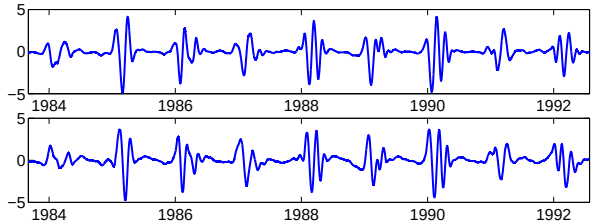
- Observed variables u_1, u_2: MJO 1 and MJO 2 indices from NLSA.
- Hidden variables v, ω: stochastic damping and stochastic phase.

Physics-Constrained Low-Order Stochastic Model

$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\, dt + \sigma_\omega\, dW_\omega,
\end{aligned}
$$

with

$$
d_u(t) = d_{u0} + d_{u1} \sin(\omega_f t + \phi).
$$



Conditional Gaussian framework

$$
\begin{aligned}
d\mathbf{u_I} &= [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t), \\
d\mathbf{u_{II}} &= [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),
\end{aligned}
$$

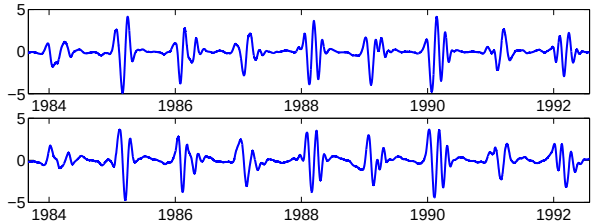
- Observed variables u_1, u_2: MJO 1 and MJO 2 indices from NLSA.
- Hidden variables v, ω: stochastic damping and stochastic phase.
- Energy-conserving nonlinear interactions between (u_1, u_2) and (v, ω).
  (Majda, Harlim, 2012)

Physics-Constrained Low-Order Stochastic Model

$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\, dt + \sigma_\omega\, dW_\omega,
\end{aligned}
$$

with

$$
d_u(t) = d_{u0} + d_{u1} \sin(\omega_f t + \phi).
$$



Conditional Gaussian framework

$$
\begin{aligned}
d\mathbf{u_I} &= [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t), \\
d\mathbf{u_{II}} &= [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),
\end{aligned}
$$

- Observed variables u_1, u_2: MJO 1 and MJO 2 indices from NLSA.
- Hidden variables v, ω: stochastic damping and stochastic phase.
- Energy-conserving nonlinear interactions between (u_1, u_2) and (v, ω).
  (Majda, Harlim, 2012)

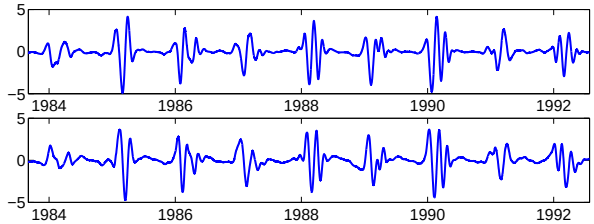
Prediction. Given the initial values of (u_1, u_2) and (v, ω), run an ensemble forecast.

Physics-Constrained Low-Order Stochastic Model

$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\, dt + \sigma_\omega\, dW_\omega,
\end{aligned}
$$

with

$$
d_u(t) = d_{u0} + d_{u1} \sin(\omega_f t + \phi).
$$



Conditional Gaussian framework

$$
\begin{aligned}
d\mathbf{u_I} &= [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t), \\
d\mathbf{u_{II}} &= [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\
&\quad + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),
\end{aligned}
$$

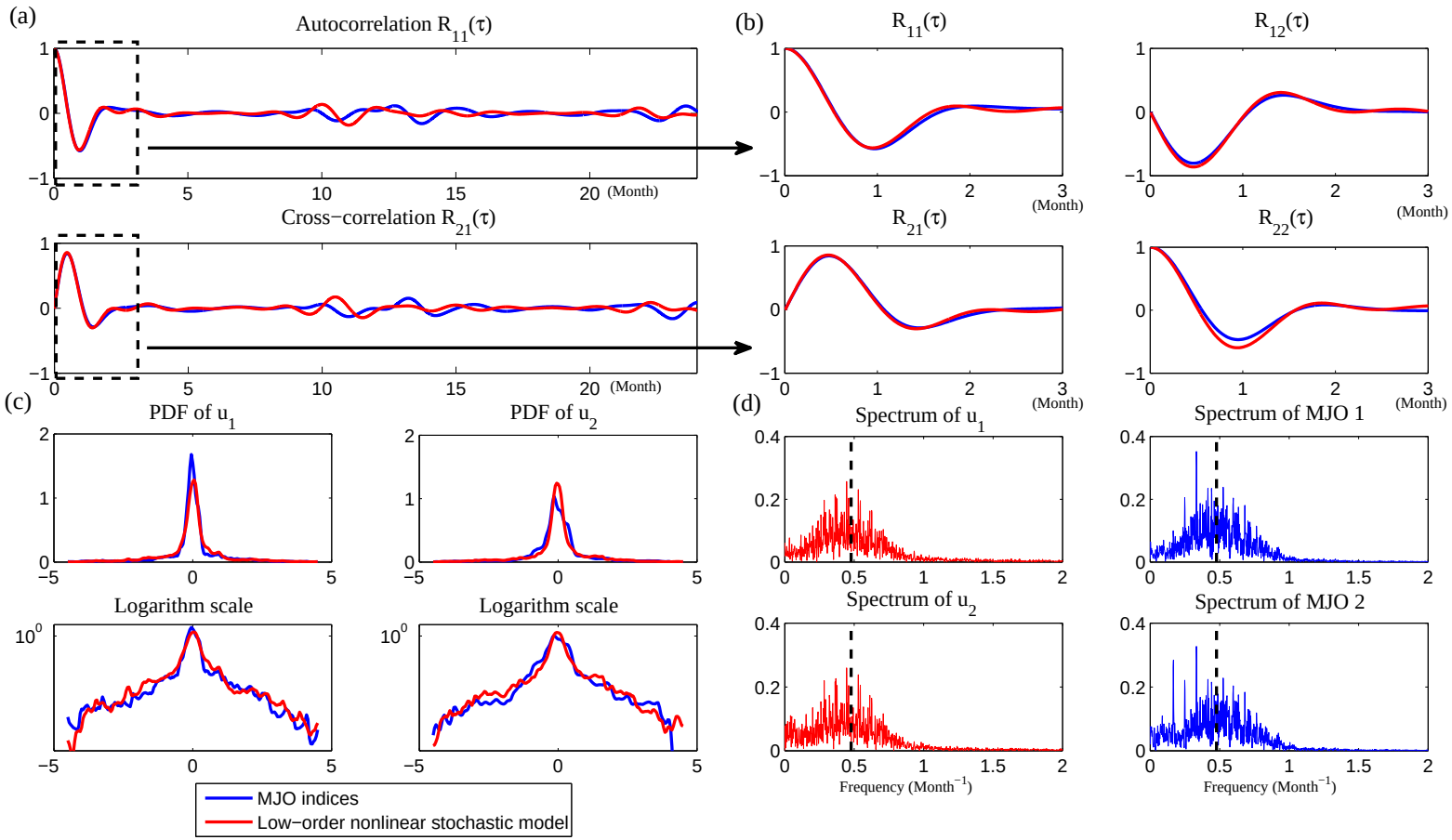
- Observed variables u_1, u_2: MJO 1 and MJO 2 indices from NLSA.
- Hidden variables v, ω: stochastic damping and stochastic phase.
- Energy-conserving nonlinear interactions between (u_1, u_2) and (v, ω).
  (Majda, Harlim, 2012)

Prediction. Given the initial values of (u_1, u_2) and (v, ω), run an ensemble forecast.

How to determine the initial values of the hidden variables?
Effective data assimilation algorithm based on conditional Gaussian framework!

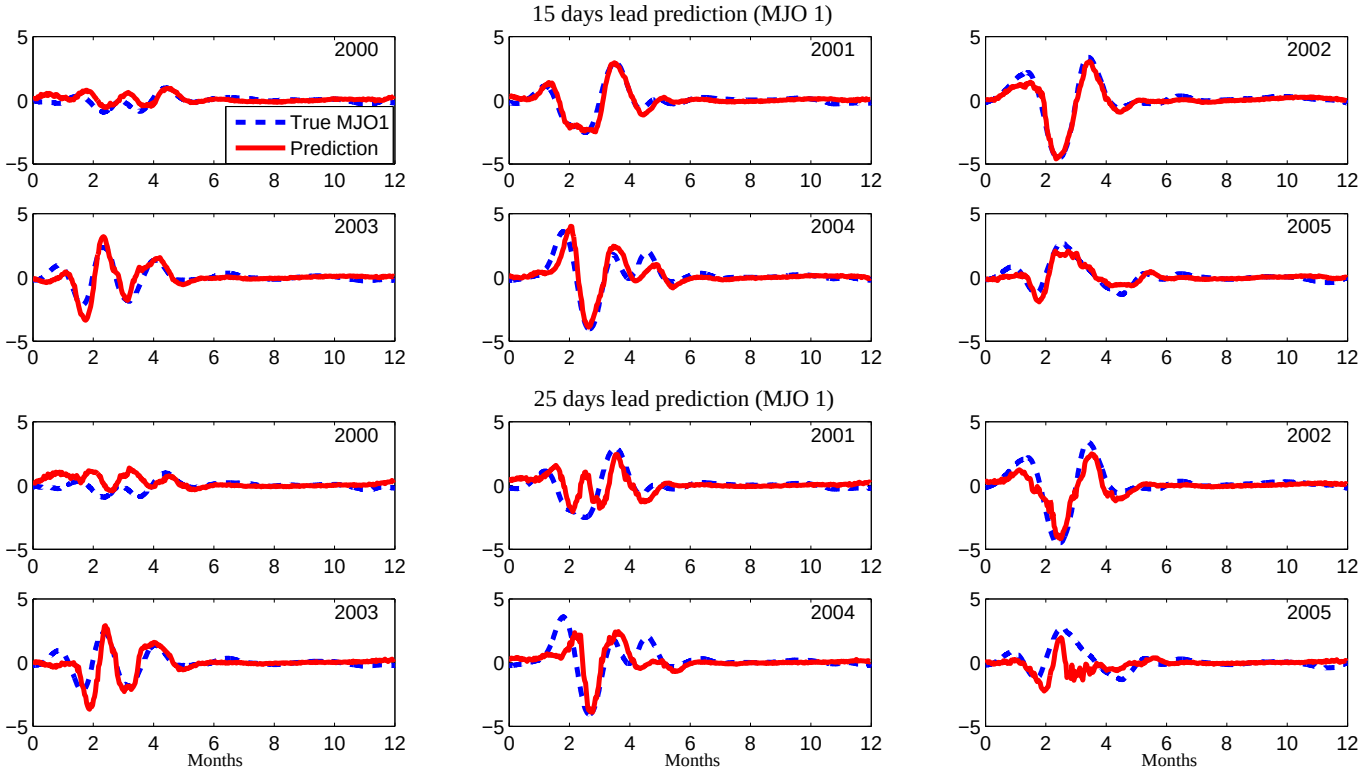
Calibration of parameters using **Information Theory** (Robust parameters)

Model vs. Observations: Non-Gaussian statistics match



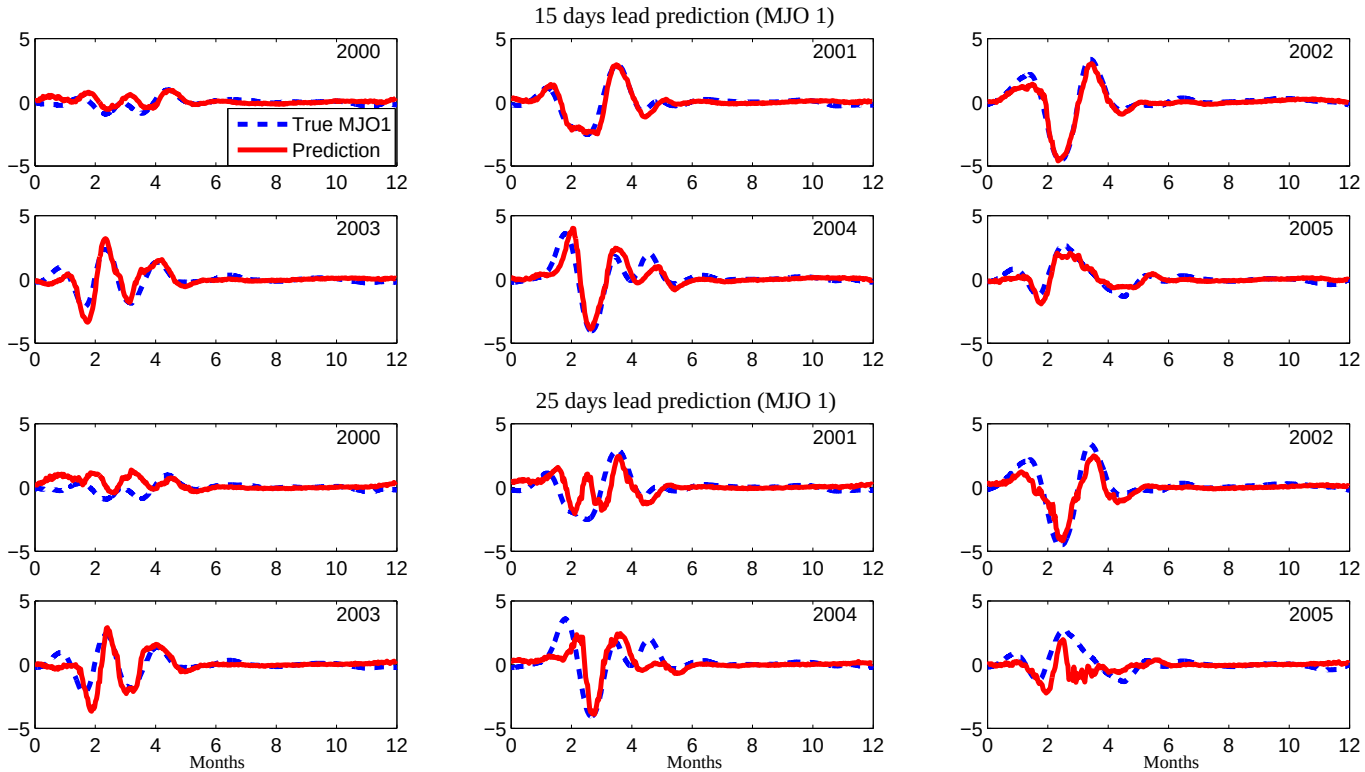
Almost perfectly match correlation functions, PDFs, and power spectrums.

Skillful prediction at 15- and 25-days lead times



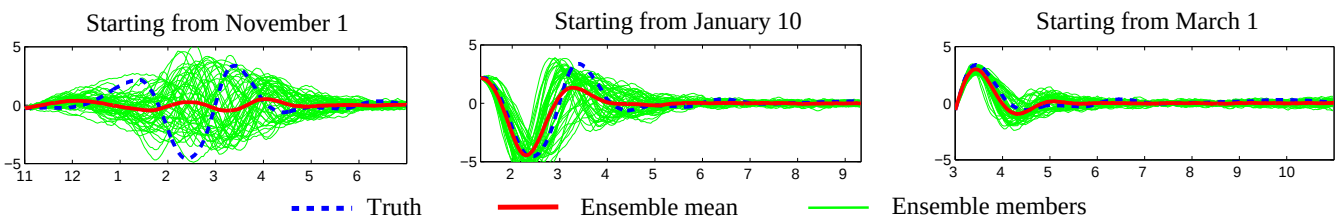
- Skillful prediction $\sim$ 25–40 days
- Reaching the predictability limit of the MJO indices (based on twin experiments).

Skillful prediction at 15- and 25-days lead times



- Skillful prediction $\sim$ 25–40 days
- Reaching the predictability limit of the MJO indices (based on twin experiments).

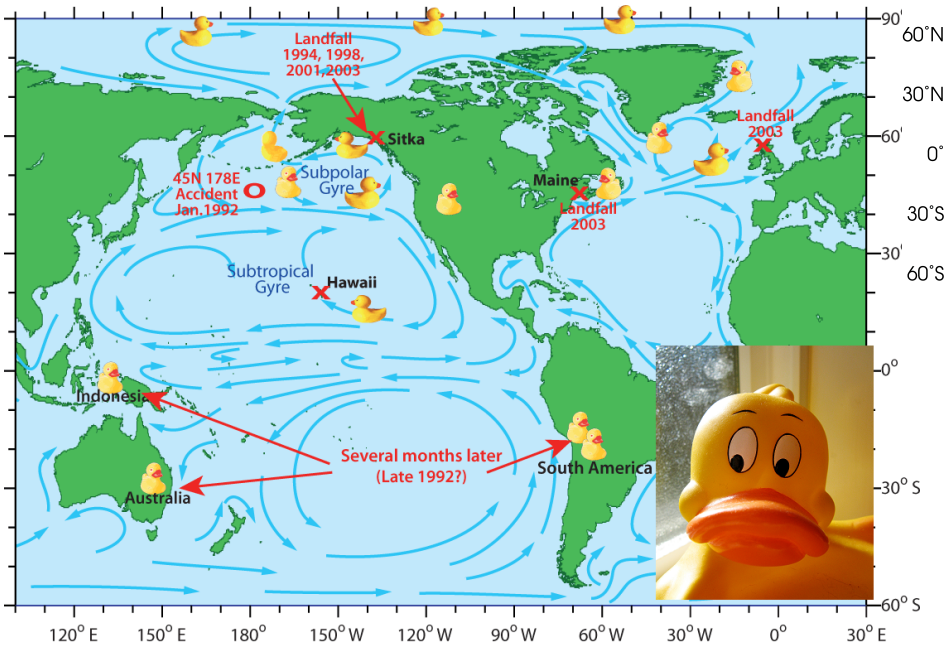
Long-range forecast starting from different dates (year 2002)

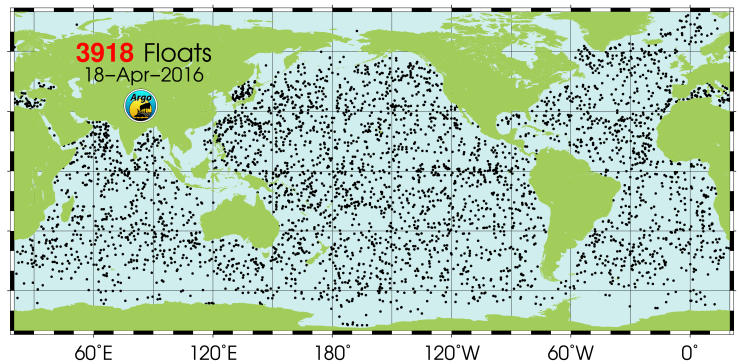


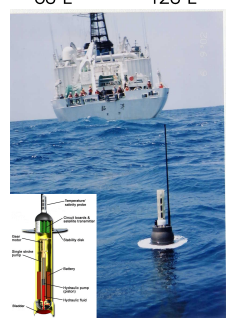
Ensemble spread $\Longleftrightarrow$ long-range forecast uncertainty is captured

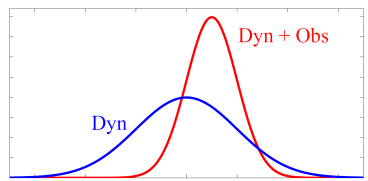
II. Noisy Lagrangian Tracers in Filtering Geophysical Flows

- Lagrangian tracers: drifters/floaters following a parcel of fluid's movement.
- **[Inverse Problems.]** Data assimilation with Lagrangian tracers: recovering the underlying velocity field with observations (from tracers).
  - Only dynamics: large uncertainty due to turbulence.
  - Dynamics + Observations: reducing error and uncertainty.









C. Jones, A. Apte, A. Stuart, ...

- What is the information gain as a function of the number of tracers?
- How to design cheap practical strategies for systems with multiscale and turbulent features?

Model set-up.

1. Underlying flows

Consider a d dimensional random flow modeled by a finite number of Fourier modes with random amplitudes in periodic domain $(0, 2\pi]^d$,

$$\vec{v}(\vec{x}, t) = \sum_{\vec{k} \in \mathbf{K}} \hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}} \cdot \vec{r}_{\vec{k}}.$$

Each $\hat{v}_{\mathbf{k}}(t)$ follows an Ornstein-Uhlenbeck (O.U.) process,

$$d\hat{v}_{\vec{k}}(t) = -d_{\vec{k}}\hat{v}_{\vec{k}}(t)dt + f_{\vec{k}}(t)dt + \sigma_{\vec{k}}dW^v_{\vec{k}}(t).$$

2. Observations

The observations are given by the trajectories of L noisy Lagrangian tracers,

$$\begin{aligned} d\vec{x}_l(t) &= \vec{v}(\vec{x}_l(t), t)dt + \sigma_x dW^x_l(t) \\ &= \sum_{\vec{k} \in \mathbf{K}} \underbrace{\hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}_l(t)} \cdot \vec{r}_{\vec{k}}}_{\text{Nonlinear!}} dt + \sigma_x dW^x_l(t), \quad l = 1, \ldots, L. \end{aligned}$$

Model set-up.

1. Underlying flows

Consider a d dimensional random flow modeled by a finite number of Fourier modes with random amplitudes in periodic domain $(0, 2\pi]^d$,

$$\vec{v}(\vec{x}, t) = \sum_{\vec{k} \in \mathbf{K}} \hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}} \cdot \vec{r}_{\vec{k}}.$$

Each $\hat{v}_{\mathbf{k}}(t)$ follows an Ornstein-Uhlenbeck (O.U.) process,

$$d\hat{v}_{\vec{k}}(t) = -d_{\vec{k}}\hat{v}_{\vec{k}}(t)dt + f_{\vec{k}}(t)dt + \sigma_{\vec{k}}dW^v_{\vec{k}}(t).$$

2. Observations

The observations are given by the trajectories of L noisy Lagrangian tracers,

$$\begin{aligned} d\vec{x}_l(t) &= \vec{v}(\vec{x}_l(t), t)dt + \sigma_x dW^x_l(t) \\ &= \sum_{\vec{k} \in \mathbf{K}} \underbrace{\hat{v}_{\vec{k}}(t) \cdot e^{i\vec{k}\cdot\vec{x}_l(t)} \cdot \vec{r}_{\vec{k}}}_{\text{Nonlinear!}} dt + \sigma_x dW^x_l(t), \quad l = 1, \ldots, L. \end{aligned}$$

3. Conditional Gaussian data assimilation framework ($d = 2$)

$\mathbf{U} = (\hat{v}_{\mathbf{1}}, \ldots \hat{v}_{\mathbf{K}})^{\mathbb{T}}$, $\quad \mathbf{X} = (x_{1,x}, x_{1,y}, \ldots, x_{L,x}, x_{L,y})^{\mathbb{T}}$

Observations:	$d\mathbf{X} = \mathbf{P}_X(\mathbf{X})\mathbf{U}dt + \Sigma_x dW_X,$
Underlying flow:	$d\mathbf{U} = -\Gamma\mathbf{U}dt + \mathbf{F}(t)dt + \Sigma_u dW_u.$

Conditional Gaussian framework

$$\begin{aligned} d\mathbf{u_I} &= [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\ &\quad + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t), \\ d\mathbf{u_{II}} &= [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt \\ &\quad + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t), \end{aligned}$$

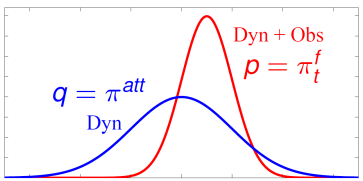
1. Recovering random incompressible flows
First rigorous math theory

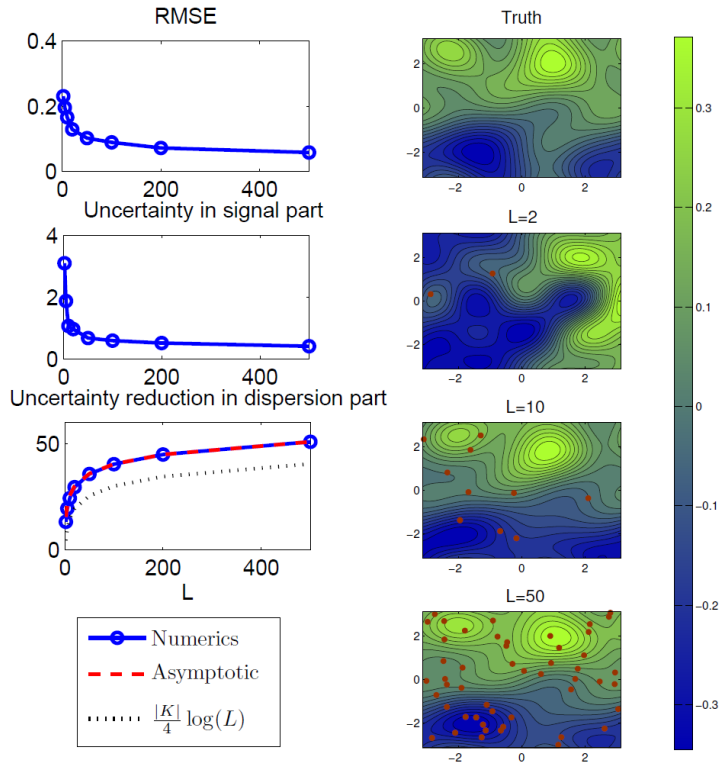
(Chen, Majda & Tong, *Nonlinearity*, 2014)

Recover or estimate the velocity $\vec{v}$ by observing L noisy trajectories $X_j(t)$,

$$\frac{dX_j}{dt} = v(X_j(t), t) + \sigma_j \dot{W}_j.$$

- Inherent nonlinearity in measurement.
- Build exact closed analytic formulas for the optimal filter for the velocity field.
- Show in a rigorous way that an exponential increase in the number of tracers for reducing the uncertainty by a fixed amount — **a practical information barrier**.





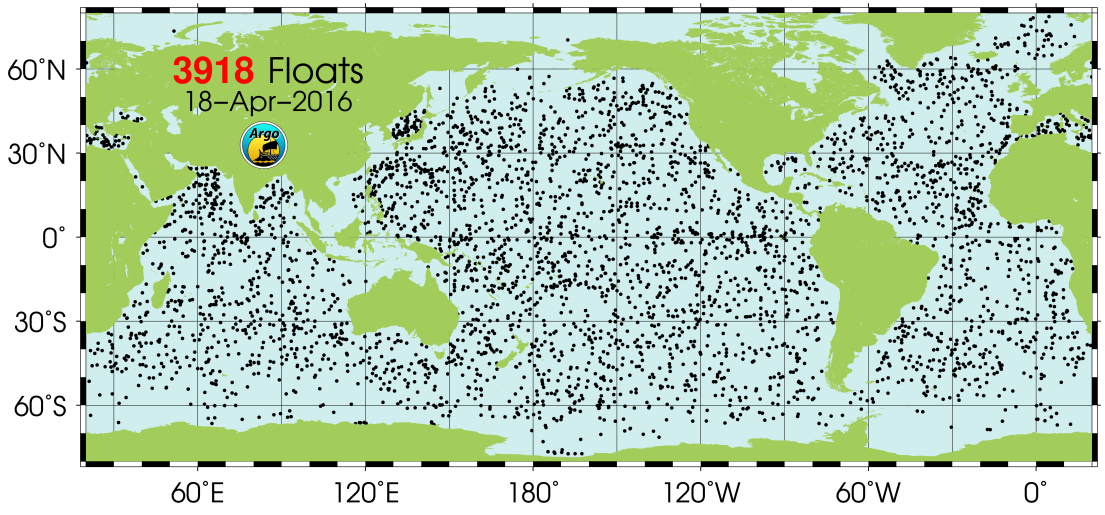
$$\mathcal{P} = \int p \ln \frac{p}{q} = \underbrace{\frac{1}{2}\left[(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T R_q^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)\right]}_{\text{signal}} + \underbrace{\frac{1}{2}\left[\text{tr}\left(R_p R_q^{-1}\right) - |\mathbf{K}| - \ln\det\left(R_p R_q^{-1}\right)\right]}_{\text{dispersion}}$$

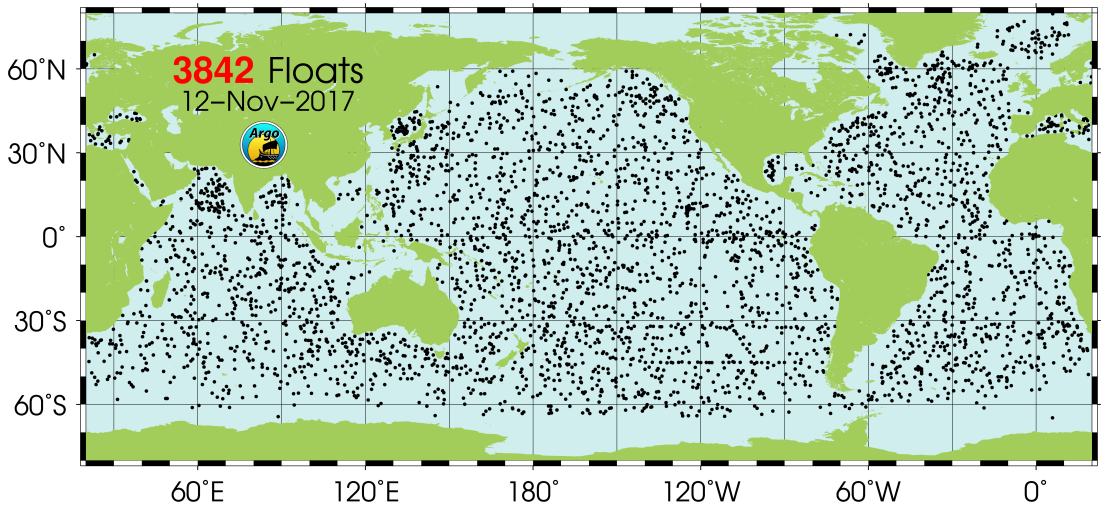
- Signal measures the lack of information in the mean weighted by model covariance.
- Dispersion involves the covariance ratio.

$$p \sim \mathcal{N}(\boldsymbol{\mu}_p, R_p),\ q \sim \mathcal{N}(\boldsymbol{\mu}_q, R_q)$$

How many tracers are in the real ocean? — Let's look at the Argo program

http://www.argo.ucsd.edu/





Great! No exponential increase of the tracer numbers!

2. Noisy Lagrangian tracers for filtering random rotating compressible flows

(Chen, Majda & Tong, *JNLS* 2015; Chen & Majda, *Monthly Weather Review*, 2016)

Starting model – 2D shallow water equation (SWE),

$$\frac{\partial \vec{u}}{\partial t} + \varepsilon^{-1}\vec{u}^{\perp} = -\varepsilon^{-1}\nabla\eta,$$

$$\frac{\partial \eta}{\partial t} + \varepsilon^{-1}\delta\nabla\cdot\vec{u} = 0.$$

- $\varepsilon = \mathrm{Ro}, \delta = \mathrm{Ro}^2\mathrm{Fr}^{-2}$.
- Ro: the Rossby number, ratio of inertial to Coriolis.
- Fr: the Froude number.

The general solution of the SWE is given by a superposition of plane waves

$$\begin{bmatrix}\vec{u}(\vec{x},t)\\ \eta(\vec{x},t)\end{bmatrix} = \sum_{\vec{k}\in\mathbb{Z}^2,\alpha\in\{B,\pm\}} \hat{z}_{\vec{k},\alpha}\exp(i\vec{k}\cdot\vec{x} - i\omega_{\vec{k},\alpha}t)\vec{r}_{\vec{k},\alpha},$$

where the two kinds of modes are:

1. Geostrophically balanced (GB) modes: $\omega_{\vec{k},B} = 0$; incompressible.
2. Gravity modes: $\omega_{\vec{k},\pm} = \pm\epsilon^{-1}\sqrt{\delta|\vec{k}|^2+1}$; compressible.

To describe the turbulent flow, we model the amplitude of each Fourier mode by an O.U. process.

Rotating shallow water models with multiscale features:

$$\begin{bmatrix} \vec{u}(\vec{x},t) \\ \eta(\vec{x},t) \end{bmatrix} = \sum_{\vec{k}\in\mathbf{K},\alpha\in\{B,\pm\}} \hat{v}_{\vec{k},\alpha}(t)\exp(i\vec{k}\cdot\vec{x})\vec{r}_{\vec{k},\alpha},$$

- Slow modes – random incompressible geostrophically balanced (GB) flows.
- Fast modes – random rotating compressible gravity waves.

$$d\hat{v}_{\vec{k},B} = (-d_B\hat{v}_{\vec{k},B} + f_{\vec{k},B}(t))dt + \sigma_{\vec{k},B}dW_{\vec{k},B},$$
$$d\hat{v}_{\vec{k},\pm} = \Big((-d_g + i\omega_{\vec{k},\pm})\hat{v}_{\vec{k},\pm} + f_{\vec{k},\pm}(t)\Big)dt + \sigma_{\vec{k},\pm}dW_{\vec{k},\pm},$$

where $\omega_{\vec{k},\pm} \propto \pm\epsilon^{-1}$ with ϵ being Rossby number.

Highly nonlinear observations mixing GB and gravity modes!

Filter Name	Forecast Model	Observations	
1. Full Filter	Full Model	Full Obs.	Practical but Expensive
2. GB Filter	GB Dynamics	GB Modes	Idealized
3. Reduced Filter I	GB Dynamics	Full Obs.	Practical and Cheap
4. Reduced Filter II (3D-VAR Filter)	GB Dynamics, and Const. Diag. Post. Cov.	Full Obs.	Practical and Cheap

- Rigorous math theory: Comparable high skill in recovering GB modes for all the filters in the geophysical scenario with small Rossby number ϵ.

III. An Efficient Statistically Accurate Algorithm for Solving the Fokker-Planck Equation in Large Dimensions

(Chen & Majda, *JCP*, 2017, *PNAS*, 2017; Chen, Majda & Tong, *SIAM UQ*, 2017)

Consider a general nonlinear dynamical system with noise,

$$d\mathbf{u} = \mathbf{F}(\mathbf{u}, t)dt + \mathbf{\Sigma}(\mathbf{u}, t)d\mathbf{W},$$

the associated Fokker-Planck equation is given by

$$\frac{\partial}{\partial t}p(\mathbf{u}, t) = -\nabla_{\mathbf{u}}\big(\mathbf{F}(\mathbf{u}, t)p(\mathbf{u}, t)\big) + \frac{1}{2}\nabla_{\mathbf{u}}\cdot\nabla_{\mathbf{u}}(\mathbf{\Sigma}\mathbf{\Sigma}^T(\mathbf{u}, t)p(\mathbf{u}, t)), \qquad \text{with } p_t\big|_{t=t_0} = p_0(\mathbf{u}).$$

- Solving the Fokker-Planck equation for both steady state and transient phases is an important topic in science, engineering, finance, and many other areas.
- Typical features of the PDFs in many applications: large dimensions and strong non-Gaussianity (geophysical turbulence, engineering, neuroscience).

no general analytical solution for the Fokker-Planck equation

numerical approaches: finite element, finite difference, direct Monte Carlo simulation

III. An Efficient Statistically Accurate Algorithm for Solving the Fokker-Planck Equation in Large Dimensions

(Chen & Majda, *JCP*, 2017, *PNAS*, 2017; Chen, Majda & Tong, *SIAM UQ*, 2017)

Consider a general nonlinear dynamical system with noise,

$$d\mathbf{u} = \mathbf{F}(\mathbf{u}, t)dt + \mathbf{\Sigma}(\mathbf{u}, t)d\mathbf{W},$$

the associated Fokker-Planck equation is given by

$$\frac{\partial}{\partial t}p(\mathbf{u}, t) = -\nabla_{\mathbf{u}}\big(\mathbf{F}(\mathbf{u}, t)p(\mathbf{u}, t)\big) + \frac{1}{2}\nabla_{\mathbf{u}}\cdot\nabla_{\mathbf{u}}(\mathbf{\Sigma}\mathbf{\Sigma}^T(\mathbf{u}, t)p(\mathbf{u}, t)), \qquad \text{with } p_t\big|_{t=t_0} = p_0(\mathbf{u}).$$

- Solving the Fokker-Planck equation for both steady state and transient phases is an important topic in science, engineering, finance, and many other areas.
- Typical features of the PDFs in many applications: large dimensions and strong non-Gaussianity (geophysical turbulence, engineering, neuroscience).

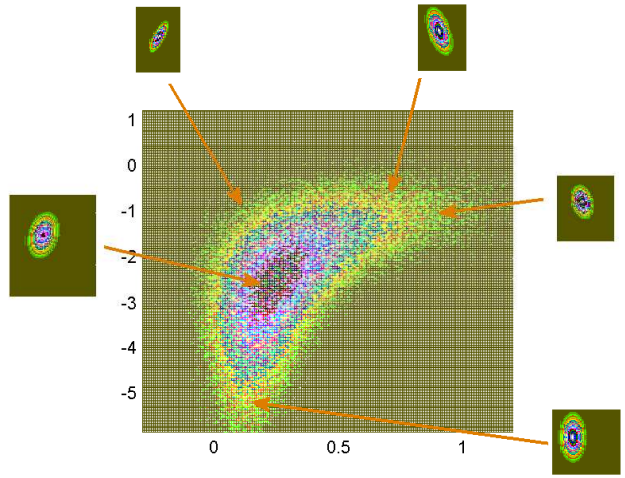
no general analytical solution for the Fokker-Planck equation

numerical approaches: finite element, finite difference, direct Monte Carlo simulation **suffering from curse of dimensionality!**

An efficient statistically accurate algorithm for conditional Gaussian systems.

(Chen & Majda, *JCP*, 2017, *PNAS*, 2017; Chen, Majda & Tong, *SIAM UQ*, 2017)

- Each sample is not a "dot" but a Gaussian distribution that covers a sufficiently large volume.
- Use dynamics to find the optimal location and the optimal volume of each sample.
- Optimization is based on semi-analytic formulae and parallel runs — computationally efficient.
- Rigorous analysis shows that a much smaller number of samples is needed compared with that in the direct MC method.



An efficient statistically accurate algorithms for conditional Gaussian systems.

Assume the dimension of $\mathbf{u_I}$ is low while that of $\mathbf{u_{II}}$ can be large,

$$d\mathbf{u_I} = [\mathbf{A}_0(t,\mathbf{u_I}) + \mathbf{A}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t,\mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t,\mathbf{u_I}) + \mathbf{a}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t,\mathbf{u_I})d\mathbf{W_{II}}(t).$$

- Sample L trajectories of $\mathbf{u_I}$ (by Monte Carlo, for example).
- $p(\mathbf{u_{II}}(t))$ is computed from running L conditional Gaussian filter in a parallel way,

$$p(\mathbf{u_{II}}(t)) = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} p\Big(\mathbf{u_{II}}(t)|\mathbf{u_I^i}(s\le t)\Big),$$

- $p(\mathbf{u_I}(t))$ is computed based on L samples with a Gaussian kernel method,

$$p\big(\mathbf{u_I}(t)\big) = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} K_{\mathbf{H}}\Big(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)\Big),$$

- The joint PDF is given by a Gaussian mixture,

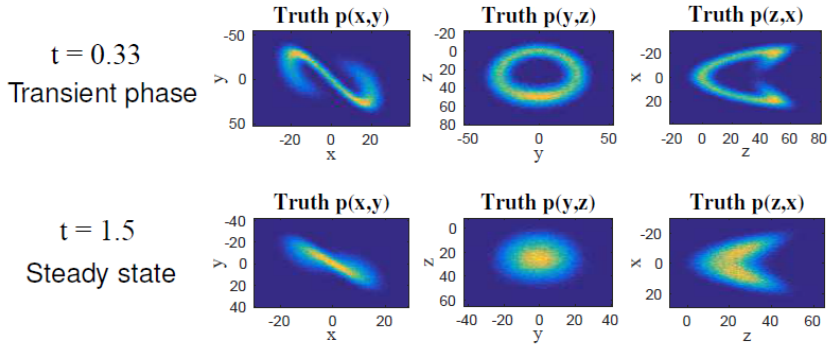
$$p(\mathbf{u_I}(t),\mathbf{u_{II}}(t)) = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} \Big(K_{\mathbf{H}}(\mathbf{u_I}(t) - \mathbf{u_I^i}(t))\cdot p(\mathbf{u_{II}}(t)|\mathbf{u_I^i}(s\le t))\Big),$$

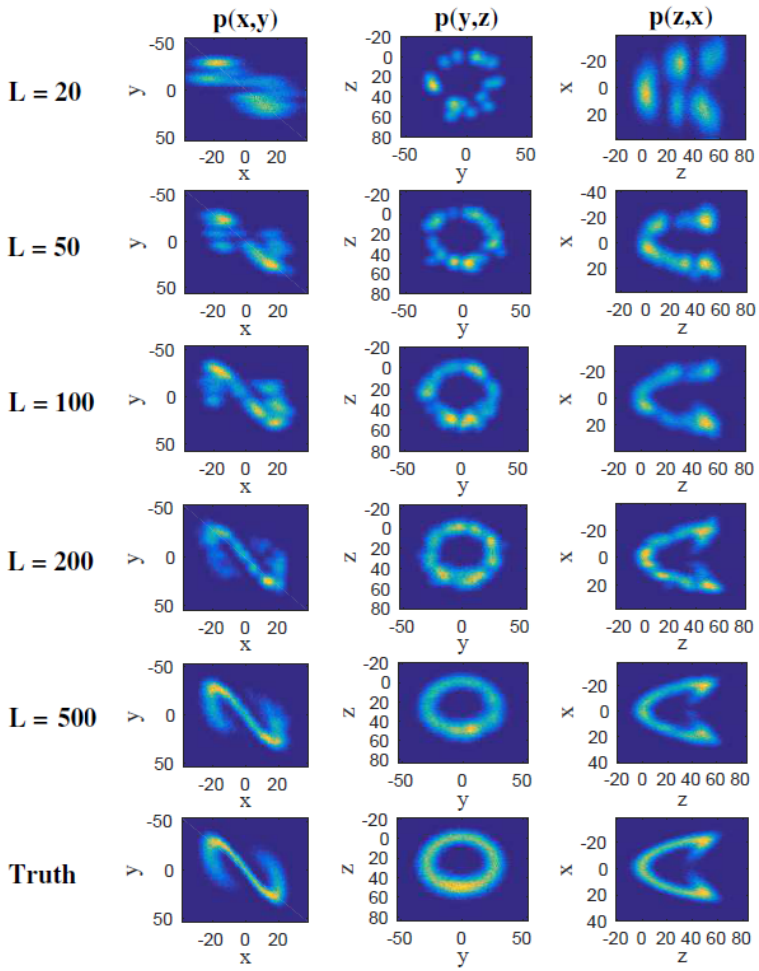
Practically, $L\sim O(100)$ is able to handle systems with $Dim(\mathbf{u_I})\le 3$ and $Dim(\mathbf{u_{II}})\sim O(10)$.

The noisy Lorenz 63 (L-63) Model:

$$dx = \sigma(y - x)dt + \sigma_x dW_x,$$
$$dy = \big(x(\rho - z) - y\big)dt + \sigma_y dW_y,$$
$$dz = (xy - \beta z)dt + \sigma_z dW_z.$$

$\sigma = 10, \rho = 28, \beta = 8/3, \sigma_x = \sigma_y = \sigma_z = 10.$

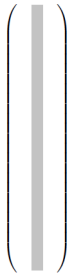




Beating the curse of dimension with block decomposition (Chen & Majda, *PNAS*, 2017).

Mean

$nm \times 1$

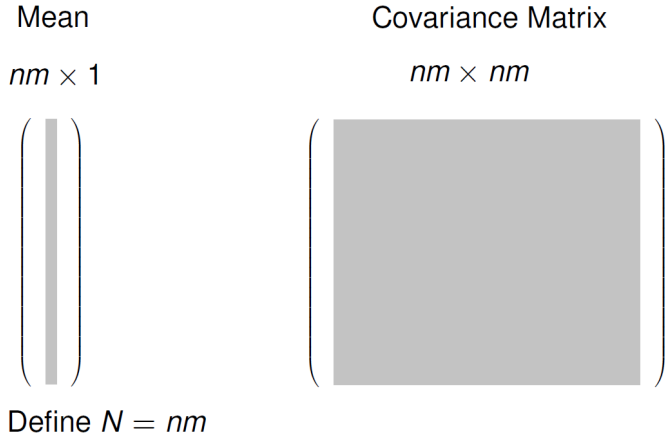


Covariance Matrix

$nm \times nm$

Define $N = nm$

Beating the curse of dimension with block decomposition (Chen & Majda, *PNAS*, 2017).



In many complex dynamical systems with *multiscale structures, multilevel dynamics or state-dependent parameterizations*, the state variables can be decomposed in the following way

$$\mathbf{u}_k = (\mathbf{u}_{\mathbf{I},k}, \mathbf{u}_{\mathbf{II},k}) \quad \text{with} \quad \mathbf{u}_{\mathbf{I},k} \in \mathbb{R}^{N_{\mathbf{I}},k} \quad \text{and} \quad \mathbf{u}_{\mathbf{II},k} \in \mathbb{R}^{N_{\mathbf{II}},k},$$

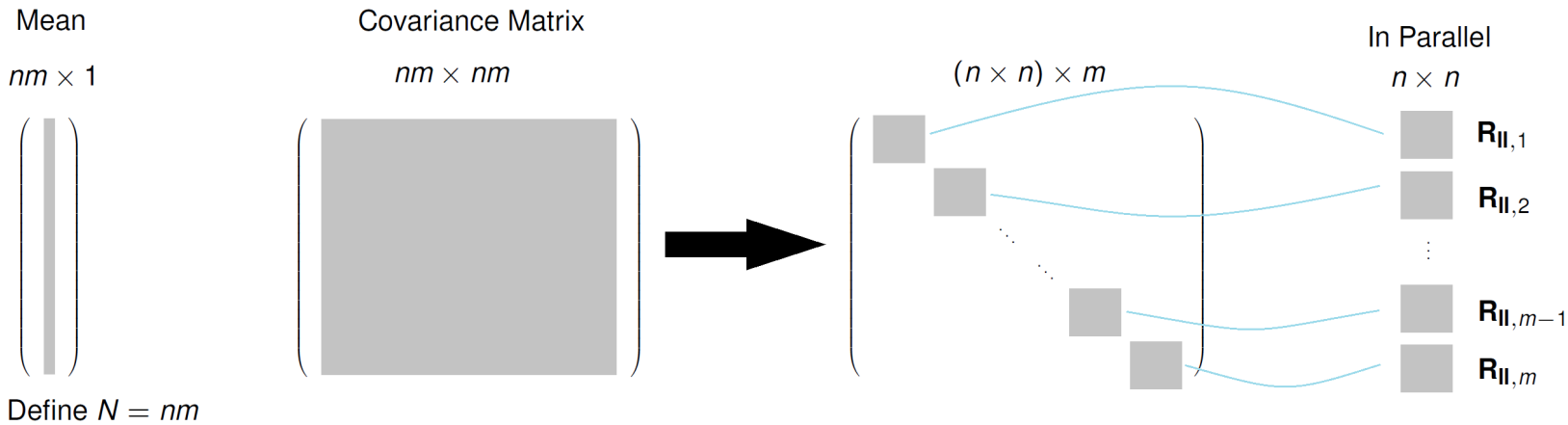
where each $\mathbf{u}_k$ satisfies

$$\begin{aligned} d\mathbf{u}_{\mathbf{I},k} &= [\mathbf{A}_{0,k}(t, \mathbf{u}_{\mathbf{I}}) + \mathbf{A}_{1,k}(t, \mathbf{u}_{\mathbf{I},k})\mathbf{u}_{\mathbf{II},k}]dt + \mathbf{\Sigma}_{\mathbf{I},k}(t, \mathbf{u}_{\mathbf{I},k})d\mathbf{W}_{\mathbf{I},k}(t), \\ d\mathbf{u}_{\mathbf{II},k} &= [\mathbf{a}_{0,k}(t, \mathbf{u}_{\mathbf{I}}) + \mathbf{a}_{1,k}(t, \mathbf{u}_{\mathbf{I},k})\mathbf{u}_{\mathbf{II},k}]dt + \mathbf{\Sigma}_{\mathbf{II},k}(t, \mathbf{u}_{\mathbf{I},k})d\mathbf{W}_{\mathbf{II},k}(t), \end{aligned}$$

and the initial values of $\mathbf{u}_k$ and $\mathbf{u}_{k'}$ with $k \neq k'$ are independent. With such block decomposition,

- The evolution of $\bar{\mathbf{u}}_{\mathbf{II},k}$ is coupled with that of all other $\bar{\mathbf{u}}_{\mathbf{II},k'}$.
- The evolution of $\mathbf{R}_{\mathbf{II},k}$ has no interaction with that of $\mathbf{R}_{\mathbf{II},k'}$ — allowing the algorithm to solve much larger dynamical systems with parallel runs.

Beating the curse of dimension with block decomposition (Chen & Majda, *PNAS*, 2017).



In many complex dynamical systems with *multiscale structures, multilevel dynamics or state-dependent parameterizations*, the state variables can be decomposed in the following way

$$\mathbf{u}_k = (\mathbf{u}_{\mathbf{I},k}, \mathbf{u}_{\mathbf{II},k}) \quad \text{with} \quad \mathbf{u}_{\mathbf{I},k} \in \mathbb{R}^{N_{\mathbf{I}},k} \quad \text{and} \quad \mathbf{u}_{\mathbf{II},k} \in \mathbb{R}^{N_{\mathbf{II}},k},$$

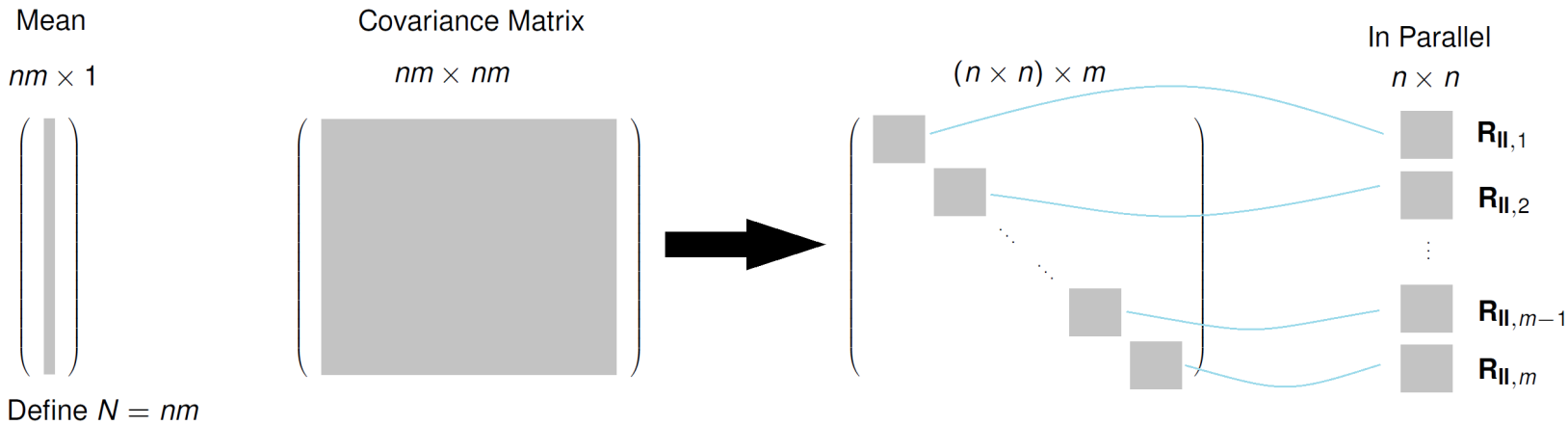
where each $\mathbf{u}_k$ satisfies

$$d\mathbf{u}_{\mathbf{I},k} = [\mathbf{A}_{0,k}(t, \mathbf{u}_{\mathbf{I}}) + \mathbf{A}_{1,k}(t, \mathbf{u}_{\mathbf{I},k})\mathbf{u}_{\mathbf{II},k}]dt + \mathbf{\Sigma}_{\mathbf{I},k}(t, \mathbf{u}_{\mathbf{I},k})d\mathbf{W}_{\mathbf{I},k}(t),$$
$$d\mathbf{u}_{\mathbf{II},k} = [\mathbf{a}_{0,k}(t, \mathbf{u}_{\mathbf{I}}) + \mathbf{a}_{1,k}(t, \mathbf{u}_{\mathbf{I},k})\mathbf{u}_{\mathbf{II},k}]dt + \mathbf{\Sigma}_{\mathbf{II},k}(t, \mathbf{u}_{\mathbf{I},k})d\mathbf{W}_{\mathbf{II},k}(t),$$

and the initial values of $\mathbf{u}_k$ and $\mathbf{u}_{k'}$ with $k \neq k'$ are independent. With such block decomposition,

- The evolution of $\bar{\mathbf{u}}_{\mathbf{II},k}$ is coupled with that of all other $\bar{\mathbf{u}}_{\mathbf{II},k'}$.
- The evolution of $\mathbf{R}_{\mathbf{II},k}$ has no interaction with that of $\mathbf{R}_{\mathbf{II},k'}$ — allowing the algorithm to solve much larger dynamical systems with parallel runs.

Beating the curse of dimension with block decomposition (Chen & Majda, *PNAS*, 2017).



In many complex dynamical systems with *multiscale structures, multilevel dynamics or state-dependent parameterizations*, the state variables can be decomposed in the following way

$$\mathbf{u}_k = (\mathbf{u}_{\mathbf{I},k}, \mathbf{u}_{\mathbf{II},k}) \quad \text{with} \quad \mathbf{u}_{\mathbf{I},k} \in \mathbb{R}^{N_{\mathbf{I}},k} \quad \text{and} \quad \mathbf{u}_{\mathbf{II},k} \in \mathbb{R}^{N_{\mathbf{II}},k},$$

where each $\mathbf{u}_k$ satisfies

$$d\mathbf{u}_{\mathbf{I},k} = [\mathbf{A}_{0,k}(t, \mathbf{u}_{\mathbf{I}}) + \mathbf{A}_{1,k}(t, \mathbf{u}_{\mathbf{I},k})\mathbf{u}_{\mathbf{II},k}]dt + \mathbf{\Sigma}_{\mathbf{I},k}(t, \mathbf{u}_{\mathbf{I},k})d\mathbf{W}_{\mathbf{I},k}(t),$$
$$d\mathbf{u}_{\mathbf{II},k} = [\mathbf{a}_{0,k}(t, \mathbf{u}_{\mathbf{I}}) + \mathbf{a}_{1,k}(t, \mathbf{u}_{\mathbf{I},k})\mathbf{u}_{\mathbf{II},k}]dt + \mathbf{\Sigma}_{\mathbf{II},k}(t, \mathbf{u}_{\mathbf{I},k})d\mathbf{W}_{\mathbf{II},k}(t),$$

and the initial values of $\mathbf{u}_k$ and $\mathbf{u}_{k'}$ with $k \neq k'$ are independent. With such block decomposition,

- The evolution of $\bar{\mathbf{u}}_{\mathbf{II},k}$ is coupled with that of all other $\bar{\mathbf{u}}_{\mathbf{II},k'}$.
- The evolution of $\mathbf{R}_{\mathbf{II},k}$ has no interaction with that of $\mathbf{R}_{\mathbf{II},k'}$ — allowing the algorithm to solve much larger dynamical systems with parallel runs.

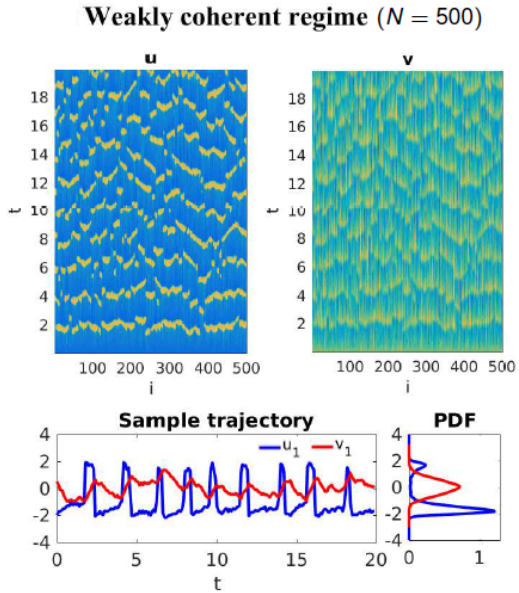
Example 1: Two-layer Lorenz 96 models (Wilks, 2005; Arnold, Moroz & Palmer, 2013)
Example 2: Stochastic coupled FitzHugh-Nagumo models (Lindner et al., 2004)

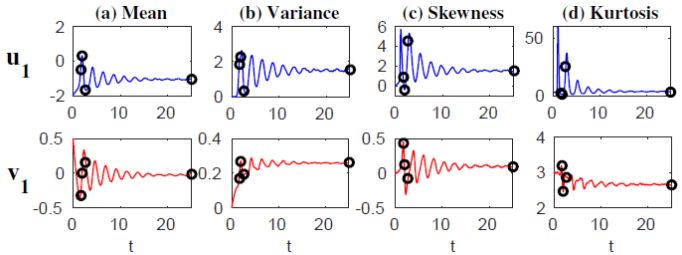
A stochastic coupled FitzHugh-Nagumo (FHN) model.

(Lindner et al., 2004; Muratov, Vanden-Eijnden & E, 2007)

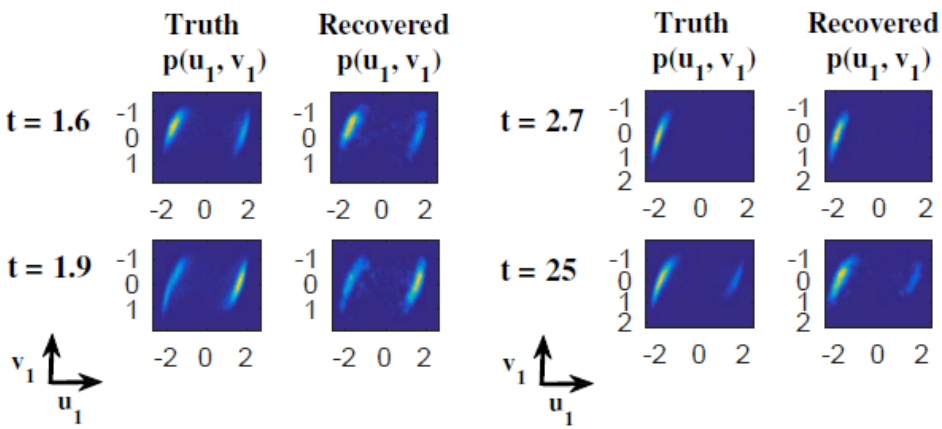
$$\epsilon \frac{du_i}{dt} = u_i - \frac{1}{3}u_i^3 + d_u(u_{i+1} + u_{i-1} - 2u_i) - v_i + \sqrt{\epsilon}\delta_1 \dot{W}_{u_i},$$

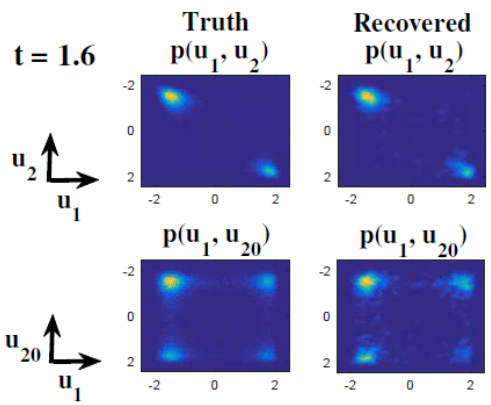
$$\frac{dv_i}{dt} = u_i + a + \delta_2 \dot{W}_{v_i}, \qquad i = 1, \ldots, N. \qquad N = 500 \longrightarrow$$





- $\epsilon = 0.01 \ll 1$: a slow-fast structure of the model.
- $a = 1.05 > 1$, $\delta_1 = 0.2$, $\delta_2 = 0.4$, $d_u = 0.5$: Random noise drives the system above the threshold level of global stability and triggers limit cycles intermittently.
- $u_i(0) = -2$, $v_i(0) = 0.5$ for all i. The model satisfies statistical symmetry.





Due to the statistical symmetric, only $L = 1$ samples is needed here!

Other Applications of Conditional Gaussian Systems

$$d\mathbf{u_I} = [\mathbf{A}_0(t,\mathbf{u_I}) + \mathbf{A}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t,\mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t,\mathbf{u_I}) + \mathbf{a}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t,\mathbf{u_I})d\mathbf{W_{II}}(t).$$

1. parameter estimation and improving stochastic parameterization.
2. understanding and predicting rare and extreme events.
3. exploring the causality between different processes using information theory (causality v.s. correlation).
4. data assimilation and prediction of spatial-extended systems

Example 1: Boussinesq equation.

$$\nabla \cdot \mathbf{u} = 0,$$
$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u} = -\frac{1}{\rho_0}\nabla p + \nu\nabla^2\mathbf{u} - g\alpha T,$$
$$\frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T = \kappa\nabla^2 T + F.$$

– Observe velocity $\mathbf{u}$ + noise.

– Recover temperature T.

Example 2: Stochastic skeleton model for the MJO.

(Thual, Majda & Stechmann, 2014)

$$u_t - yv - \theta_x = 0,$$
$$yu - \theta_y = 0,$$
$$\theta_t - u_x - v_y = \bar{H}a - s^\theta,$$
$$q_t + \bar{Q}(u_x + v_y) = -\bar{H}a + s^q,$$
$$a_t = \Gamma qa.$$

– Observe wave activity $\mathbf{a}$ + noise.

– Recover temperature q and velocity u, v.

(Chen and Majda, *Monthly Weather Review*, 2015)

Other Applications of Conditional Gaussian Systems

$$d\mathbf{u_I} = [\mathbf{A}_0(t,\mathbf{u_I}) + \mathbf{A}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t,\mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t,\mathbf{u_I}) + \mathbf{a}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t,\mathbf{u_I})d\mathbf{W_{II}}(t).$$

1. parameter estimation and improving stochastic parameterization.
2. understanding and predicting rare and extreme events.
3. exploring the causality between different processes using information theory (causality v.s. correlation).
4. data assimilation and prediction of spatial-extended systems

Example 3: A simple dynamical model for the El Niño (Chen & Majda, *PNAS*, 2017).

Atmosphere	$-yv - \partial_x\theta = 0$
	$yu - \partial_y\theta = 0$
	$-(\partial_x u + \partial_y v) = E_q/(1-\overline{Q})$
Ocean	$\partial_\tau U - c_1 Y V + c_1\partial_x H = c_1\tau_x$
	$YU + \partial_Y H = 0$
	$\partial_\tau H + c_1(\partial_x U + \partial_Y V) = 0$
SST	$\partial_\tau T + \mu\partial_x(UT) = -c_1\zeta E_q + c_1\eta H$

Latent heating:

$$E_q = \alpha_q T.$$

Wind stress:

$$\tau_x = \gamma(u + u_p),$$

Wind Bursts & easterly mean trade wind:

$$u_p = a_p(\tau)s_p(x)\phi_0(y),$$
$$\frac{da_p}{d\tau} = -d_p(a_p - \hat{a}_p) + \sigma_p(T_W)\dot{W}(\tau).$$

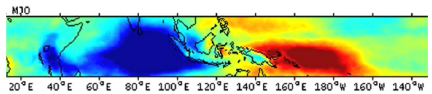
– Observe sea surface temperature (SST) T and wind burst noise a_p.

– Recover $u, U, H, \theta\ldots$.

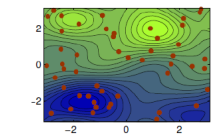
Summary

A conditional Gaussian framework for data assimilation and prediction is introduced. Despite the conditional Gaussianity, the system remains highly nonlinear and is able to capture the non-Gaussian features as observed in nature.

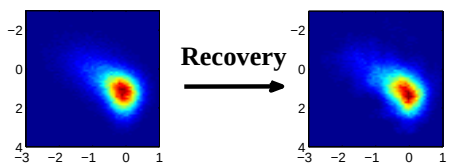
- Predicting the large-scale MJO via a physics-constrained low-order nonlinear stochastic model.



- Understanding the information barrier and data assimilation skill of recovering ocean flows with noisy Lagrangian tracers.



- An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions with strongly non-Gaussian features.



- Other applications: parameter estimation, spatial extended physical systems ...

Thank you

(nan.chen@nyu.edu)

Reserve Slides

Appendix 1: More details of parameter estimation.

1. Estimating one additive parameter γ^* in a linear scalar model,

$$du = (A_0 u + A_1 \gamma^*)dt + \sigma_u dW_u.$$

	Direct approach	Stochastic parameterized equations
Convergence rate	algebraic	exponential
Error as $t \to \infty$	zero	usually non-zero
$\sigma_u \downarrow$	convergence rate $\uparrow$	convergence rate $\uparrow$

2. Estimating one multiplicative parameter γ^* in a linear scalar model,

$$du = (A_0 - \gamma^* u)dt + \sigma_u dW_u,$$

	Direct approach	Stochastic parameterized equations
$\sigma_u \downarrow$ with $A_0 \neq 0$	convergence rate $\uparrow$	convergence rate $\uparrow$
$\sigma_u \downarrow$ with $A_0 = 0$	independent of σ_u	convergence rate $\uparrow$

3. Estimating one multiplicative parameter γ^* in a cubic nonlinear scalar model,

$$du = (A_0 - \gamma^* u^3)dt + \sigma_u dW_u,$$

	Direct approach	Stochastic parameterized equations
$\sigma_u \downarrow$	convergence rate $\downarrow$	convergence rate $\uparrow$

4. Estimating four different parameters a^*, b^*, c^* and f^* in a cubic nonlinear scalar model,

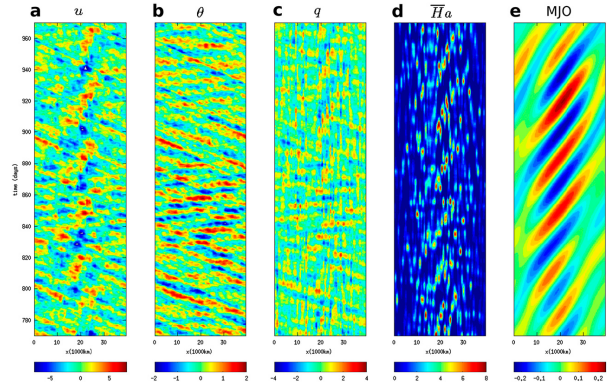
$$du = (a^* u + b^* u^2 - c^* u^3 + f^*)dt + \sigma_u dW_u,$$

	Direct approach	Stochastic parameterized equations
$\sigma_u \downarrow$	may not converge to the truth	convergence rate $\uparrow$

Appendix 2: Data assimilation and prediction of spatial extended turbulent systems.

a. Stochastic skeleton model for the MJO (Majda & Stechmann, PNAS 2009; Thual, M & S, JAS 2014)

$$
\begin{aligned}
u_t - yv - \theta_x &= 0,\\
yu - \theta_y &= 0,\\
\theta_t - u_x - v_y &= \bar{H}a - s^{\theta},\\
q_t + \bar{Q}(u_x + v_y) &= -\bar{H}a + s^{q},\\
a &= \text{stochastic birth-death process},
\end{aligned}
$$



The expectation of convective activity a satisfies $a_t = \Gamma q a$.

b. Meridional (y direction) truncation + characteristic form. $u, \theta \Longleftrightarrow K, R$.

$$K_t + K_x = (S^{\theta} - \bar{H}A)/2, \qquad R_t - R_x/3 = (S^{\theta} - \bar{H}A)/3.$$

c. Design nonlinear filter with judicious model error

Observed: $$\frac{d\hat{A}_k}{dt} = \Gamma \sum_{-M+1 \le s \le M} \hat{Q}_s \hat{A}_{k-s} + \sigma_k^A \dot{W}_k^A,$$

Unobserved: $$\frac{d\hat{K}_k}{dt} = (-il_k - \bar{d}_k^K)\hat{K}_k + \frac{1}{2}\left(\hat{S}_k^{\theta} - \bar{H}\hat{A}_k\right) + \sigma_k^K \dot{W}_k^K,$$

$$\frac{d\hat{R}_k}{dt} = \ldots, \qquad \frac{d\hat{Q}_k}{dt} =$$

[Conditional Gaussian system!]

Recover the initial values of K, R and Q and run the dynamical model for prediction.

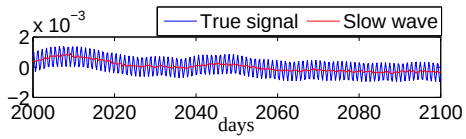
(Chen & Majda, Monthly Weather Review 2015)

d. Further applying an effectively reduced filter for small-scale waves ($k \gg 1$).

$$\frac{d\hat{K}_k}{dt} = \underbrace{-il_k\hat{K}_k}_{\text{fast inertial oscillation}} + \underbrace{\frac{1}{2}\left(\hat{S}_k^{\theta} - \bar{H}\hat{A}_k\right)}_{\text{slow external forcing}}.$$

Average out the fast oscillations,

$$\tilde{\hat{K}}_k = \frac{\hat{S}_k^{\theta} - \bar{H}\hat{A}_k}{2il_k}.$$



Appendix 3: Derivations of the efficient statistically accurate algorithm.

First, the joint distribution of $\mathbf{u_I}$ and $\mathbf{u_{II}}$ at time t can be written as

$$p\Big(\mathbf{u_I}(t), \mathbf{u_{II}}(t)\Big) = \int p\Big(\mathbf{u_{II}}(t), \mathbf{u_I}(t) \,|\, \mathbf{u_I}(s \le t)\Big)\, p\Big(\mathbf{u_I}(s \le t)\Big)\, d\mathbf{u_I}(s \le t) \tag{3}$$

Here, according to the basic probability relationship $p(x, y|z) = p(x|y, z)\, p(y|z)$, we have the following

$$p\Big(\mathbf{u_{II}}(t), \mathbf{u_I}(t) \,|\, \mathbf{u_I}(s \le t)\Big) = p\Big(\mathbf{u_{II}}(t) \,|\, \mathbf{u_I}(s \le t)\Big)\, p\Big(\mathbf{u_I}(t) \,|\, \mathbf{u_I}(s \le t)\Big). \tag{4}$$

The second term on the right hand side of (4) is actually a delta function peaking at the conditioned value of $\mathbf{u_I}$ at time t. In fact, if we replace the condition inside the PDF $\mathbf{u_I}(s \le t)$ by $\mathbf{u_I^i}(s \le t)$, we have

$$p\Big(\mathbf{u_I}(t) \,|\, \mathbf{u_I^i}(s \le t)\Big) = \delta(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)) \tag{5}$$

In addition,

$$p(\mathbf{u_I}(s \le t)) = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \delta\left(\mathbf{u_I}(s \le t) - \mathbf{u_I^i}(s \le t)\right). \tag{6}$$

Therefore, inserting (4)–(6) into (3) yields

$$\begin{aligned} p\Big(\mathbf{u_I}(t), \mathbf{u_{II}}(t)\Big) &= \int p\Big(\mathbf{u_{II}}(t), \mathbf{u_I}(t) \,|\, \mathbf{u_I}(s \le t)\Big)\, p\Big(\mathbf{u_I}(s \le t)\Big)\, d\mathbf{u_I}(s \le t) \\ &= \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \delta\Big(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)\Big)\, p\Big(\mathbf{u_{II}}(t) \,|\, \mathbf{u_I^i}(s \le t)\Big) \end{aligned} \tag{7}$$

Next, we make use of the kernel approximation $K_{\mathbf{H}}(\mathbf{u_I}(t) - \mathbf{u_I^i}(t))$ for $\delta\Big(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)\Big)$. Note that in the limit $L \to \infty$ the bandwidth goes to zero and the kernel approximation converges to $\delta\Big(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)\Big)$, which leads to (7) that is consistent with solving the Fokker-Planck equation for the joint PDF.

Appendix 4: Rigorous analysis of the error in the efficient statistically accurate algorithms.

Kernel density estimation for the joint PDF.

$$\tilde{p}_t(\mathbf{u_I}, \mathbf{u_{II}}) = \frac{1}{L}\sum_{i=1}^{L} K_H((\mathbf{u_I}, \mathbf{u_{II}}) - (\mathbf{u_I}^i(t), \mathbf{u_{II}}^i(t))),$$

$$\text{with } K_H(\mathbf{u_I}, \mathbf{u_{II}}) = (2\pi H)^{-\frac{N_I+N_{II}}{2}} \exp\left(-\frac{1}{2H}\sum_{i=1}^{N_I} c_i^2 \mathbf{u}_{\mathbf{I},i}^2 - \frac{1}{2H}\sum_{i=1}^{N_{II}} c_{i+N_I}^2 \mathbf{u}_{\mathbf{II},i}^2\right).$$

Hybrid method — kernel density estimation for $\mathbf{u_I}$ *and conditional Gaussian mixture for* $\mathbf{u_{II}}$.

$$\hat{p}_t(\mathbf{u_I}, \mathbf{u_{II}}) = \frac{1}{L}\sum_{i=1}^{L} K_H(\mathbf{u_I} - \mathbf{u_I}^i(t)) p(\mathbf{u_{II}}|\mathbf{u_I}^i(s \le t)).$$

$$\text{with } K_H(\mathbf{u_I}) = (2\pi H)^{-\frac{N_I}{2}} \exp\left(-\frac{1}{2H}\sum_{i=1}^{N_I} c_i^2 \mathbf{u}_{\mathbf{I},i}^2\right).$$

The mean integrated squared error (MISE) (of the hybrid method) is the average L^2 distance to the true density:

$$\begin{aligned}\text{MISE} &= \mathbb{E}\int |p_t(\mathbf{u_I}, \mathbf{u_{II}}) - \hat{p}_t(\mathbf{u_I}, \mathbf{u_{II}})|^2 d\mathbf{u_I} d\mathbf{u_{II}} \\ &= \underbrace{\mathbb{E}\int |\hat{p}_t(\mathbf{u_I}, \mathbf{u_{II}}) - \bar{p}_t(\mathbf{u_I}, \mathbf{u_{II}})|^2 d\mathbf{u_I} d\mathbf{u_{II}}}_{\text{Bias}} + \underbrace{\int |p_t(\mathbf{u_I}, \mathbf{u_{II}}) - \bar{p}_t(\mathbf{u_I}, \mathbf{u_{II}})|^2 d\mathbf{u_I} d\mathbf{u_{II}}}_{\text{Variance}}\end{aligned}$$

Theorem (C., Majda, Tong): error estimation of the hybrid method.

The two parts of MISE for the hybrid method are bounded:

$$\begin{aligned}
\hat{p}_t \text{ Variance} &\leq \frac{1}{L}\mathbb{E}\left(\prod_{i=1}^{N_{\mathbf{I}}}(\pi H c_i^2)\det(\pi \mathbf{R}_{\mathbf{II}}(t))\right)^{-\frac{1}{2}}, \\
\hat{p}_t \text{ Bias} &\leq \frac{1+\delta}{4}H^2 J\left(\sum_{i=1}^{N_{\mathbf{I}}} c_i^2 \partial^2_{\mathbf{u}_{\mathbf{I}},i}\, p_t(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})\right) + \frac{1+\delta^{-1}}{2}M^2 H^3\left(\sum_{i=1}^{N_{\mathbf{I}}} c_i^2\right)^3 J(M(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})),
\end{aligned} \tag{8}$$

where $J(f(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}}))$ denotes the integral $\int f^2(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}}) d\mathbf{u}_{\mathbf{I}} d\mathbf{u}_{\mathbf{II}}$. The function $M(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})$ is an upper bound of the third order directional derivative of p_t in the direction of $\mathbf{u}_{\mathbf{I}}$ around $(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})$.

By taking δ close to zero and ignoring the higher order term in the bias upper bound, we recover an upper bound similar to the asymptotic MISE (AMISE) , except that our method also consists a random component of $\mathbf{R}_{\mathbf{II}}(t)$:

$$\text{AMISE} \leq \frac{1}{L}\mathbb{E}\left(\prod_{i=1}^{N_{\mathbf{I}}}(\pi H c_i^2)\det(\pi \mathbf{R}_{\mathbf{II}}(t))\right)^{-\frac{1}{2}} + \frac{1}{4}H^2 J\left(\sum_{i=1}^{N_{\mathbf{I}}} c_i^2 \partial^2_{\mathbf{u}_{\mathbf{I}},i}\, p_t(\mathbf{u}_{\mathbf{I}}, \mathbf{u}_{\mathbf{II}})\right),$$

which gives

$$H \sim O\left(L^{-\frac{2}{4+N_{\mathbf{I}}}}\right) \quad \text{and consequentially} \quad \text{MISE} \sim O\left(L^{-\frac{4}{4+N_{\mathbf{I}}}}\right).$$

MISE has no dependence on $\mathbf{N}_{\mathbf{II}}$ — the dimensional of $\mathbf{u}_{\mathbf{II}}$.

$\tilde{p}_t$: Kernel density estimation for the joint PDF.

$\hat{p}_t$: Hybrid method.

Theorem (C., Majda, Tong): Comparison of the two methods.

The error in the bias

$$\tilde{p}_t \text{ Bias bound} \geq \hat{p}_t \text{ Bias bound},$$

and in the variance

$$\frac{\tilde{p}_t \text{ Variance bound}}{\hat{p}_t \text{ Variance bound}} = \frac{H^{-\frac{N_{\mathbf{II}}}{2}} \prod_{i=1}^{N_{\mathbf{II}}} c_{i+N_{\mathbf{I}}}}{\mathbb{E}\sqrt{\det(\mathbf{R}_{\mathbf{II}}(t))}^{-1}}.$$

We have:

$$\tilde{p}_t : \quad \text{MISE} \sim O\left(L^{-\frac{4}{4+N_{\mathbf{I}}+N_{\mathbf{II}}}}\right)$$

$$\hat{p}_t : \quad \text{MISE} \sim O\left(L^{-\frac{4}{4+N_{\mathbf{I}}}}\right)$$

If one wants the performance of the direct kernel method to be the same as the hybrid method, then the sample size needs to increase to

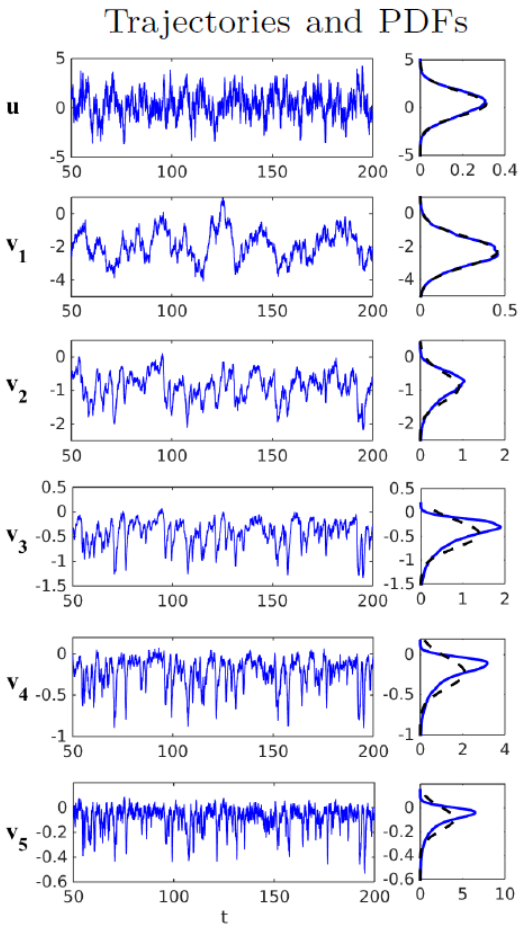
$$\widetilde{L} = L^{\frac{4+N_{\mathbf{I}}+N_{\mathbf{II}}}{4+N_{\mathbf{I}}}},$$

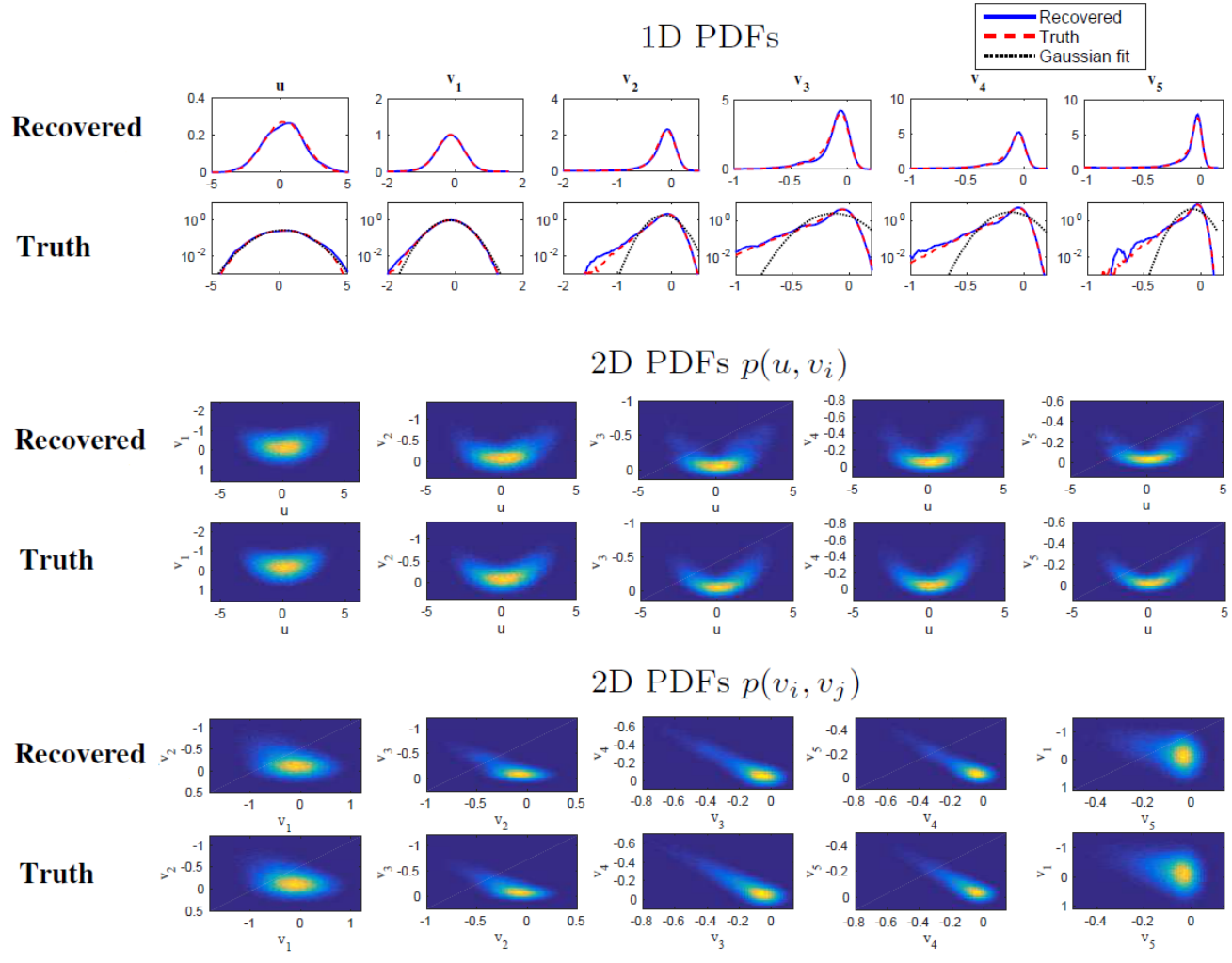
which can be many magnitudes larger than L.

Recovery of the PDFs of a 6D conceptual dynamical model for turbulence:

$$du = (-d_u u + F_u + \sum_i^5 \gamma_i\, u\, v_i)dt + \sigma_u dW_u,$$

$$dv_i = (-d_{v_i} v_i - \gamma_i\, u^2)dt + \sigma_{v_i} dW_{v_i}, \qquad i = 1, \ldots, 5.$$





Appendix 5: Data assimilation of ocean flows using Lagrangian tracers
More realistic scenario — nonlinear coupling of GB and gravity modes:

$$d\hat{v}_{\vec{k},B} = (-d_B \hat{v}_{\vec{k},B} + f_{\vec{k},B}(t))dt + \sigma_{\vec{k},B} dW_{\vec{k},B}(t),$$

$$d\hat{v}_{\vec{k},\pm} = \Big((-d_g + i\omega_{\vec{k},\pm} \underline{+ i\hat{v}_{\vec{k},B}})\hat{v}_{\vec{k},\pm} + f_{\vec{k},\pm}(t) \Big) dt + \sigma_{\vec{k},\pm} dW_{\vec{k},\pm}(t).$$

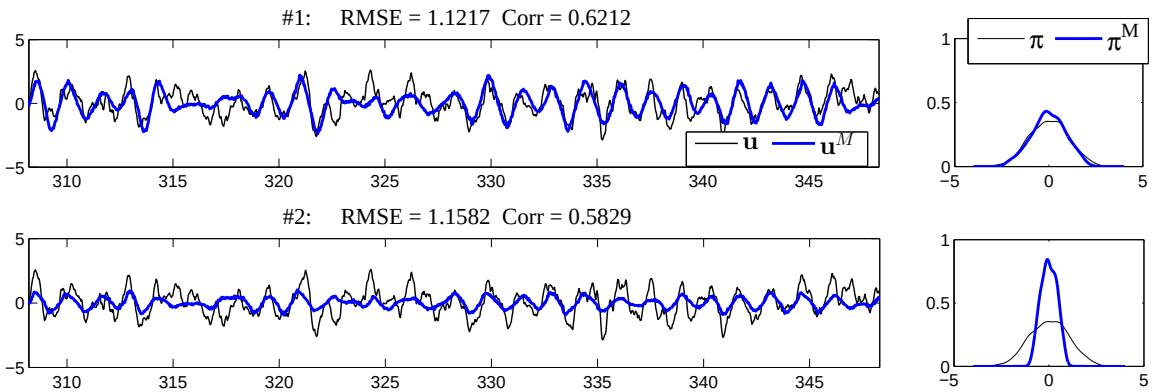
Linear models without $\underline{i\hat{v}_{\vec{k},B}}$ are used as imperfect forecast models such that the corresponding filters belong to the conditional Gaussian framework.

- Assessing model error for approximate filters through information theory.

Combination of three information measures (Chen & Majda, 2015; Branicki & Majda, 2014).

1. Shannon entropy of residual $\sim$ root-mean-square error.
2. Mutual information $\sim$ pattern correlation.
3. Relative entropy: an indicator of assessing the disparity in the amplitudes and peaks — important in quantifying extreme events!

$$\mathcal{P}(\pi, \pi^M) = \int \pi(\mathbf{u}) \ln \frac{\pi(\mathbf{u})}{\pi^M(\mathbf{u})} d\mathbf{u}$$



Appendix 6: Nonlinear Laplacian Spectrum Analysis (NLSA).

- NLSA is a nonlinear data analysis technique that combines ideas from lagged embedding (Packard et al. 1980; Sauer et al. 1991), machine learning (Coifman and Lafon 2006; Belkin and Niyogi 2003), adaptive weights and spectral entropy criteria to extract spatiotemporal modes of variability from high-dimensional time series.

- These modes are computed utilizing the eigenfunctions of a discrete analog of Laplace-Beltrami operator, which can be thought of as a local analog of the temporal covariance matrix employed in EOF and EEOF techniques, but adapted to the nonlinear geometry of data generated by complex dynamical systems.

- NLSA by design requires no ad hoc pre-processing of data such as detrending or spatiotemporal filtering of the full data set and it captures both intermittency and low frequency variability.

- The NLSA modes have higher memory and predictability compared with those extracted via EEOF analysis.

Procedure:

1. construct a time lagged embedding dataset utilizing Takens' method of delay (Takens et al. 1981). Denote q the lagged embedding window size. Then the lagged embedding matrix can be written as

$$X = \begin{pmatrix} z_1 & z_2 & \cdots & z_{n-q+1} \\ z_2 & z_3 & \cdots & z_{n-q+2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{q-1} & z_q & \cdots & z_{n-1} \\ z_q & z_{q+1} & \cdots & z_n \end{pmatrix}.$$

2. Compute the kernel matrix K with entries $K_{ij} = K(X(t_i), X(t_j))$ given by

$$K(X(t_i), X(t_j)) = \exp\left(-\frac{\|X(t_i) - X(t_j)\|^2}{\epsilon \xi(t_i)\xi(t_j)}\right),$$

where $\xi(t_i) = \|X(t_i) - X(t_{i-1})\|$ and $X(t_i) = (z_i, \ldots, z_{i+q-1})^T$.
The kernel matrix K can be thought as a nonlinear analogy of the temporal covariance matrix in the singular spectrum analysis (Ghil et al. 2002), while this nonlinearity is crucial in capturing both intermittency and low-frequent variability.

3. The NLSA temporal patterns $\phi(t_i)$ are then determined by the eigenvectors of the Laplacian matrix $L = I - P$,

$$L\phi_k = \lambda_k \phi_k, \qquad \phi_k = (\phi_{1k}, \phi_{2k}, \ldots \phi_{Sk})^T,$$

where

$$q_i = \sum_{j=1}^{S} K_{ij}, \qquad K'_{ij} = \frac{K_{ij}}{q_i q_j}, \qquad d_i = \sum_{j=1}^{S} K ij', \qquad P_{ij} = \frac{K_{ij}}{d_i}, \qquad S = n - q.$$