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

Nan Chen

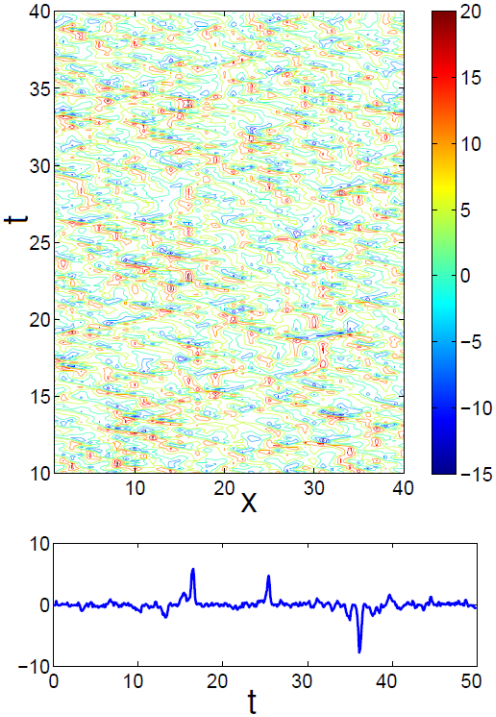
Center for Atmosphere Ocean Science (CAOS)
Courant Institute of Mathematical Sciences
New York University

SIAM Conference on Uncertainty Quantification
April 16–19, 2018
Hyatt Regency Orange County
Garden Grove, California, USA

Introduction

Turbulent dynamical systems

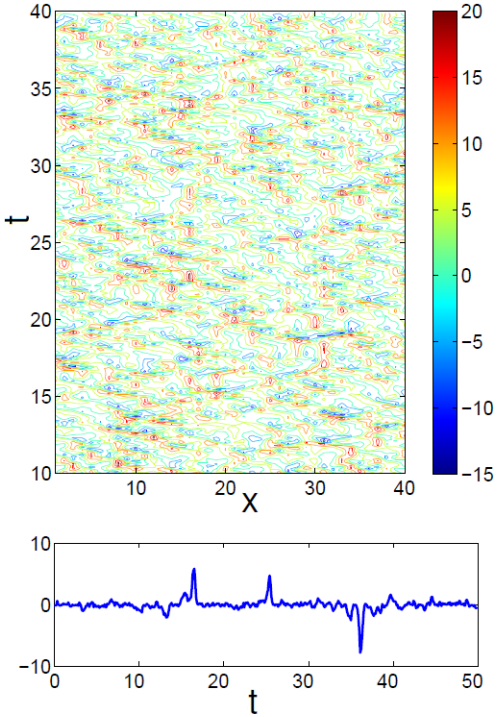
- ubiquitous in geoscience, engineering, neural and material sciences
- characterized by a large dimensional phase space and a large dimensional space of strong instabilities, which transfer energy throughout the system



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 from noisy observations — data assimilation/filtering
- effective predictions with improved initial values from 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 (Chen & Majda, 2016),

$$d\mathbf{u_I} = [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t) \tag{1a}$$

$$d\mathbf{u_{II}} = [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t) \tag{1b}$$

Once $\mathbf{u_I}(s)$ for $s \leq t$ is given, $\mathbf{u_{II}}(t)$ conditioned on $\mathbf{u_I}(s)$ becomes a Gaussian process,

$$p(\mathbf{u_{II}}(t)|\mathbf{u_I}(s \leq t)) \sim \mathcal{N}(\bar{\mathbf{u}}_{\mathbf{II}}(t), \mathbf{R_{II}}(t)). \tag{2}$$

Conditional Gaussian Nonlinear Systems

Many turbulent dynamical systems belong to conditional Gaussian framework.

The conditional Gaussian systems have the following abstract form (Chen & Majda, 2016),

$$d\mathbf{u_I} = [\mathbf{A}_0(t,\mathbf{u_I}) + \mathbf{A}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t,\mathbf{u_I})d\mathbf{W_I}(t) \tag{1a}$$

$$d\mathbf{u_{II}} = [\mathbf{a}_0(t,\mathbf{u_I}) + \mathbf{a}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t,\mathbf{u_I})d\mathbf{W_{II}}(t) \tag{1b}$$

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

$$p(\mathbf{u_{II}}(t)|\mathbf{u_I}(s \le t)) \sim \mathcal{N}(\bar{\mathbf{u}}_{\mathbf{II}}(t), \mathbf{R_{II}}(t)). \tag{2}$$

- 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 Gaussian distribution in (2) has closed analytic form (Liptser & Shiryaev 2001):

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

$$d\mathbf{R_{II}}(t) = \left\{\mathbf{a}_1(t,\mathbf{u_I})\mathbf{R_{II}} + \mathbf{R_{II}}\mathbf{a}_1^*(t,\mathbf{u_I}) + (\mathbf{\Sigma_{II}}\mathbf{\Sigma_{II}}^*)(t,\mathbf{u_I}) - (\mathbf{R_{II}}\mathbf{A}_1^*(t,\mathbf{u_I}))(\mathbf{\Sigma_I}\mathbf{\Sigma_I}^*)^{-1}(t,\mathbf{u_I})(\mathbf{R_{II}}\mathbf{A}_1^*(t,\mathbf{u_I}))^*\right\}dt.$$

Examples of conditional Gaussian systems.

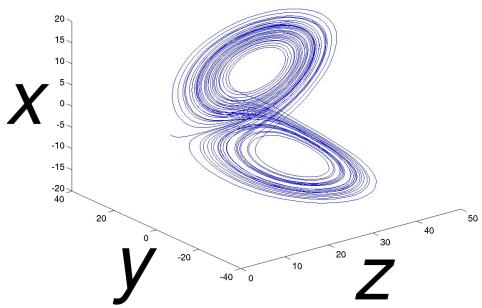
$$d\mathbf{u_I} = [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),$$

Noisy Lorenz 63 model

— A simple chaotic system

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

$\rho = 28$
$\sigma = 10$
$\beta = 8/3$



Examples of conditional Gaussian systems.

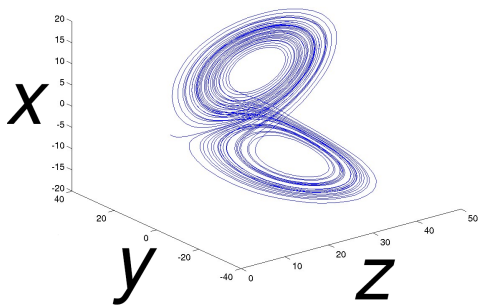
$$d\mathbf{u_I} = [\mathbf{A}_0(t, \mathbf{u_I}) + \mathbf{A}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t, \mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t, \mathbf{u_I}) + \mathbf{a}_1(t, \mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t, \mathbf{u_I})d\mathbf{W_{II}}(t),$$

Noisy Lorenz 63 model

— A simple chaotic system

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

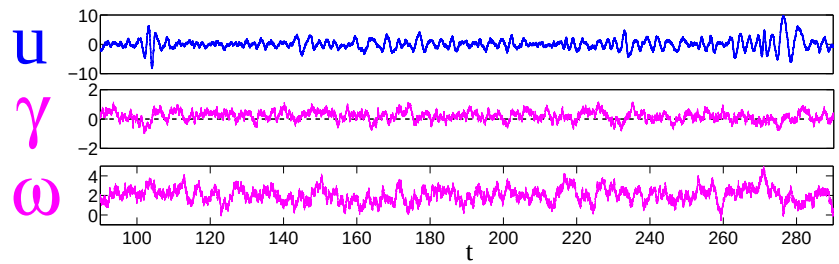
$\rho = 28$
$\sigma = 10$
$\beta = 8/3$



Stochastic parameterizations

— Capturing intermittency and random phases

$$du = (-\gamma + i\omega)udt + \sigma_u dW_u,$$
$$d\gamma = -d_\gamma(\gamma - \hat{\gamma})dt + \sigma_\gamma dW_\gamma$$
$$d\omega = -d_\omega(\omega - \hat{\omega})dt + \sigma_\omega dW_\omega$$



Examples of conditional Gaussian systems.

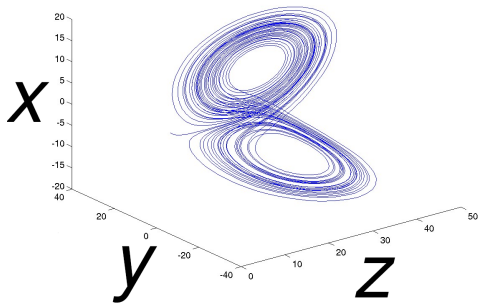
$$d\mathbf{u_I} = [\mathbf{A}_0(t,\mathbf{u_I}) + \mathbf{A}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t,\mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t,\mathbf{u_I}) + \mathbf{a}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t,\mathbf{u_I})d\mathbf{W_{II}}(t),$$

Noisy Lorenz 63 model

— A simple chaotic system

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

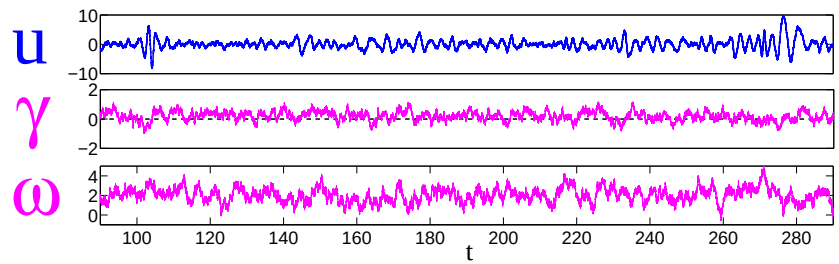
$\rho = 28$
$\sigma = 10$
$\beta = 8/3$



Stochastic parameterizations

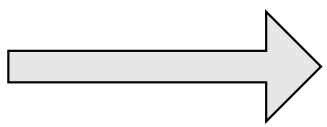
— Capturing intermittency and random phases

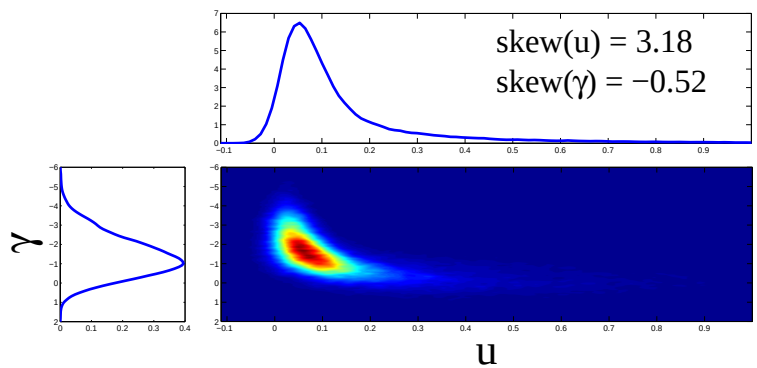
$$du = (-\gamma + i\omega)udt + \sigma_u dW_u,$$
$$d\gamma = -d_\gamma(\gamma - \hat{\gamma})dt + \sigma_\gamma dW_\gamma$$
$$d\omega = -d_\omega(\omega - \hat{\omega})dt + \sigma_\omega dW_\omega$$



Conditional Gaussian ≠ Gaussian

$$du = (-d_u u + vu + f)dt + \sigma_u dW_u,$$
$$dv = (-d_v v - u^2)dt + \sigma_v dW_v$$





Outline

1. Quantifying the uncertainty reduction and understanding the data assimilation skill of recovering ocean flows with noisy Lagrangian tracers.

2. An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions.

3. Predicting an important atmospheric phenomenon via a low-order nonlinear stochastic model.

Outline

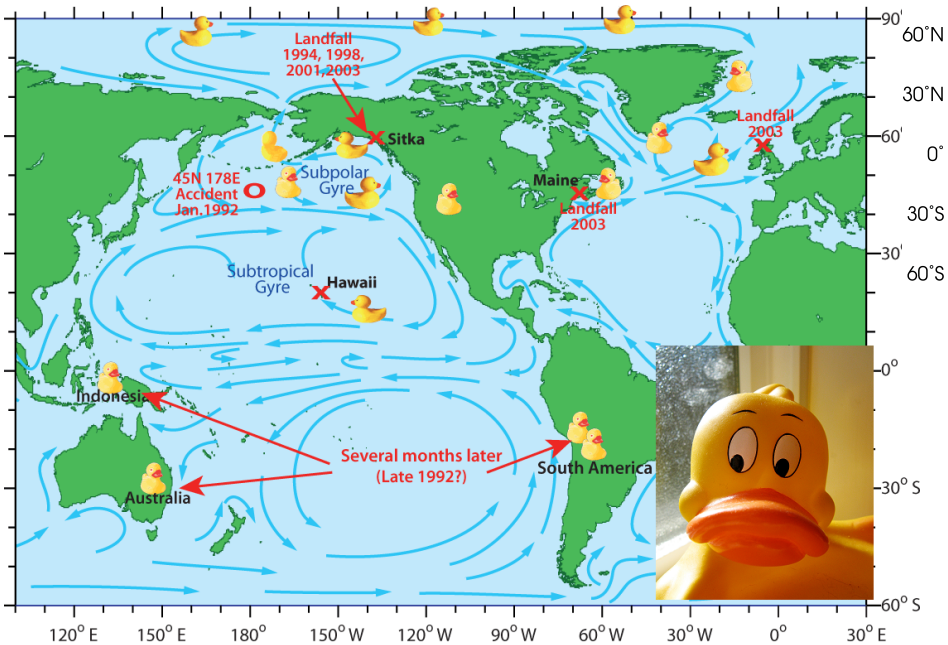
1. Quantifying the uncertainty reduction and understanding the data assimilation skill of recovering ocean flows with noisy Lagrangian tracers.

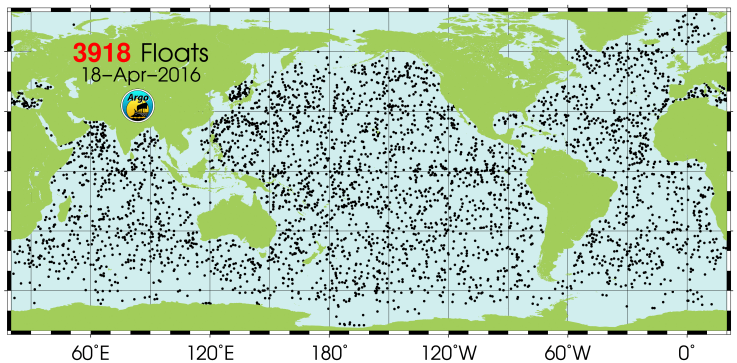
2. An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions.

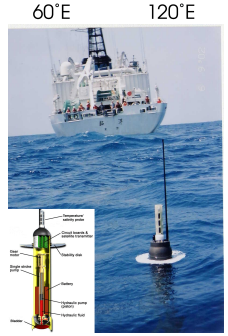
3. Predicting an important atmospheric phenomenon via a low-order nonlinear stochastic model.

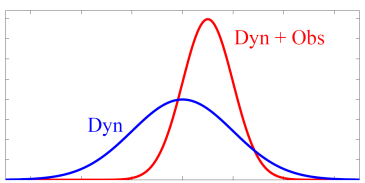
I. Recovering Geophysical Flows with Lagrangian Tracers

- 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.









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

Model set-up.

1. Underlying flows

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

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

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

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

Model set-up.

1. Underlying flows

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

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

Each $\hat{v}_{\vec{k}}(t)$ 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 random flow modeled by a finite number of Fourier modes with random amplitudes in periodic domain $(0, 2\pi]^2$,

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

Each $\hat{v}_{\vec{k}}(t)$ 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 — $p(\mathbf{U}|\mathbf{X})$.

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

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

1. Recovering random incompressible flows
First rigorous math theory

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

Prior distribution	based only on the model	$p(\mathbf{U}_t) \sim \mathcal{N}(\mathbf{m}_t^{att}, \mathbf{R}_t^{att})$
Posterior distribution	combining model and obs	$p(\mathbf{U}_t \vert \mathbf{X}_{s\le t}) \sim \mathcal{N}(\mathbf{m}_t, \mathbf{R}_t)$

To quantify the uncertainty reduction in the posterior distribution $p(\mathbf{U}_t|\mathbf{X}_{s\le t})$ related to the prior $p(\mathbf{U}_t)$, the *relative entropy* is adopted:

$$\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t)) = \int p(\mathbf{U}_t|\mathbf{X}_{s\le t}) \ln \frac{p(\mathbf{U}_t|\mathbf{X}_{s\le t})}{p(\mathbf{U}_t)}$$

For Gaussian distributions,

$$\begin{aligned}
&\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t)) \\
&= \frac{1}{2}\left[(\mathbf{m}_t - \mathbf{m}_t^{att})^*(\mathbf{R}_t^{att})^{-1}(\mathbf{m}_t - \mathbf{m}_t^{att})\right] && \cdots \text{Signal} \\
&+ \frac{1}{2}\left[\mathrm{tr}(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1}) - |\mathbf{K}| - \ln\det(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})\right] && \cdots \text{Dispersion}
\end{aligned}$$

Uncertainty reduction:

$$\begin{aligned}
&\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t)) \\
&\quad = \frac{1}{2}\left[(\mathbf{m}_t - \mathbf{m}_t^{att})^*(\mathbf{R}_t^{att})^{-1}(\mathbf{m}_t - \mathbf{m}_t^{att})\right] && \cdots \text{Signal} \\
&\quad + \frac{1}{2}\left[\mathrm{tr}(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1}) - |\mathbf{K}| - \ln\det(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})\right] && \cdots \text{Dispersion}
\end{aligned}$$

Theorem (Uncertainty Reduction)

As $L \to \infty$, there exists a fixed time $s_0 > 0$ such that for a.s. $\vec{v}_{s\le t}$

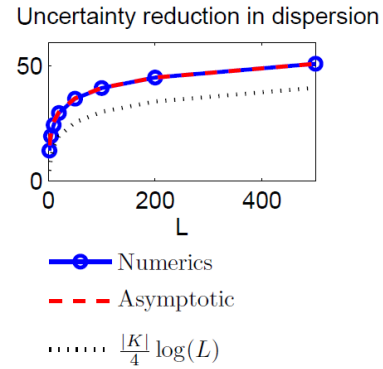
For any $t > s_0$, $\quad \text{Signal} \to \frac{1}{2}(\mathbf{U}_t - \mathbf{m}_t^{att})^*\mathbf{R}_{att}^{-1}(\mathbf{U}_t - \mathbf{m}_t^{att})$ in $\mathbf{P}_{\vec{v}_{s\le t}}$,

For any $t > 0$, $\quad \dfrac{\text{Dispersion}}{\frac{|\mathbf{K}|+2}{4}\ln L} \to 1$ in $\mathbf{P}_{\vec{v}_{s\le t}}$.

Reducing the uncertainty by a fixed amount requires an exponential increase in the number of tracers — **A practical information barrier**!

Uncertainty reduction:

$$\mathcal{P}(p(\mathbf{U}_t|\mathbf{X}_{s\le t}), p(\mathbf{U}_t))$$
$$= \frac{1}{2}\left[(\mathbf{m}_t - \mathbf{m}_t^{att})^*(\mathbf{R}_t^{att})^{-1}(\mathbf{m}_t - \mathbf{m}_t^{att})\right] \quad \cdots \text{Signal}$$
$$+ \frac{1}{2}\left[\mathrm{tr}(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1}) - |\mathbf{K}| - \ln\det(\mathbf{R}_t(\mathbf{R}_t^{att})^{-1})\right] \quad \cdots \text{Dispersion}$$

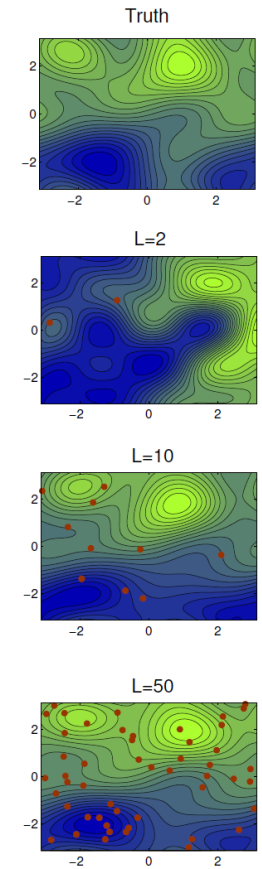


Theorem (Uncertainty Reduction)

As $L \to \infty$, there exists a fixed time $s_0 > 0$ such that for a.s. $\vec{v}_{s\le t}$

For any $t > s_0$, $\quad$ Signal $\to \frac{1}{2}(\mathbf{U}_t - \mathbf{m}_t^{att})^*\mathbf{R}_{att}^{-1}(\mathbf{U}_t - \mathbf{m}_t^{att})$ in $\mathbf{P}_{\vec{v}_{s\le t}}$,

For any $t > 0$, $\quad \dfrac{\text{Dispersion}}{\frac{|\mathbf{K}|+2}{4}\ln L} \to 1$ in $\mathbf{P}_{\vec{v}_{s\le t}}$.



Reducing the uncertainty by a fixed amount requires an exponential increase in the number of tracers — **A practical information barrier**!

2. Noisy Lagrangian tracers for recovering random rotating compressible flows

(Chen, Majda & Tong, *JNLS* 2015; Chen & Majda, *MWR*, 2016)

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

$$\frac{\partial \vec{u}}{\partial t} + \varepsilon^{-1}\vec{u}^{\perp} = -\varepsilon^{-1}\nabla\eta,$$
$$\frac{\partial \eta}{\partial t} + \varepsilon^{-1}\nabla\cdot\vec{u} = 0, \qquad \varepsilon: \text{ the Rossby number.}$$

Two types of modes in the SWE:

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

– Rotating shallow water models with multiscale features:

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

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

– Highly nonlinear observations mixing GB and gravity modes!

Designing cheap practical strategy (filter) to recover the GB flows

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	GB Dynamics	Full Obs	Practical and Cheap

Notations for the posterior estimation of the GB flows:

Full filter: $\mathcal{N}(\mathbf{m}_t^F, \mathbf{R}_t^F)$ GB filter: $\mathcal{N}(\mathbf{m}_t^G, \mathbf{R}_t^G)$ Reduced filter: $\mathcal{N}(\mathbf{m}_t^R, \mathbf{R}_t^R)$

Designing cheap practical strategy (filter) to recover the GB flows

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	GB Dynamics	Full Obs	Practical and Cheap

Notations for the posterior estimation of the GB flows:

Full filter: $\mathcal{N}(\mathbf{m}_t^F, \mathbf{R}_t^F)$ GB filter: $\mathcal{N}(\mathbf{m}_t^G, \mathbf{R}_t^G)$ Reduced filter: $\mathcal{N}(\mathbf{m}_t^R, \mathbf{R}_t^R)$

Theorem (Recovering the large-scale GB flows)

Assume initially the same Gaussian distribution and no correlation between GB and gravity parts. For any fixed $q \geq 1$, $T \geq 0$ there is an ϵ-uniform constant M,

$$\left[\mathbb{E}_{\vec{V}^F} \sup_{t\leq T} \|\mathbf{m}_t^F - \mathbf{m}_t^G\|^{2q}\right]^{\frac{1}{2q}} \leq \epsilon M \qquad \left[\mathbb{E}_{\vec{V}^F} \sup_{t\leq T} \|\mathbf{R}_t^F - \mathbf{R}_t^G\|^{2q}\right]^{\frac{1}{2q}} \leq \epsilon M$$

$$\left[\mathbb{E} \sup_{t\leq T} \|\mathbf{m}_t^F - \mathbf{m}_t^R\|^{2q}\right]^{\frac{1}{2q}} \leq \epsilon M \qquad \left[\mathbb{E} \sup_{t\leq T} \|\mathbf{R}_t^F - \mathbf{R}_t^R\|^{2q}\right]^{\frac{1}{2q}} \leq \epsilon M$$

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

Outline

1. Quantifying the uncertainty reduction and understanding the data assimilation skill of recovering ocean flows with noisy Lagrangian tracers.

2. An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions.

3. Predicting an important atmospheric phenomenon via a low-order nonlinear stochastic model.

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

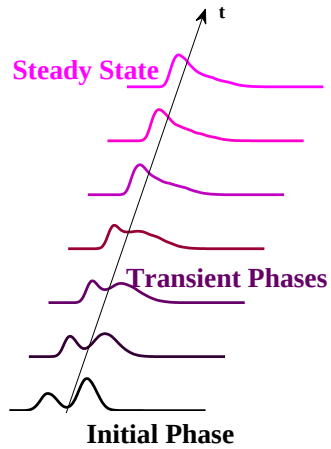
(Chen & Majda, *JCP*, 2017, *PNAS*, 2017; Chen, Majda & Tong, *SIAM UQ*, 2017)

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 Fokker-Planck equation describes the time evolution of the probability density function (PDF) associated with $\mathbf{u}$,

$$\frac{\partial}{\partial t}p(\mathbf{u}, t) = -\nabla_{\mathbf{u}}\big(\mathbf{F}(\mathbf{u}, t)p(\mathbf{u}, t)\big) + \frac{1}{2}\nabla_{\mathbf{u}} \cdot \nabla_{\mathbf{u}}(\mathbf{\Sigma}\mathbf{\Sigma}^T(\mathbf{u}, t)p(\mathbf{u}, t)).$$



- important in solving $p(\mathbf{u}, t)$ for both steady state and transient phases.
- features in many applications: large dimensions and strong non-Gaussianity.

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 with a Hybrid Strategy

(Chen & Majda, *JCP*, 2017)

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

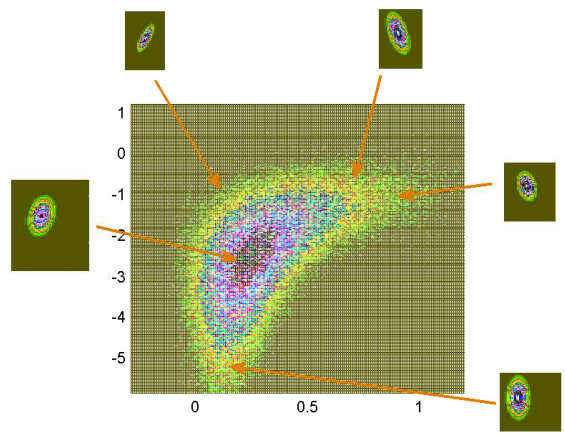
$$d\mathbf{u_I} = [\mathbf{A}_0(t,\mathbf{u_I}) + \mathbf{A}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_I}(t,\mathbf{u_I})d\mathbf{W_I}(t),$$
$$d\mathbf{u_{II}} = [\mathbf{a}_0(t,\mathbf{u_I}) + \mathbf{a}_1(t,\mathbf{u_I})\mathbf{u_{II}}]dt + \mathbf{\Sigma_{II}}(t,\mathbf{u_I})d\mathbf{W_{II}}(t).$$

Step 1. Sample L trajectories of $\mathbf{u_I}$ (e.g., by Monte Carlo), namely $\mathbf{u_I}^i(s \le t), i = 1, \ldots, L$.

Step 2. $p(\mathbf{u_{II}}(t))$ is given by L conditional Gaussian distributions,

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

- Each $p\big(\mathbf{u_{II}}(t)|\mathbf{u_I}^i(s \le t)\big)$ is optimal.
- Each $p\big(\mathbf{u_{II}}(t)|\mathbf{u_I}^i(s \le t)\big)$ has closed analytic form.
- $p\big(\mathbf{u_{II}}(t)|\mathbf{u_I}^i(s \le t)\big)$ with different i can be solved in a parallel way.



Computationally efficient and accurate!

Practically, only a small number of samples L is needed.

An Efficient Statistically Accurate Algorithm with a Hybrid Strategy

(Chen & Majda, *JCP*, 2017)

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

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

Step 3. $p(\mathbf{u_I}(t))$ is computed using a Gaussian kernel method (Botev et al., 2010),

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

Step 4. The joint PDF is given by a Gaussian mixture,

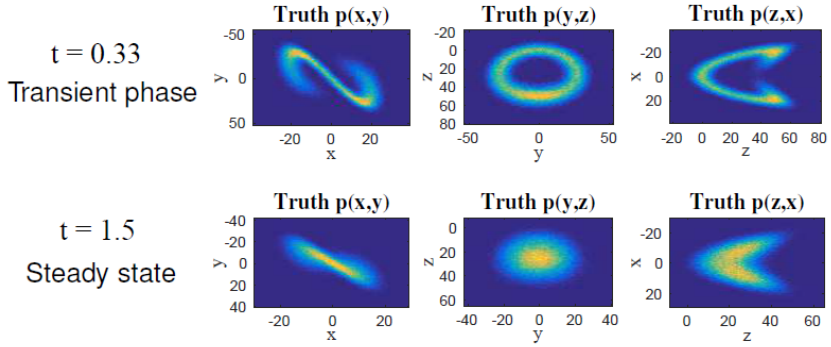
$$p(\mathbf{u_I}(t), \mathbf{u_{II}}(t)) = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} \Big(K_{\mathbf{H}}(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)) \cdot p(\mathbf{u_{II}}(t)|\mathbf{u_I^i}(s \le t))\Big).$$

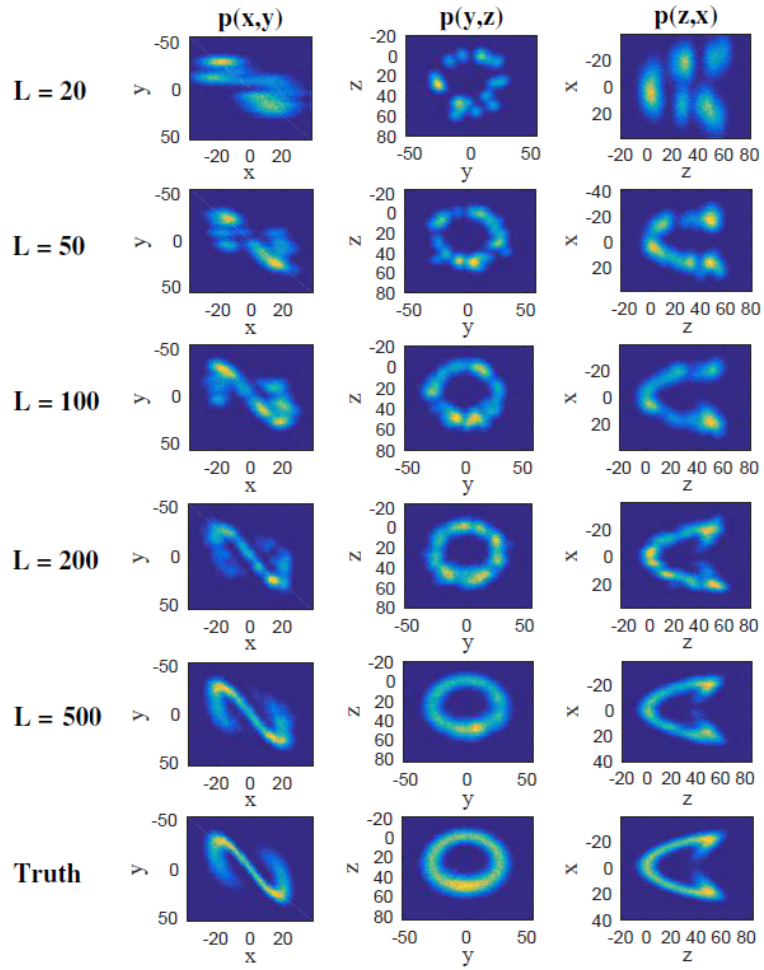
- Practically, $L \sim O(100)$ is able to handle systems with $\mathrm{Dim}(\mathbf{u_I}) \le 3$ and $\mathrm{Dim}(\mathbf{u_{II}}) \sim O(10)$.
- Rigorous analysis shows that a much smaller L is needed compared with direct Monte Carlo methods especially when $\mathrm{Dim}(\mathbf{u_{II}})$ is large (Chen, Majda & Tong 2018).
- For system with much larger dimensions, block decomposition and statistical symmetry can be applied. See Majda's talk in MS6 (Chen & Majda, *PNSA*, 2017).

A simple numerical illustration: The noisy Lorenz 63 Model.

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

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





$\Longleftarrow$ Recovering the PDFs at a transient phase $t = 0.33$.

$L_{MC} = 20,000$ samples is needed using direct Monte Carlo for reaching the same accuracy as $L = 200$ in the algorithm here!

Outline

1. Quantifying the uncertainty reduction and understanding the data assimilation skill of recovering ocean flows with noisy Lagrangian tracers.

2. An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions.

3. Predicting an important atmospheric phenomenon via a low-order nonlinear stochastic model.

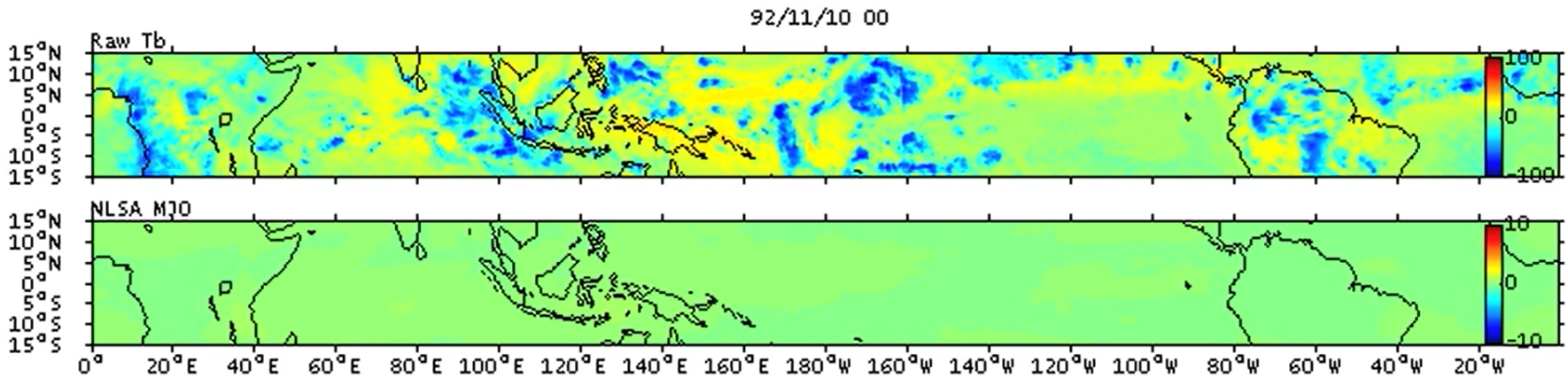
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

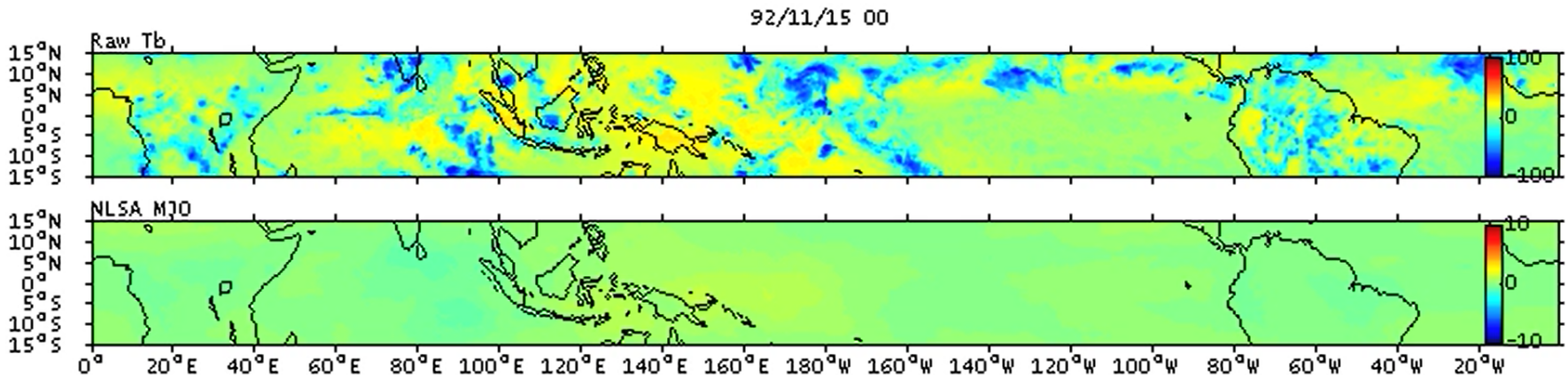
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

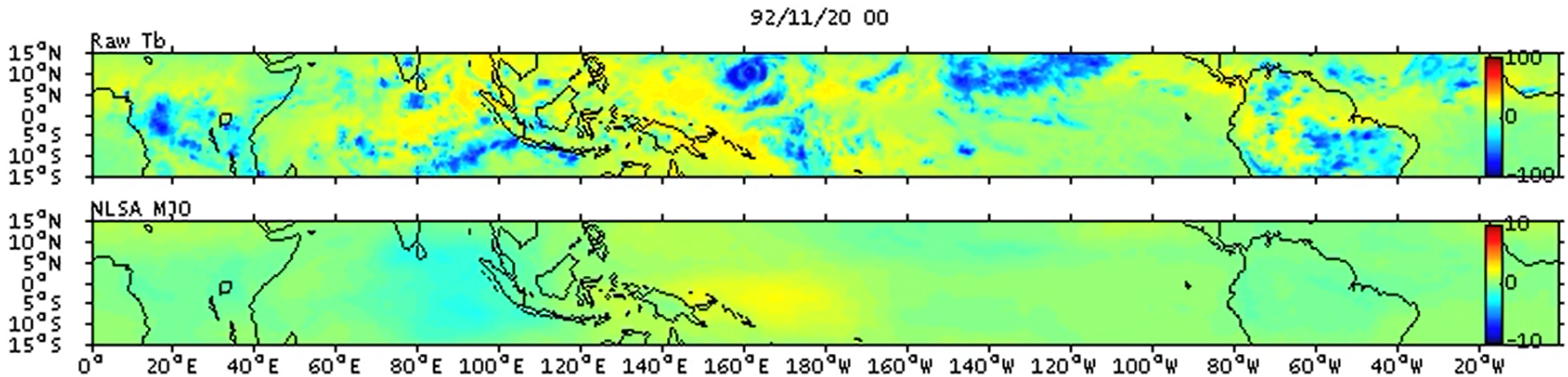
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

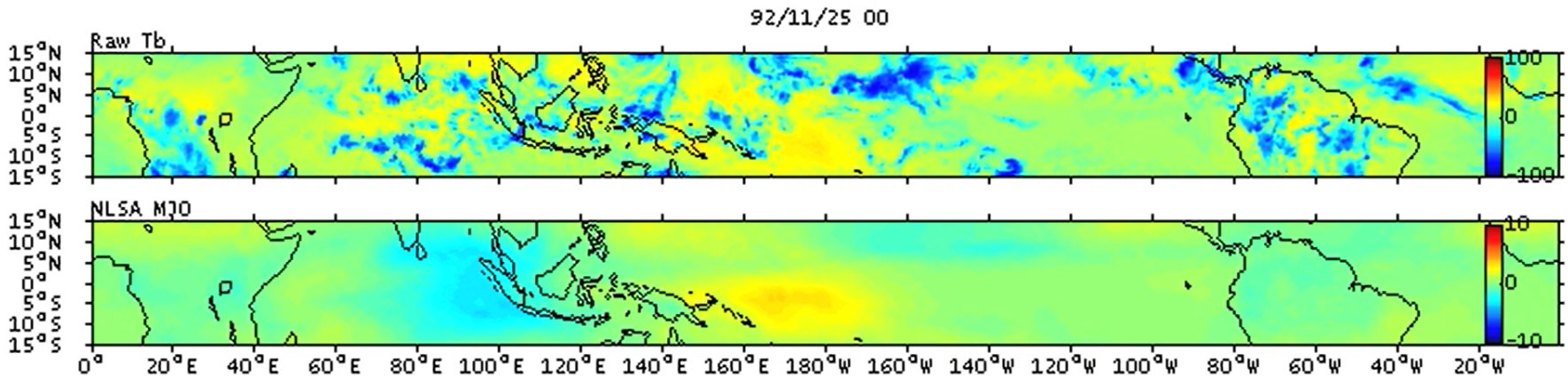
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

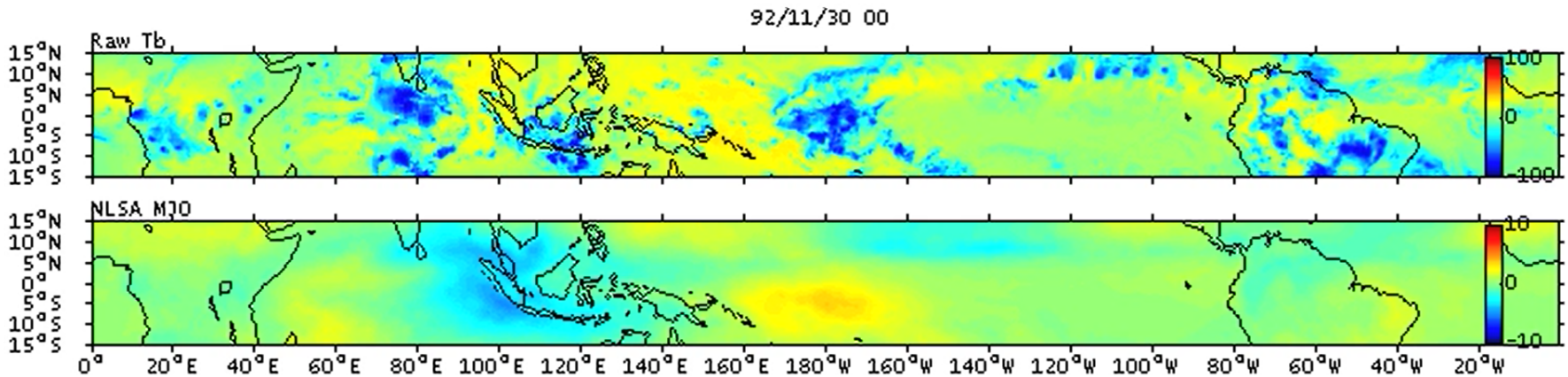
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

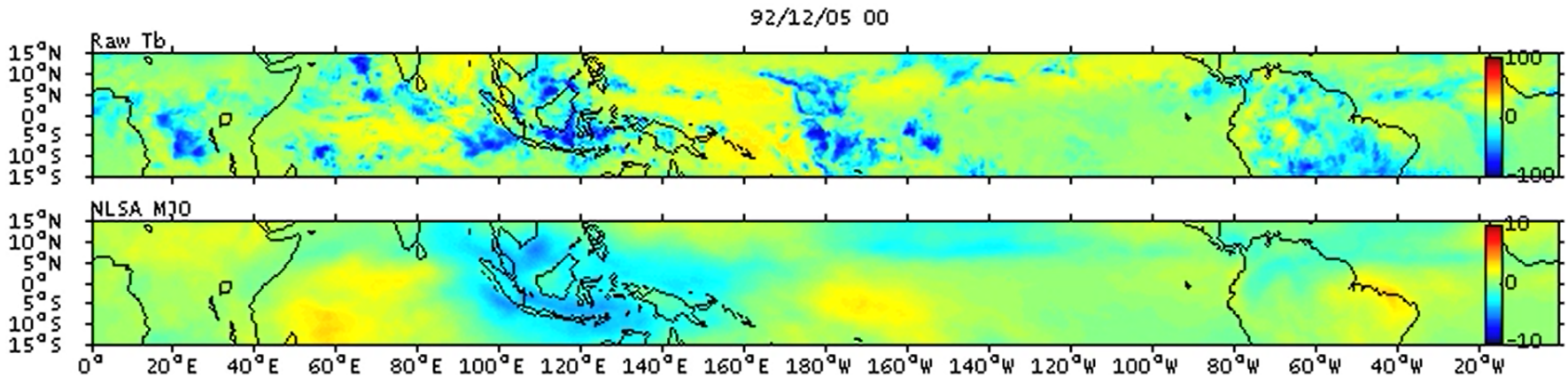
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

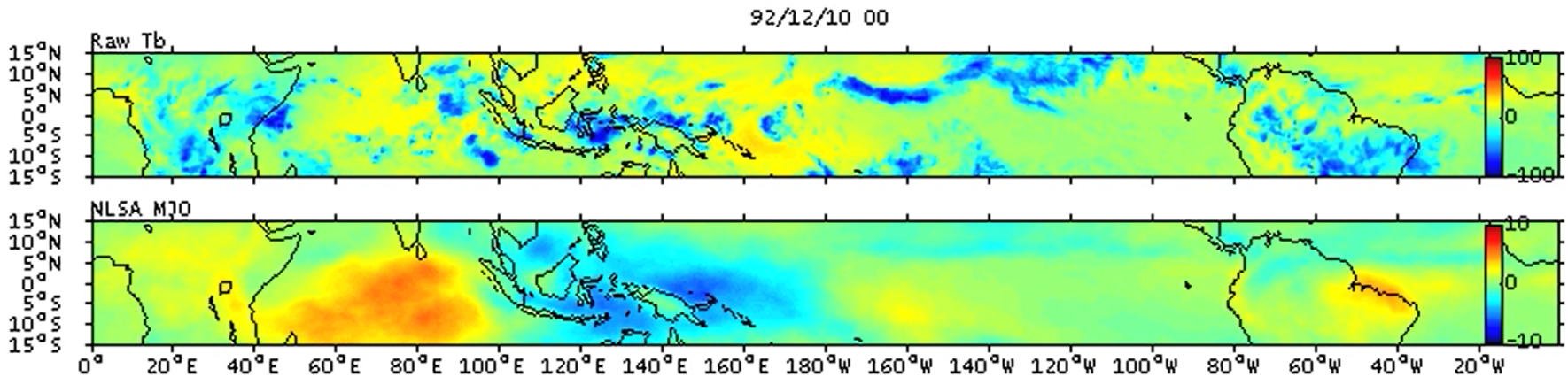
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

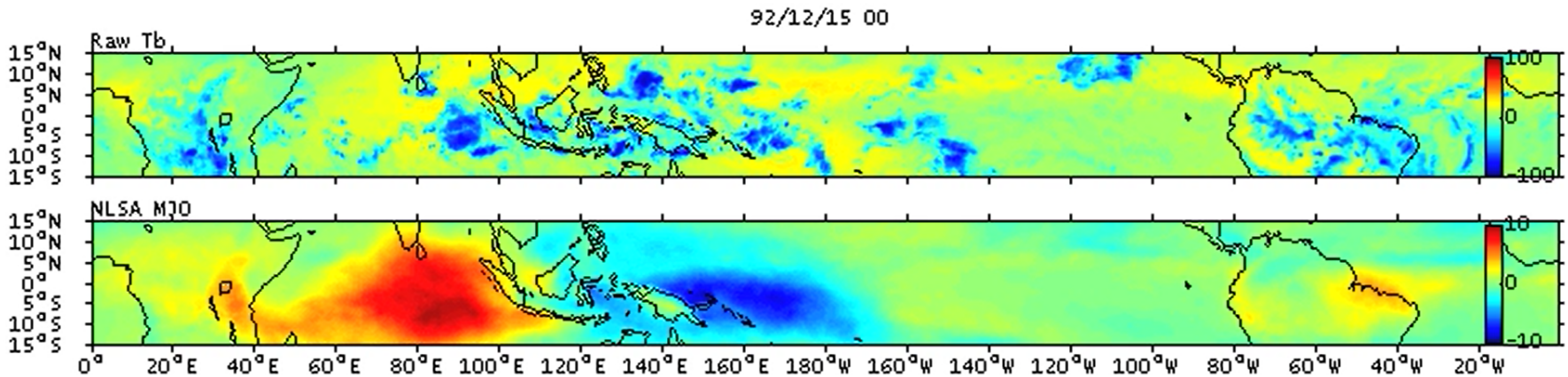
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

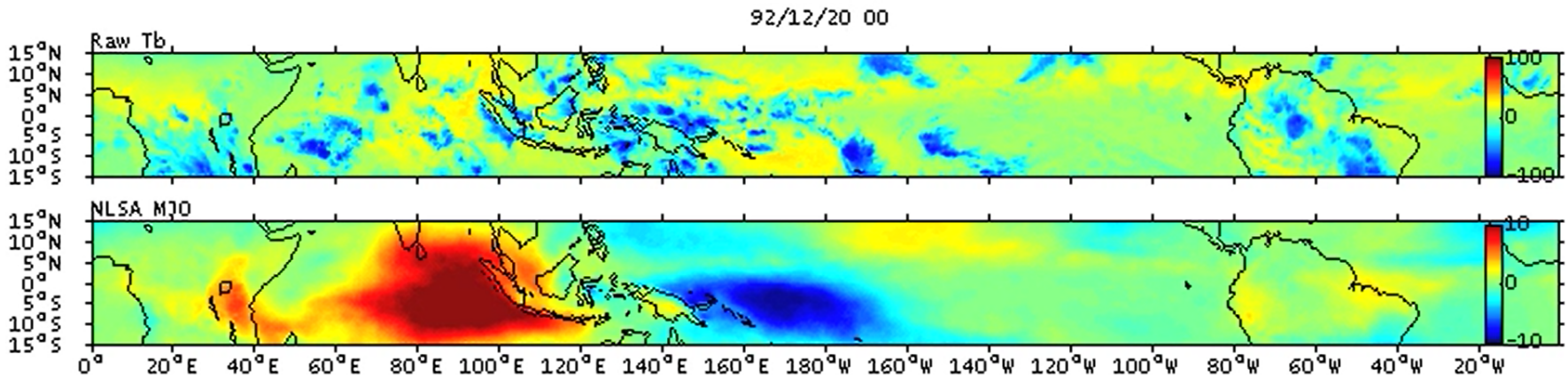
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

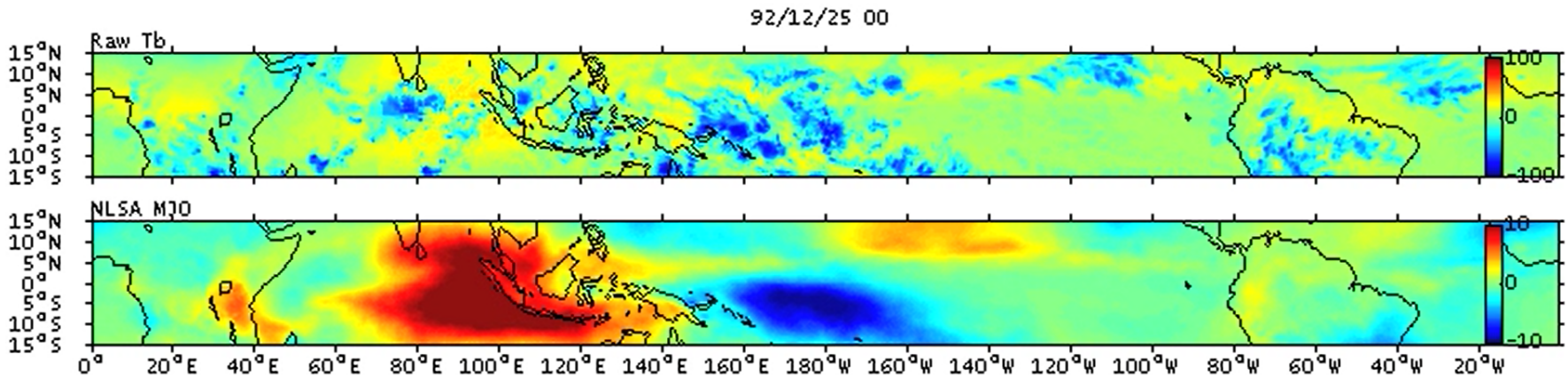
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

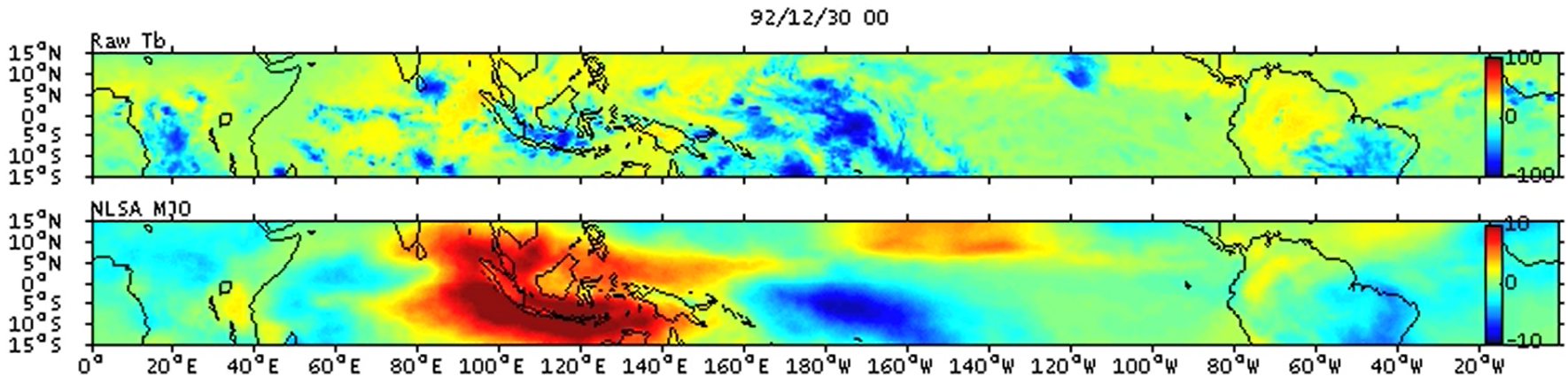
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

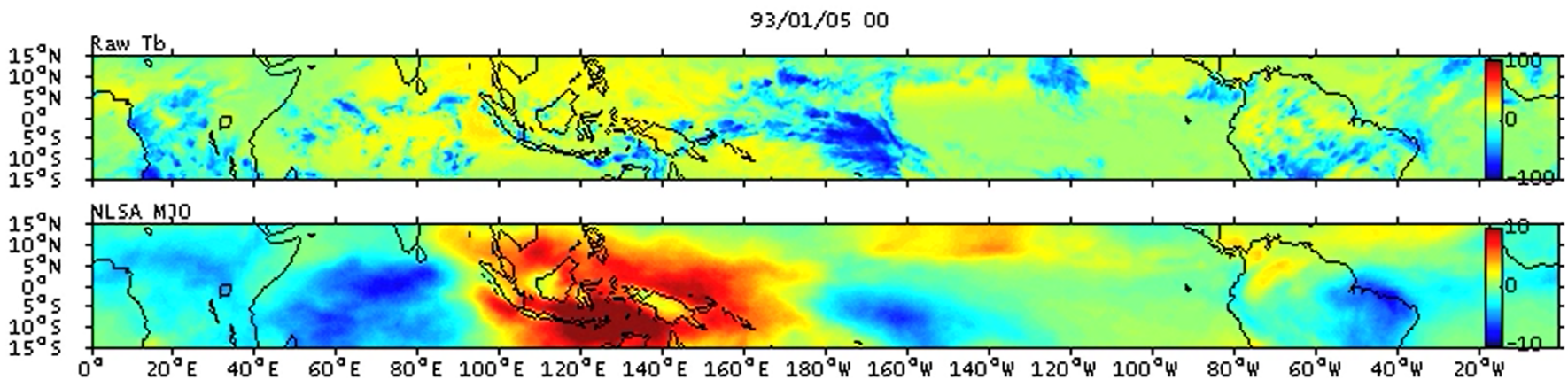
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

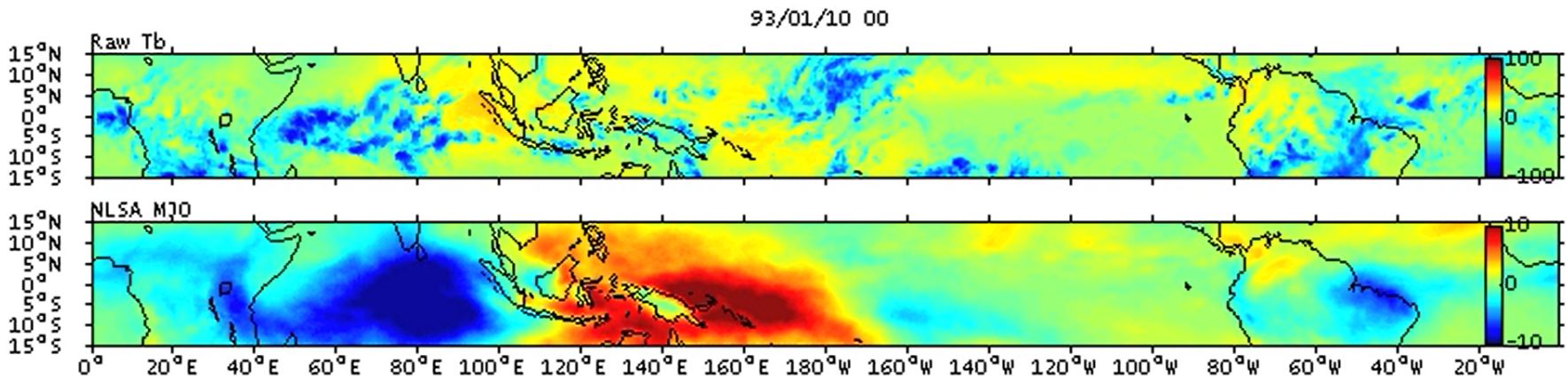
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

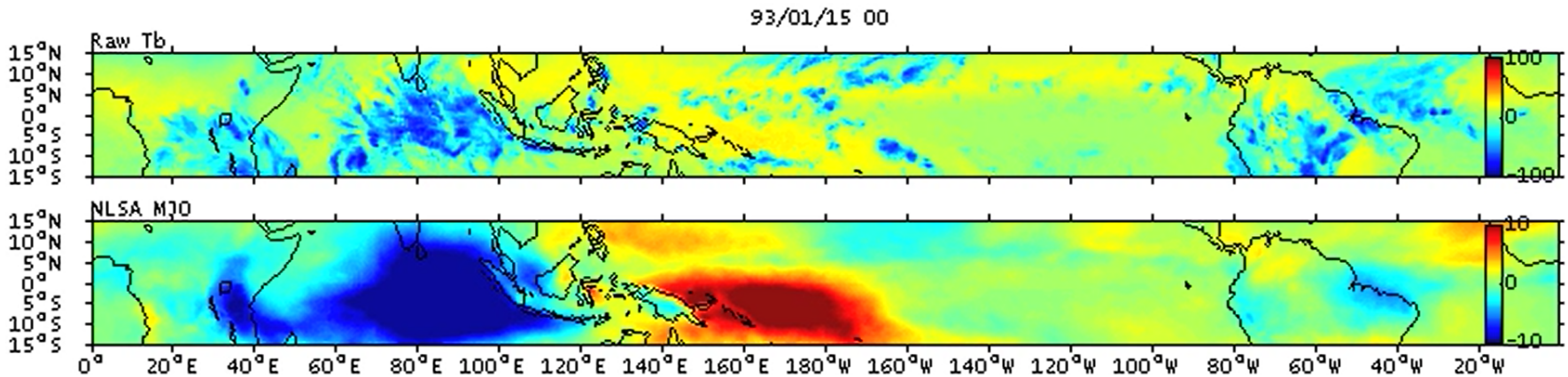
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

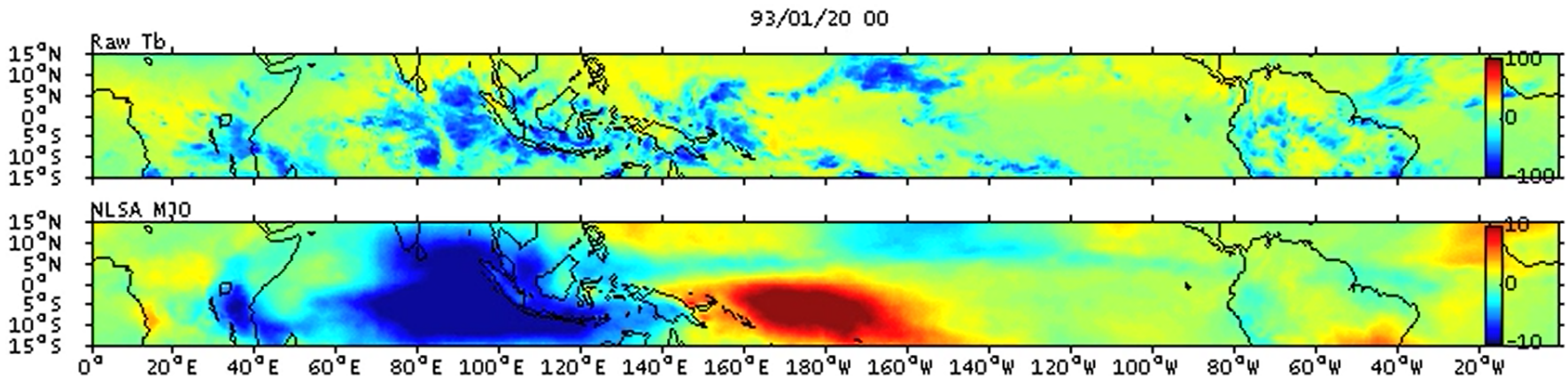
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

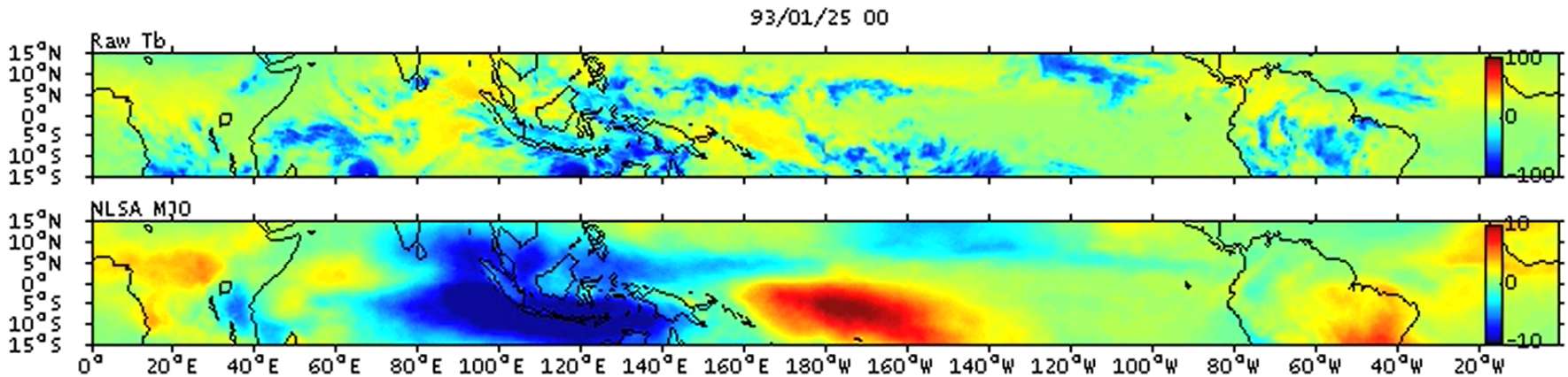
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

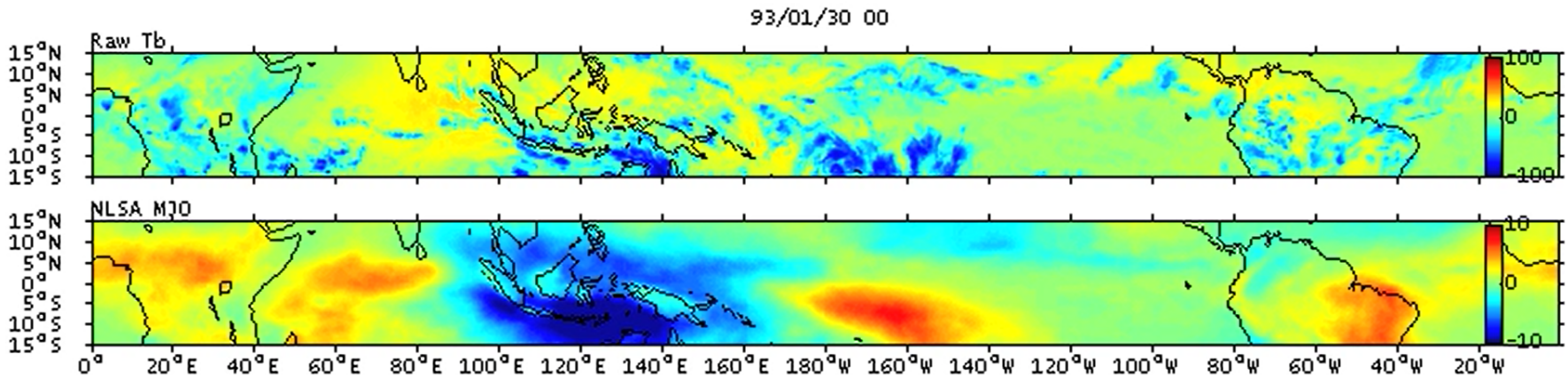
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

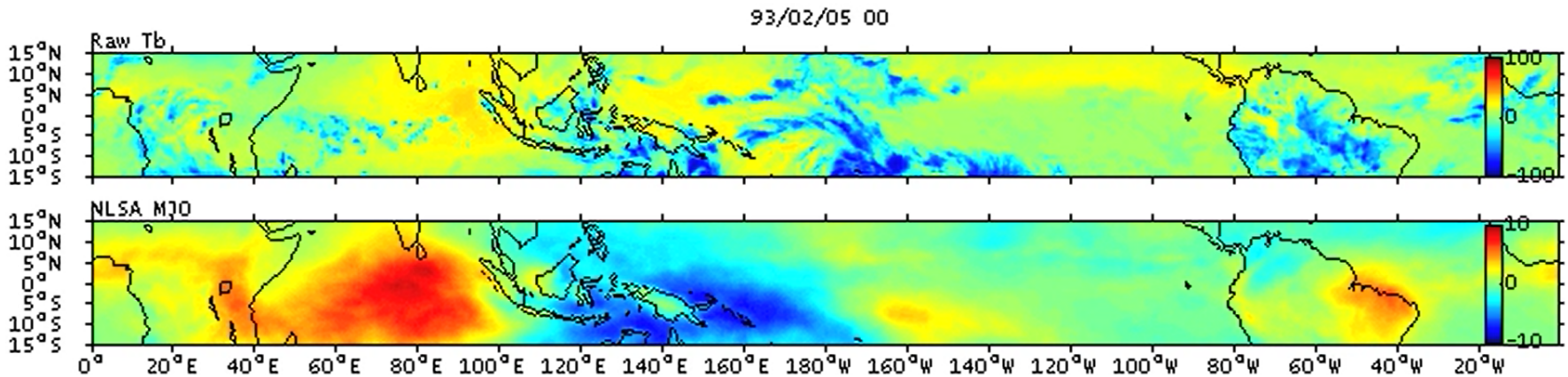
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

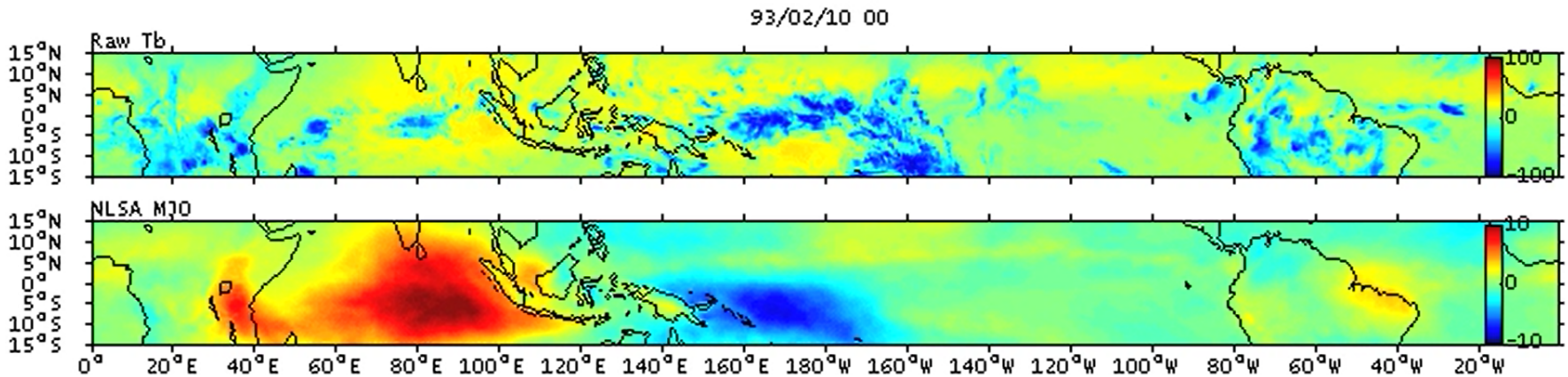
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

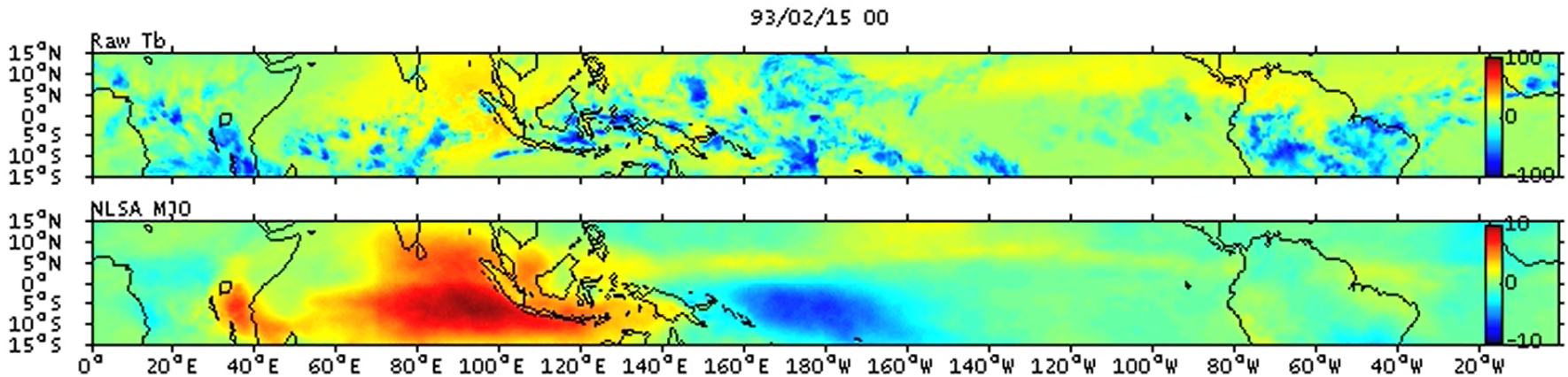
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

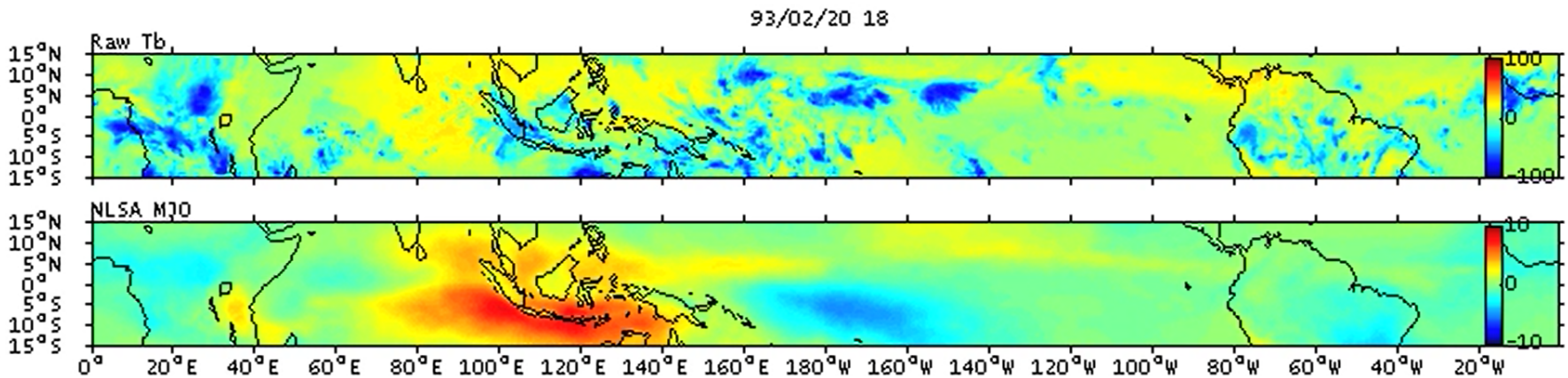
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

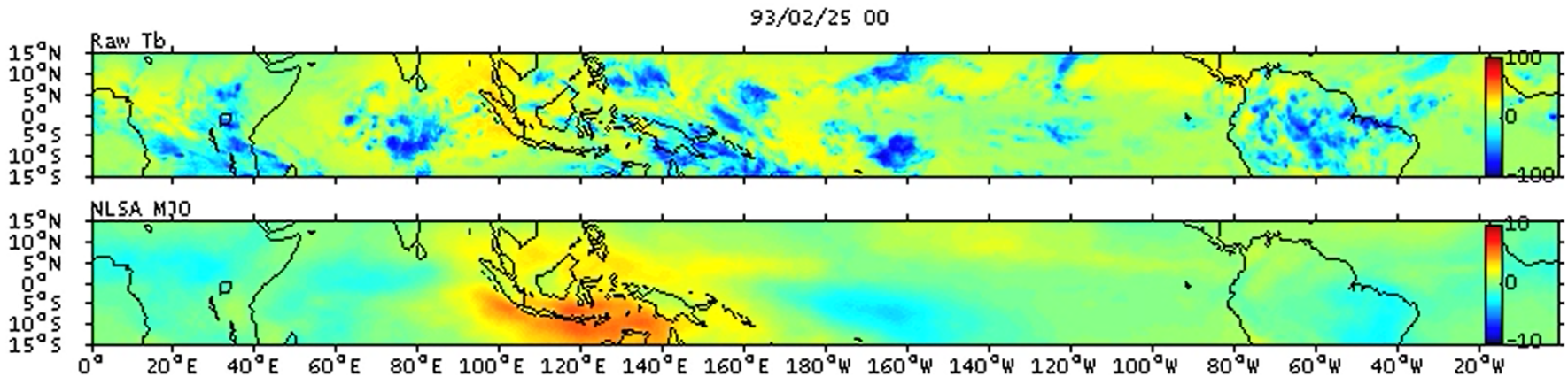
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

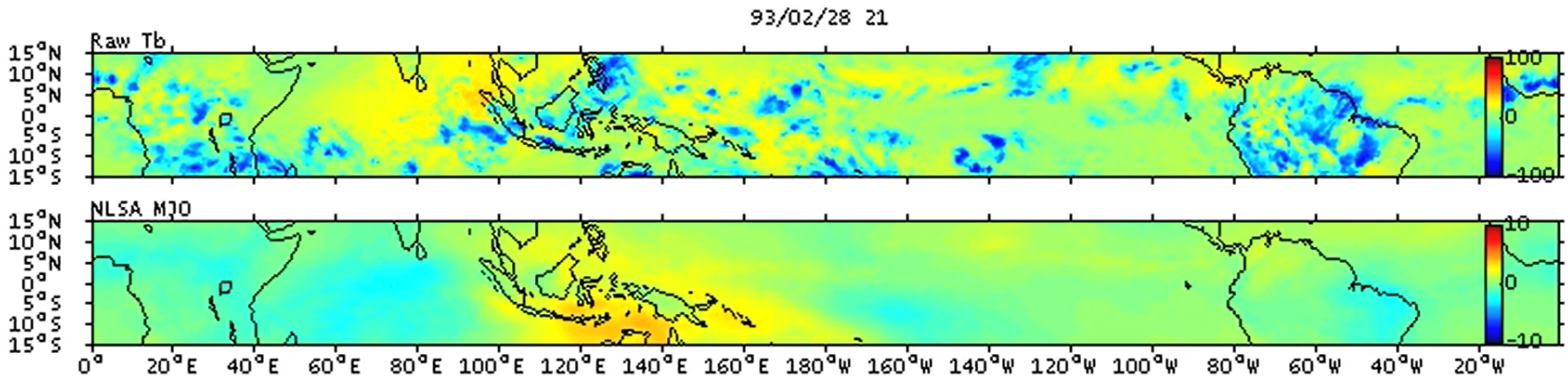
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

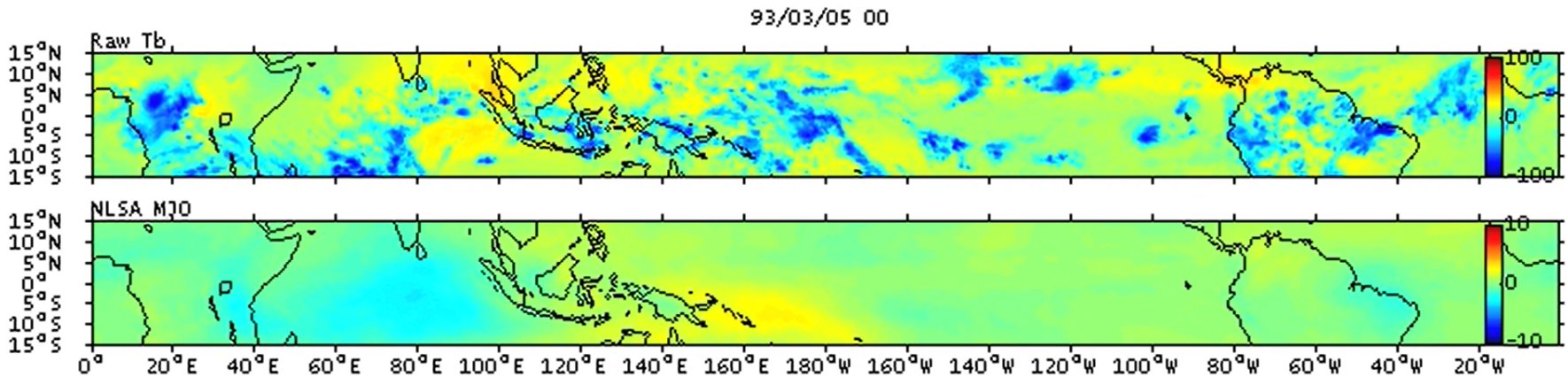
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

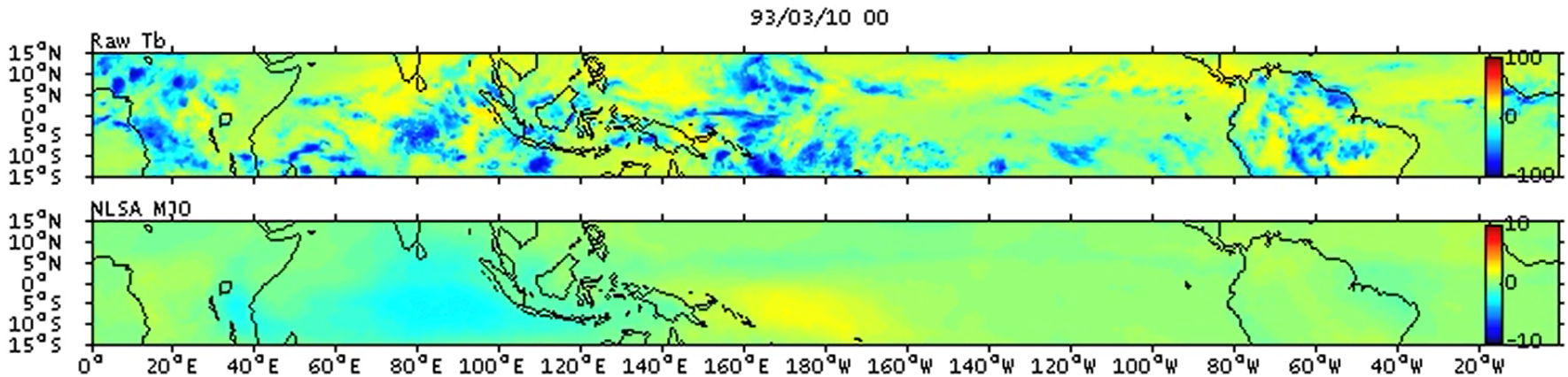
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

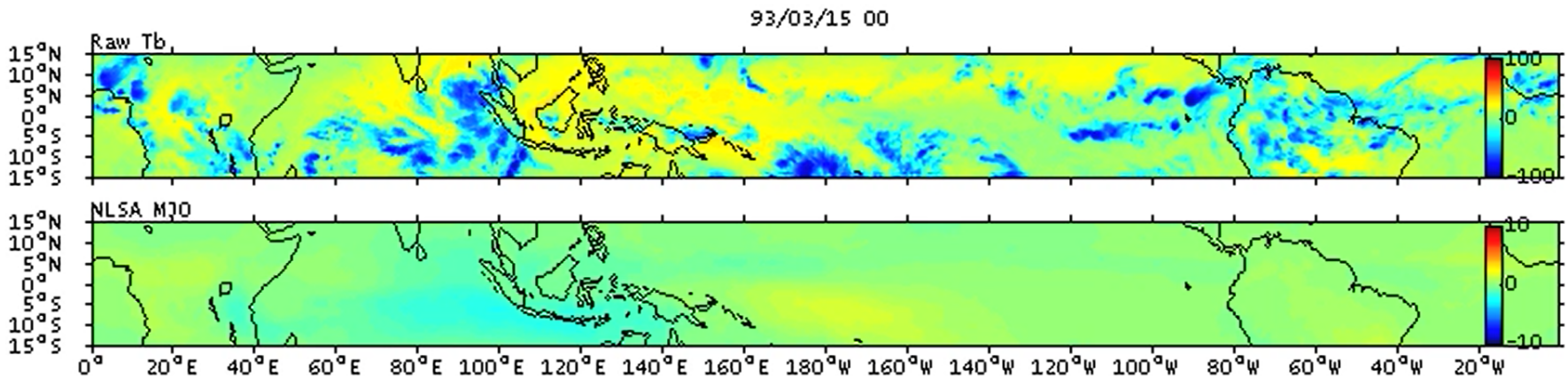
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

- 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 field campaign of 1992-1993 (Webster & Lukas).

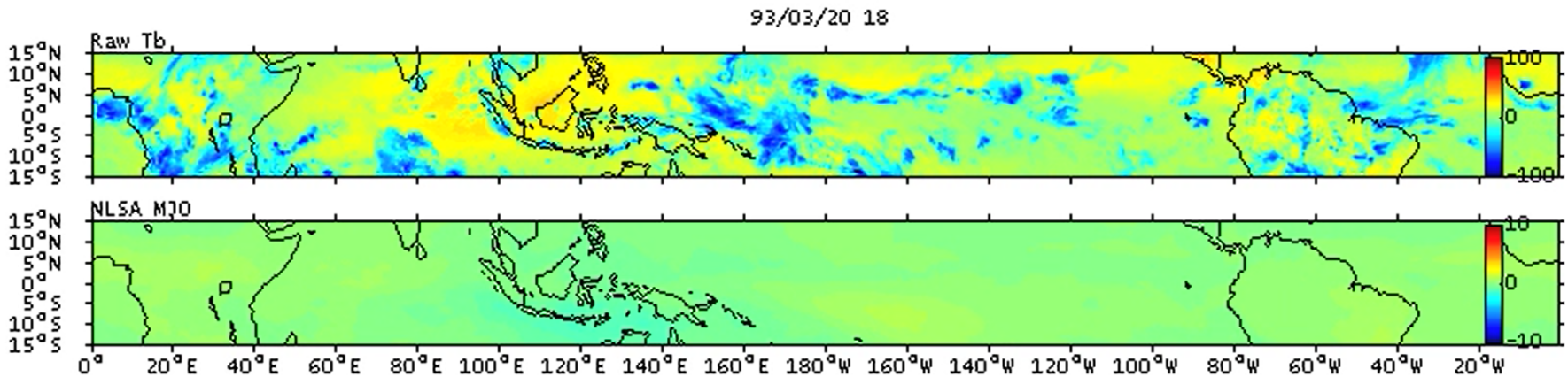
III. Predicting the Large-Scale Madden-Julian Oscillation

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

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

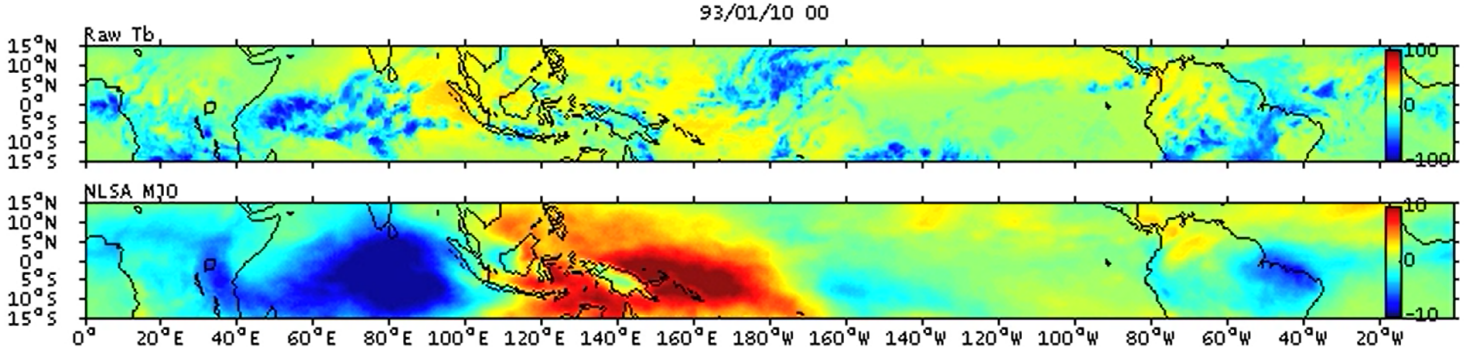
- A novel nonlinear techniques, **Nonlinear Laplacian Spectral Analysis (NLSA)**, is applied to the cloudiness data of dimensions $O(10^5)$ (Giannakis & Majda, *PNAS*, 2012).
- NLSA captures nonlinear dynamical features such as intermittency and extreme events.



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

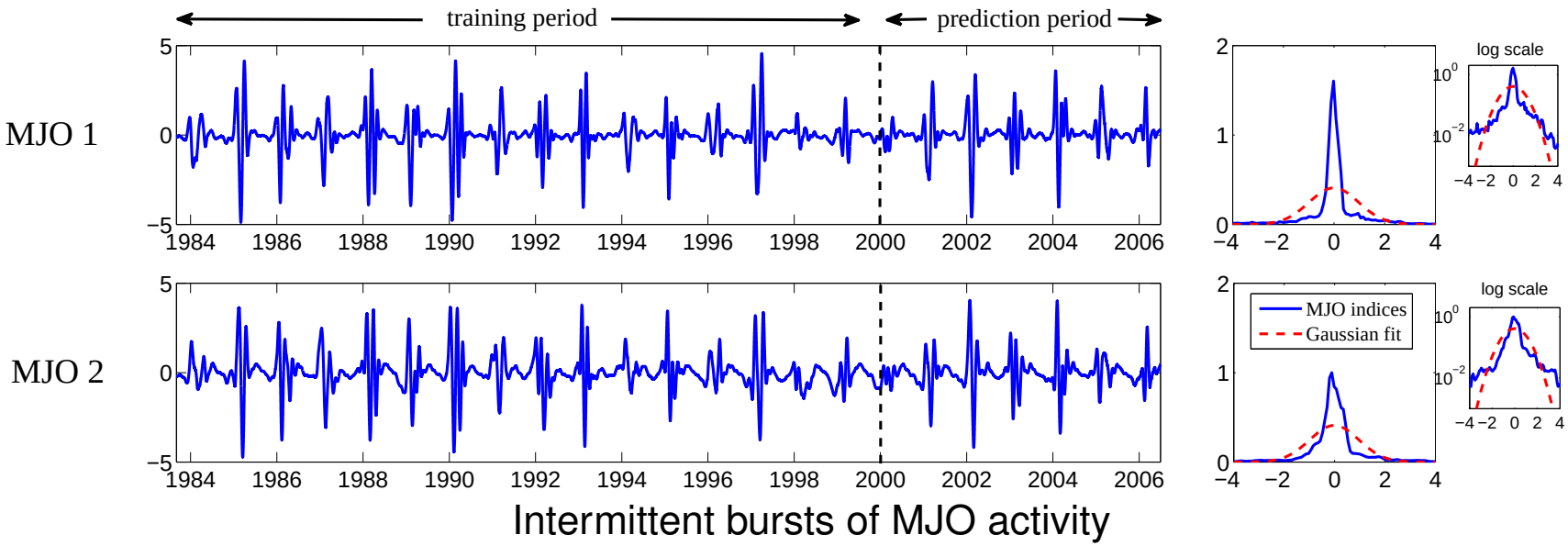
red: weak convection (clear sky). blue: strong convection (heavy rainfall).

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



NLSA Large-Scale MJO Patterns
= Spatial Basis (1) $\times$ Time Series (1) + Spatial Basis (2) $\times$ Time Series (2)

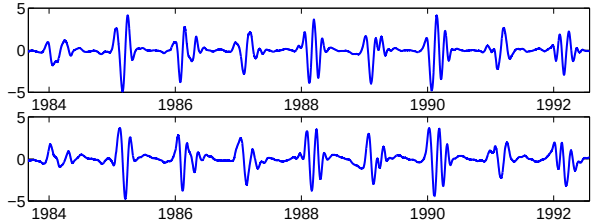
NLSA Time-Series Techniques $\Longrightarrow$ 2 components of MJO Cloud Patterns



Intermittent bursts of MJO activity

Physics-Constrained Low-Order Stochastic Model

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



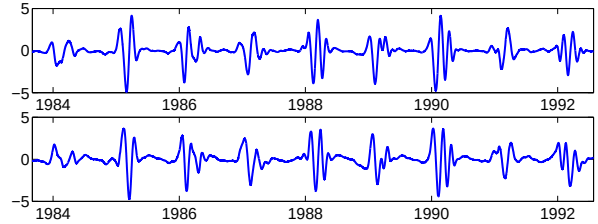
,

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



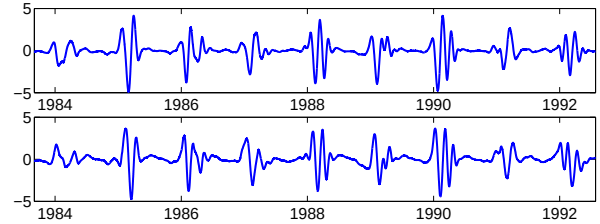
$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v \qquad\qquad\qquad\quad)\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\,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

with

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

- 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



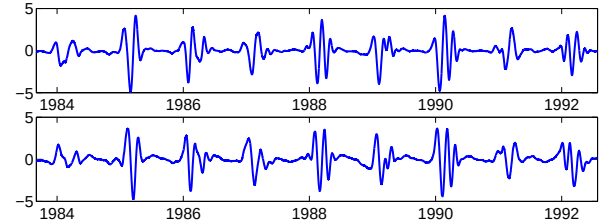
$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\,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

with

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

- 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



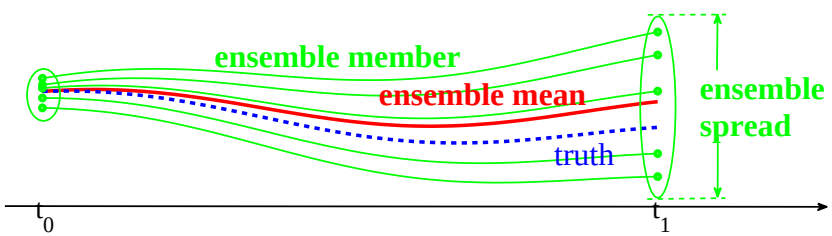
$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\,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

with

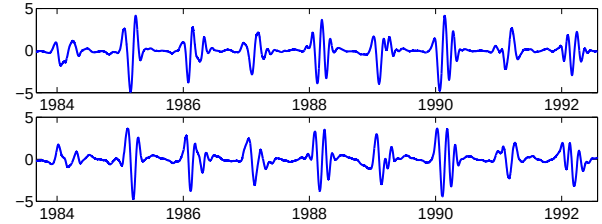
$$d_u(t) = d_{u0} + d_{u1} \sin(\omega_f t + \phi).$$

- 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



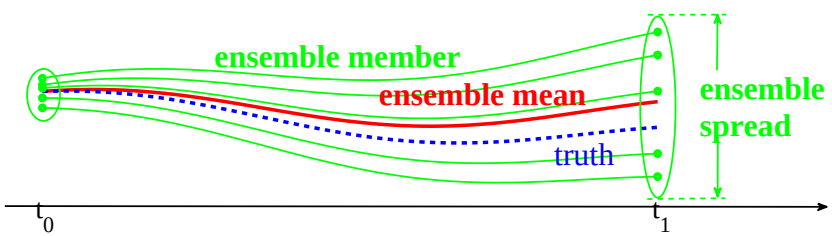
$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\,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

with

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

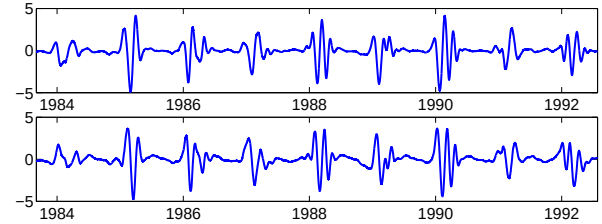
- Observed variables u_1, u_2: MJO 1 and MJO 2 indices from NLSA.
- Hidden variables v, ω: stochastic damping and stochastic phase.
- Energy-conserving nonlinear interactions between (u_1, u_2) and (v, ω).
  (Majda, Harlim, 2012)

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 v, ω?

Physics-Constrained Low-Order Stochastic Model



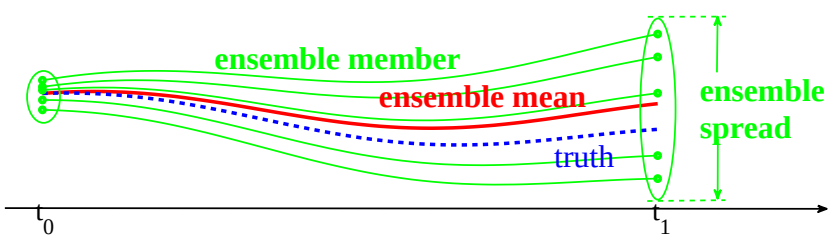
$$
\begin{aligned}
du_1 &= (-d_u(t)\, u_1 + \gamma\, v\, u_1 - \omega\, u_2)\, dt + \sigma_u\, dW_{u_1}, \\
du_2 &= (-d_u(t)\, u_2 + \gamma\, v\, u_2 + \omega\, u_1)\, dt + \sigma_u\, dW_{u_2}, \\
dv &= (-d_v\, v - \gamma\, (u_1^2 + u_2^2))\, dt + \sigma_v\, dW_v, \\
d\omega &= (-d_\omega \omega + \hat{\omega})\,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

with

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

- 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)

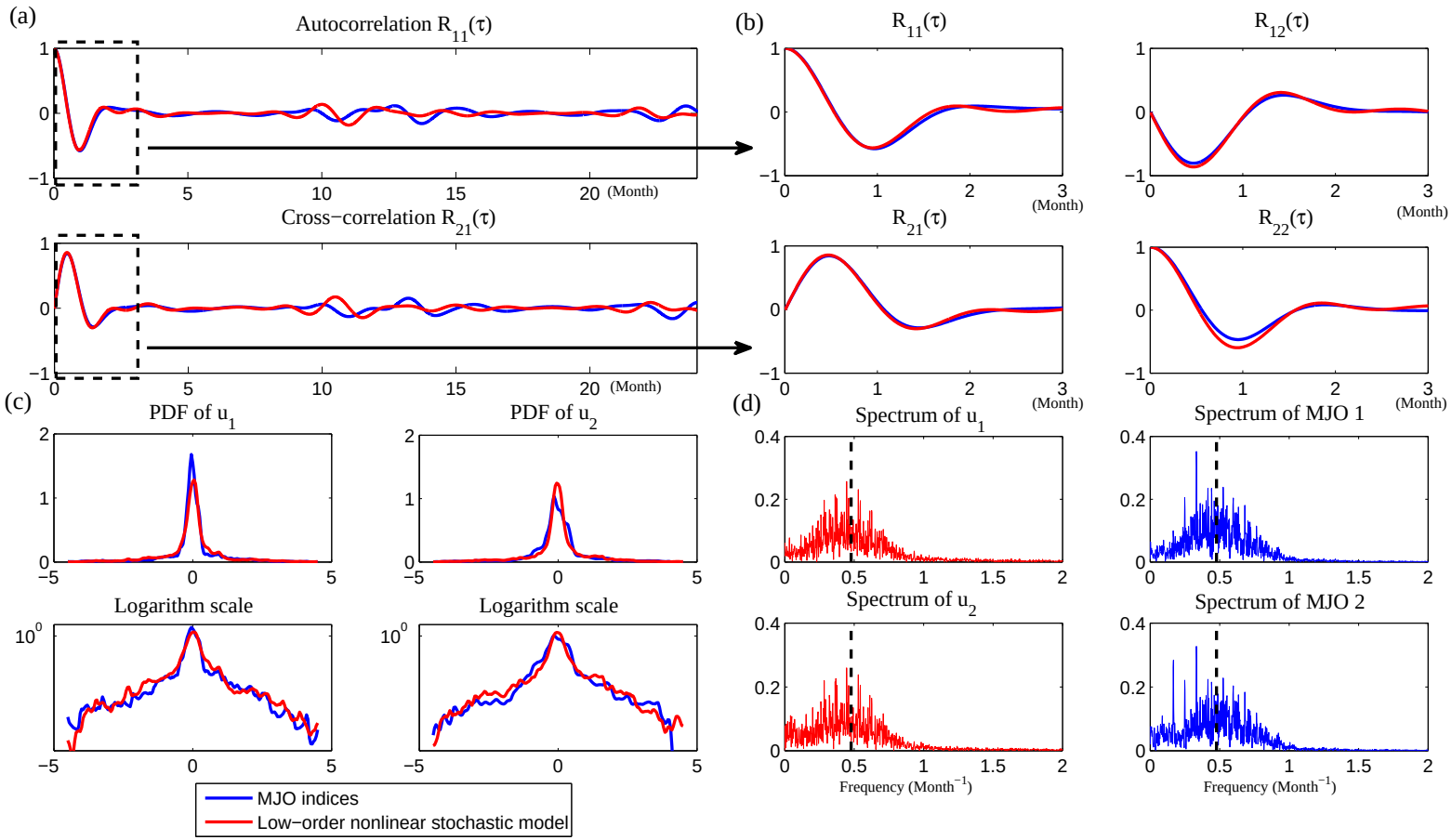
Prediction. Given the initial values of (u_1, u_2) and (v, ω), run an ensemble forecast.



How to determine the initial values of the hidden variables v, ω?

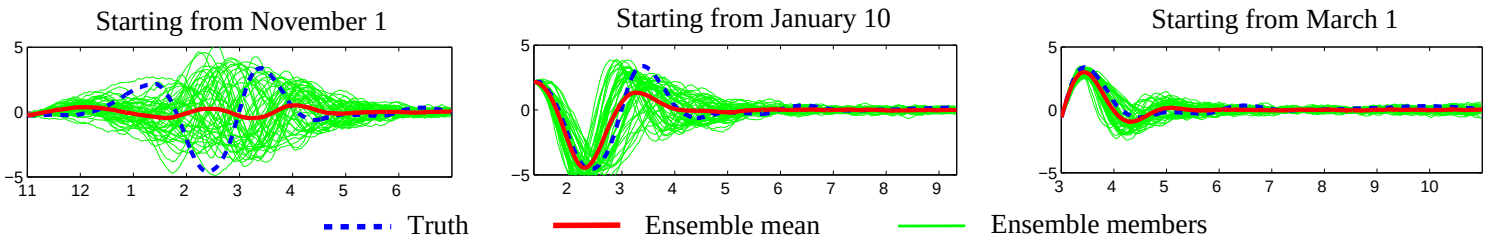
Effective data assimilation of $p(\mathbf{u_{II}}(t)|\mathbf{u_I}(s \le t))$ based on the conditional Gaussian framework!

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



almost perfectly capturing the observational statistics
(correlation functions, highly non-Gaussian PDFs, and power spectrums)

Example: forecasting in year 2002



- Ensemble mean prediction is skillful for 25-50 days in different years.
- Ensemble spread captures the onset and demise of the MJO events and the long-range forecast uncertainty.
- Small uncertainty spread at short terms indicates the accuracy of the ensemble mean prediction.

Other Topics 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
2. predicting rare and extreme events
3. exploring the causality between different processes
4. data assimilation and the prediction of spatial-extended systems

Example 1: Boussinesq equation.

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

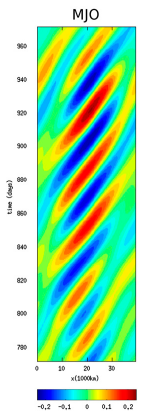
– Observe velocity **u** + noise.

– 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 **a** + noise.

– Recover temperature q and velocity u, v.

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

Summary

A conditional Gaussian framework for uncertainty quantification, data assimilation and prediction is developed,

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

Despite the conditional Gaussianity, the system remains highly nonlinear and is able to capture the non-Gaussian features as observed in nature.

1. Quantifying the uncertainty reduction and understanding the data assimilation skill of recovering ocean flows with noisy Lagrangian tracers
   - first rigorous theory showing the practical information barrier
   - cheap practical strategies for recovering the multiscale turbulent geophysical flows
2. An efficient statistically accurate algorithm for solving the Fokker-Planck equation in high dimensions with strongly non-Gaussian features
   - computationally efficient and accurate, much cheaper than Monte Carlo
   - overcoming the curse of dimensionality
3. Predicting the large-scale MJO via a low-order nonlinear stochastic model
   - energy-conserving nonlinear interactions between the observed and hidden variables
   - almost perfectly capturing the observed highly non-Gaussian statistics
   - skillful prediction and long-term uncertainty quantification

Conditional Gaussian framework

Chen and Majda. "Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics." *Monthly Weather Review* 144.12 (2016): 4885-4917.

1. Data assimilation of turbulent flows using Lagrangian tracers

Chen, Majda and Tong. "Information barriers for noisy Lagrangian tracers in filtering random incompressible flows." *Nonlinearity* 27.9 (2014): 2133.

Chen, Majda and Tong. "Noisy Lagrangian tracers for filtering random rotating compressible flows." *Journal of Nonlinear Science* 25.3 (2015): 451-488.

Chen and Majda. "Model error in filtering random compressible flows utilizing noisy Lagrangian tracers." *Monthly Weather Review* 144.11 (2016): 4037-4061.

2. Efficient algorithms for the Fokker-Planck equation in large dimensions

Chen and Majda. "Efficient statistically accurate algorithms for the Fokker-Planck equation in large dimensions." *Journal of Computational Physics* 354 (2018): 242-268.

Chen and Majda. "Beating the curse of dimension with accurate statistics for the Fokker¨CPlanck equation in complex turbulent systems." *Proceedings of the National Academy of Sciences* (2017): 201717017.

Chen, Majda, and Tong. "Rigorous analysis for efficient statistically accurate algorithms for solving Fokker-Planck equations in large dimensions." *SIAM/ASA Journal on Uncertainty Quantification* (2017).

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

Chen, Majda and Giannakis. "Predicting the cloud patterns of the Madden-Julian Oscillation through a low-order nonlinear stochastic model." *Geophysical Research Letters* 41.15 (2014): 5612-5619.

Thank you

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,

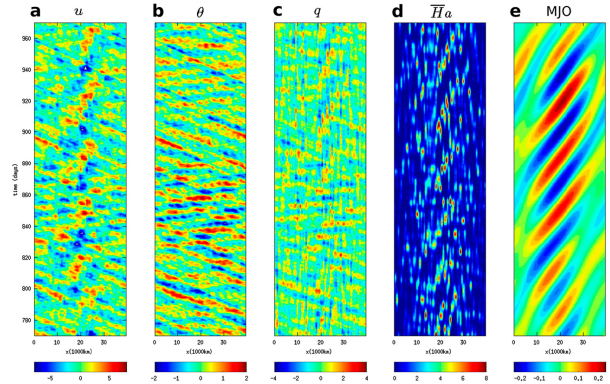
$$du = (a^* u + b^* u^2 - c^* u^3 + f^*)dt + \sigma_u dW_u,$$

	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.

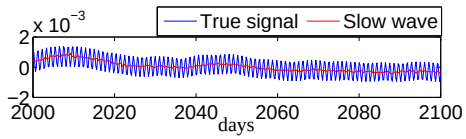
(Chen & Majda, Monthly Weather Review 2015)

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

$$p\Big(\mathbf{u_I}(t), \mathbf{u_{II}}(t)\Big) = \int p\Big(\mathbf{u_{II}}(t), \mathbf{u_I}(t) \,|\, \mathbf{u_I}(s \le t)\Big)\, p\Big(\mathbf{u_I}(s \le t)\Big)\, d\mathbf{u_I}(s \le t) \tag{2}$$

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

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

The second term on the right hand side of (3) 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

$$p\Big(\mathbf{u_I}(t) \,|\, \mathbf{u_I^i}(s \le t)\Big) = \delta(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)) \tag{4}$$

In addition,

$$p(\mathbf{u_I}(s \le t)) = \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \delta\left(\mathbf{u_I}(s \le t) - \mathbf{u_I^i}(s \le t)\right). \tag{5}$$

Therefore, inserting (3)–(5) into (2) yields

$$\begin{aligned} p\Big(\mathbf{u_I}(t), \mathbf{u_{II}}(t)\Big) &= \int p\Big(\mathbf{u_{II}}(t), \mathbf{u_I}(t) \,|\, \mathbf{u_I}(s \le t)\Big)\, p\Big(\mathbf{u_I}(s \le t)\Big)\, d\mathbf{u_I}(s \le t) \\ &= \lim_{L\to\infty} \frac{1}{L} \sum_{i=1}^{L} \delta\Big(\mathbf{u_I}(t) - \mathbf{u_I^i}(t)\Big)\, p\Big(\mathbf{u_{II}}(t) \,|\, \mathbf{u_I^i}(s \le t)\Big) \end{aligned} \tag{6}$$

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 (6) that is consistent with solving the Fokker-Planck equation for the joint PDF.

Appendix 4: Rigorous analysis of the error.

The mean integrated squared error (MISE) of any estimated density $\hat{p}_t(\mathbf{u_I}, \mathbf{u_{II}})$ is the average L^2 distance to the true density $p_t(\mathbf{u_I}, \mathbf{u_{II}})$:

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

Theorem: Error estimation in two different methods.
Consider the following two ways of estimating the density p_t.

- $\tilde{p}_t$: Kernel density estimation for the **joint PDF**.
- $\hat{p}_t$: Hybrid method — Kernel density **+** conditional Gaussian estimation.

With the same L, the error in the bias

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

and in the variance

$$\frac{\tilde{p}_t \text{ Variance bound}}{\hat{p}_t \text{ Variance bound}} = \frac{H^{-\frac{N_{II}}{2}} C}{\mathbb{E}\sqrt{\det(\mathbf{R_{II}}(t))}^{-1}}.$$

Here $\mathbb{E}\sqrt{\det(\mathbf{R_{II}}(t))}$ **does not decrease as** L **but** H **does!!**

When H shrinks and N_{II} becomes large, $H^{-\frac{N_{II}}{2}}$ increases dramatically.

Theorem: Error estimation in two different methods.
(Continued)

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

$\hat{p}_t$: Hybrid method — Kernel density **+** conditional Gaussian estimation.

We have the following MISE estimations:

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

The error in the hybrid method **does not depend on $N_{\mathbf{II}}$**.
— Beating the curse of dimensionality in $\mathbf{u}_{\mathbf{II}}$!

Theorem: Error estimation in two different methods.
(Continued)

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

$\hat{p}_t$: Hybrid method — Kernel density **+** conditional Gaussian estimation.

We have the following MISE estimations:

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

The error in the hybrid method **does not depend on** $N_{\mathbf{II}}$.
— Beating the curse of dimensionality in $\mathbf{u}_{\mathbf{II}}$!

To reach the same accuracy, the sample number $\widetilde{L}$ used in $\tilde{p}_t$ and the sample number $\hat{L}$ used in $\hat{p}_t$ must satisfy

$$\widetilde{L} = \hat{L}^{1+\frac{N_{\mathbf{II}}}{4+N_{\mathbf{I}}}}.$$

$\widetilde{L}$ can be much larger than $\hat{L}$ especially when $N_{\mathbf{II}} :=\text{Dim}(\mathbf{u}_{\mathbf{II}})$ is large.

Theorem: Error estimation in two different methods.
(Continued)

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

$\hat{p}_t$: Hybrid method — Kernel density **+** conditional Gaussian estimation.

We have the following MISE estimations:

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

The error in the hybrid method **does not depend on** $N_{\mathbf{II}}$.
— Beating the curse of dimensionality in $\mathbf{u}_{\mathbf{II}}$!

To reach the same accuracy, the sample number $\widetilde{L}$ used in $\tilde{p}_t$ and the sample number $\hat{L}$ used in $\hat{p}_t$ must satisfy

$$\widetilde{L} = \hat{L}^{1+\frac{N_{\mathbf{II}}}{4+N_{\mathbf{I}}}}.$$

$\widetilde{L}$ can be much larger than $\hat{L}$ especially when $N_{\mathbf{II}} :=\text{Dim}(\mathbf{u}_{\mathbf{II}})$ is large.

Example:
$\text{Dim}(\mathbf{u}_{\mathbf{I}}) = 3$ and $\text{Dim}(\mathbf{u}_{\mathbf{II}}) = 7$.
$\hat{L} = 1,000$ samples in $\hat{p}_t$ requires $\widetilde{L} = 1,000,000$ samples in $\tilde{p}_t$.

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

Updating the full covariance matrix is expensive for very large dimensional systems.

In many multiscale or multilevel dynamical systems, the state variables under certain conditions can be decomposed as

$$\mathbf{u} = \bigcup_k \mathbf{u}_k, \qquad \mathbf{u}_k = (\mathbf{u}_{\mathbf{I},k}, \mathbf{u}_{\mathbf{II},k}),$$

- the evolution of $\bar{\mathbf{u}}_{\mathbf{II},k}$ is fully coupled with that of all other $\bar{\mathbf{u}}_{\mathbf{II},k'}$, and
- 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 decompositions (Chen & Majda, *PNAS*, 2017)

Updating the full covariance matrix is expensive for very large dimensional systems.

In many multiscale or multilevel dynamical systems, the state variables under certain conditions can be decomposed as

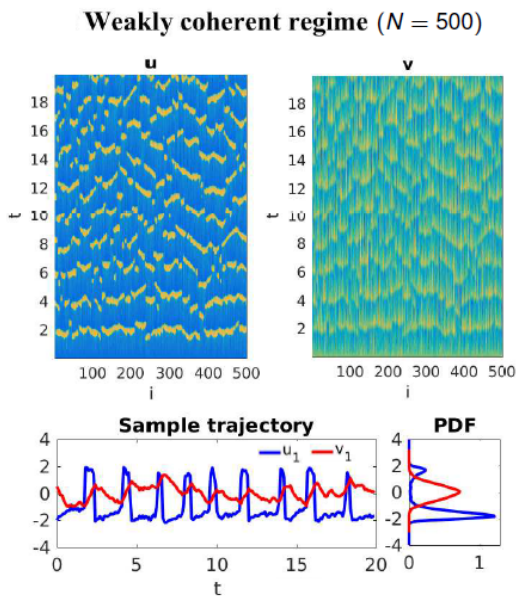
$$\mathbf{u} = \bigcup_k \mathbf{u}_k, \qquad \mathbf{u}_k = (\mathbf{u}_{\mathrm{I},k}, \mathbf{u}_{\mathrm{II},k}),$$

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

Example: A stochastic coupled FitzHugh-Nagumo model (Lindner et al., 2004).

a prototype model in excitable media, describing activation and deactivation dynamics of spiking neurons by external input currents

$$\epsilon \frac{du_i}{dt} = u_i - \frac{1}{3}u_i^3 + d_u(u_{i+1} + u_{i-1} - 2u_i) - v_i + \sqrt{\epsilon}\delta_1 \dot{W}_{u_i},$$
$$\frac{dv_i}{dt} = u_i + a + \delta_2 \dot{W}_{v_i}, \qquad i = 1, \ldots, N. \qquad N = 500 \longrightarrow$$



With constant parameters and homogeneous initial values, the model satisfies statistical symmetry.

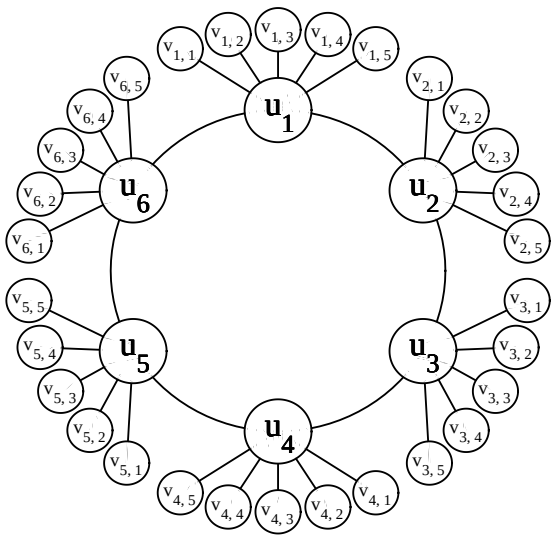
Due to the statistical symmetry, only L = 1 samples is needed in the algorithm to recover two-point statistics!

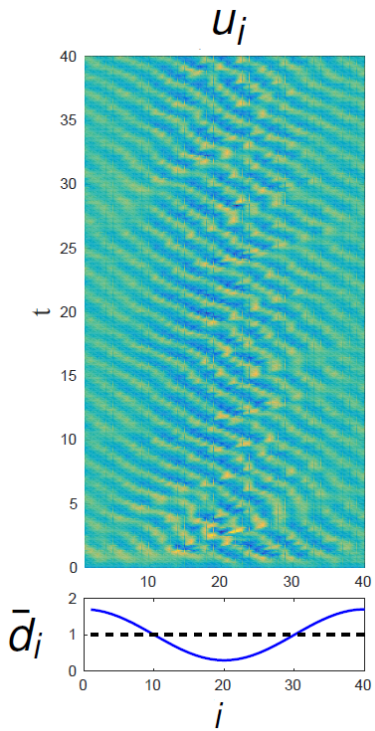
Example: A two-layer Lorenz 96 model (modified from Arnold, Moroz & Palmer, 2013)

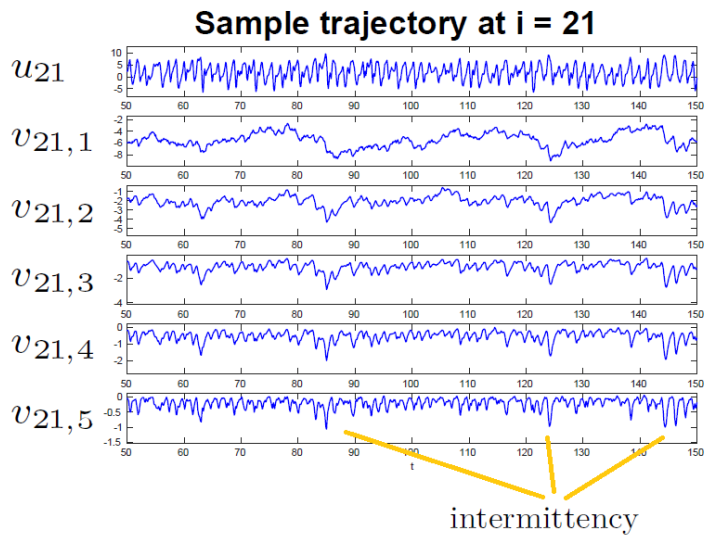
$$\frac{du_i}{dt} = u_{i-1}(u_{i+1} - u_{i-2}) + \sum_{j=1}^{J} \gamma_{i,j} u_i v_{i,j} - \bar{d}_i u_i + F + \sigma_u \dot{W}_{u_i}, \qquad i = 1, \ldots, I,$$

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

with $I = 40$ and $J = 5$. The total number of dimension is 240.





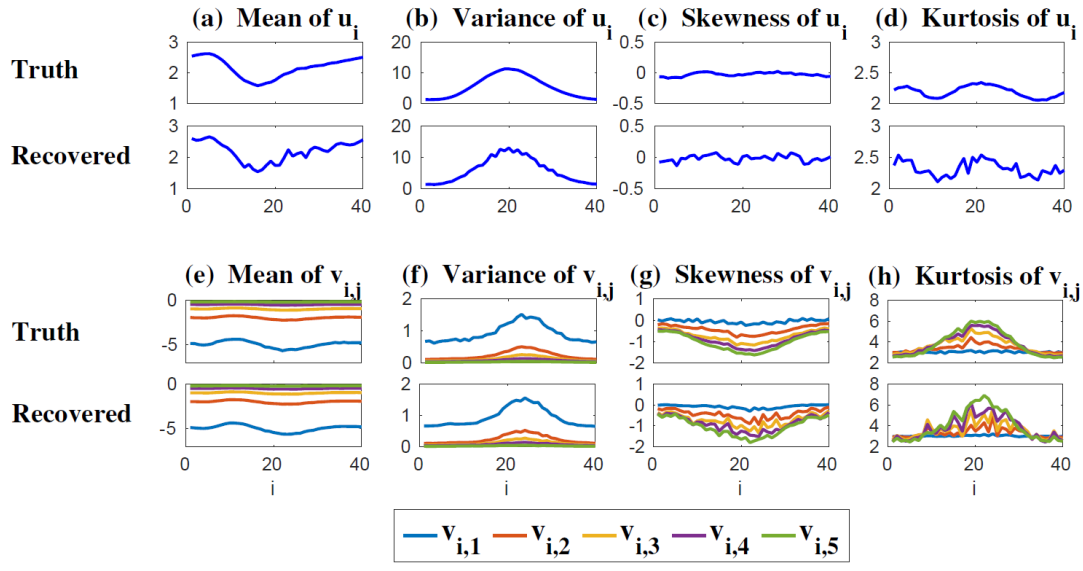


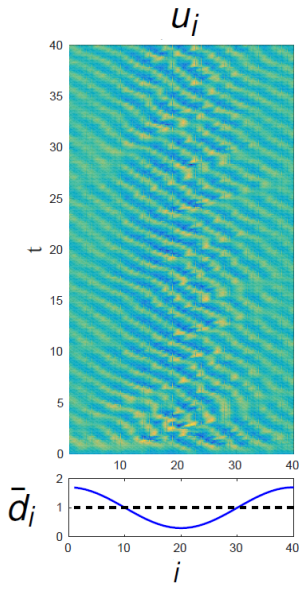
- $L = 500$ is used in the algorithm for recovering one-point and two-point statistics.
- As comparison, truth is generated using Monte Carlo with $L_{MC} = 150,000$.

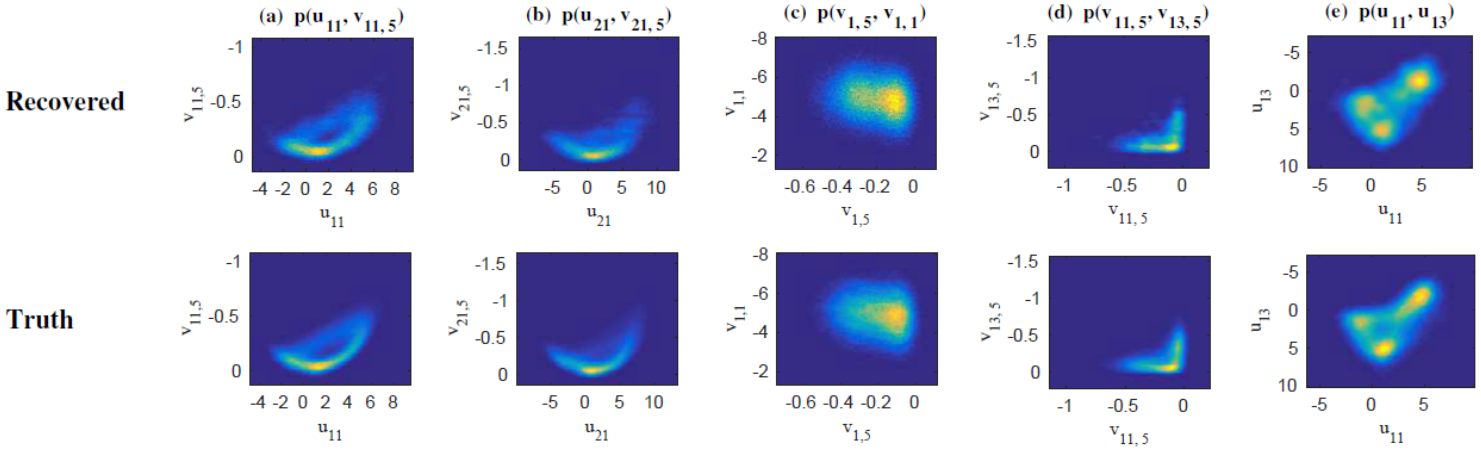
$$\frac{du_i}{dt} = u_{i-1}(u_{i+1} - u_{i-2}) + \sum_{j=1}^{J} \gamma_{i,j} u_i v_{i,j} - \bar{d}_i u_i + F + \sigma_u \dot{W}_{u_i},$$

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

$\sigma_{k,1}$	$\sigma_{k,2}$	$\sigma_{k,3}$	$\sigma_{k,4}$	$\sigma_{k,5}$
0.5	0.2	0.1	0.1	0.1
$d_{v_{k,1}}$	$d_{v_{k,2}}$	$d_{v_{k,3}}$	$d_{v_{k,4}}$	$d_{v_{k,5}}$
0.2	0.5	1	2	5
σ_u		$\gamma_{i,j} = \gamma_i$		
1		$0.1 + 0.025\cos(2\pi i/I)$		
F		$\bar{d}_i$		
8		$1 + 0.7\cos(2\pi i/I)$		







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

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

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

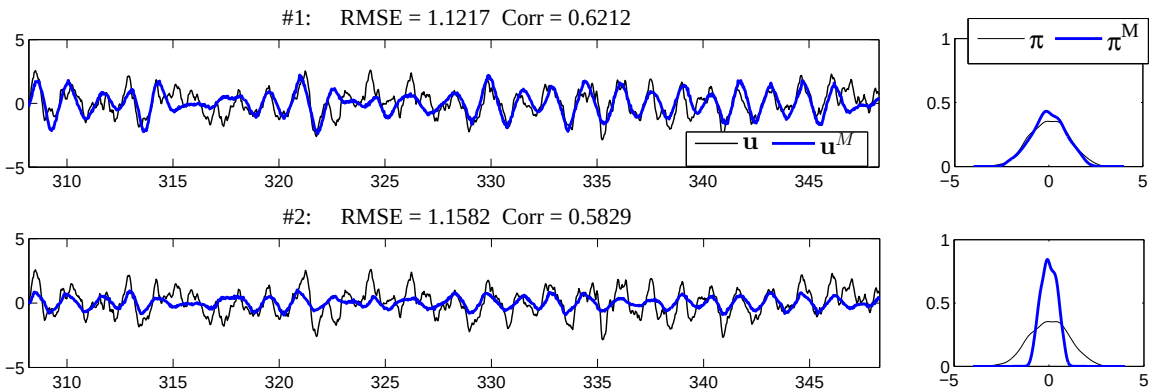
Linear models without $\underline{i\hat{v}_{\vec{k},B}}$ are used as imperfect forecast models such that the corresponding filters belong to the conditional Gaussian framework.

- 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

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