Capturing intermittent and low-frequency variability in high-dimensional data through nonlinear Laplacian spectral analysis

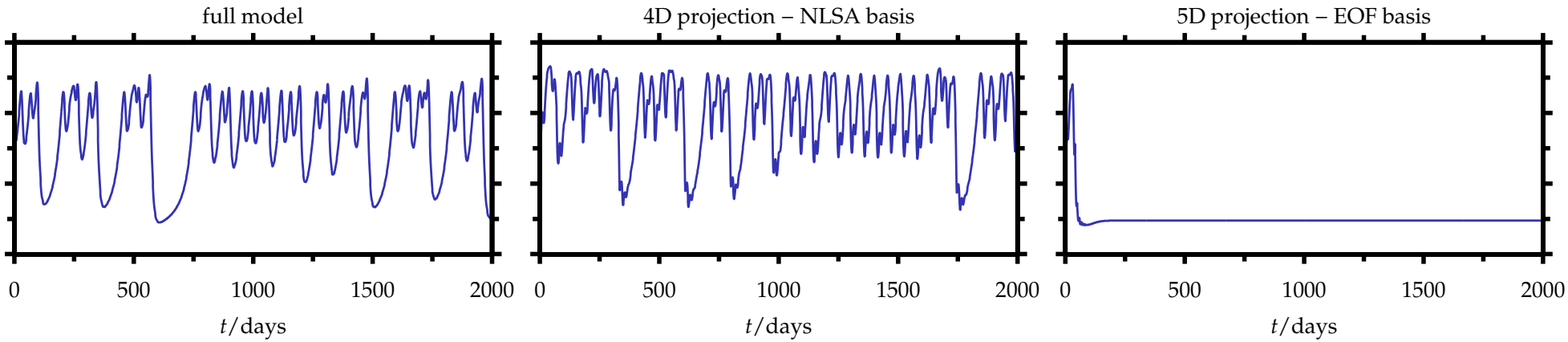
Dimitris Giannakis
Center for Atmosphere Ocean Science
Courant Institute of Mathematical Sciences, NYU

Adaptive Data Analysis and Sparsity
IPAM, 01/31/2013

Motivation

Decompose high-dimensional signal x_t into spatiotemporal patterns x_t^k to reveal intrinsically-nonlinear dynamical processes

$$x_t = \sum_k x_t^k$$



- *Dynamically-important modes are not necessarily those that carry high variance**

---

*Aubry et al. (1993), *SIAM J. Sci. Comput.*, 14, 483; Crommelin & Majda (2004), *J. Atmos. Sci.* 61, 2206

Outline

1. Nonlinear Laplacian spectral analysis (NLSA)
2. 6D chaotic ODE system
3. North Pacific SST variability in comprehensive climate models
4. Analysis of infrared brightness temperature satellite data for tropical dynamics

Acknowledgments

- Andrew Majda (Courant)

- Wen-wen Tung (Purdue)

1. Nonlinear Laplacian spectral analysis (NLSA)

Overview of NLSA algorithms

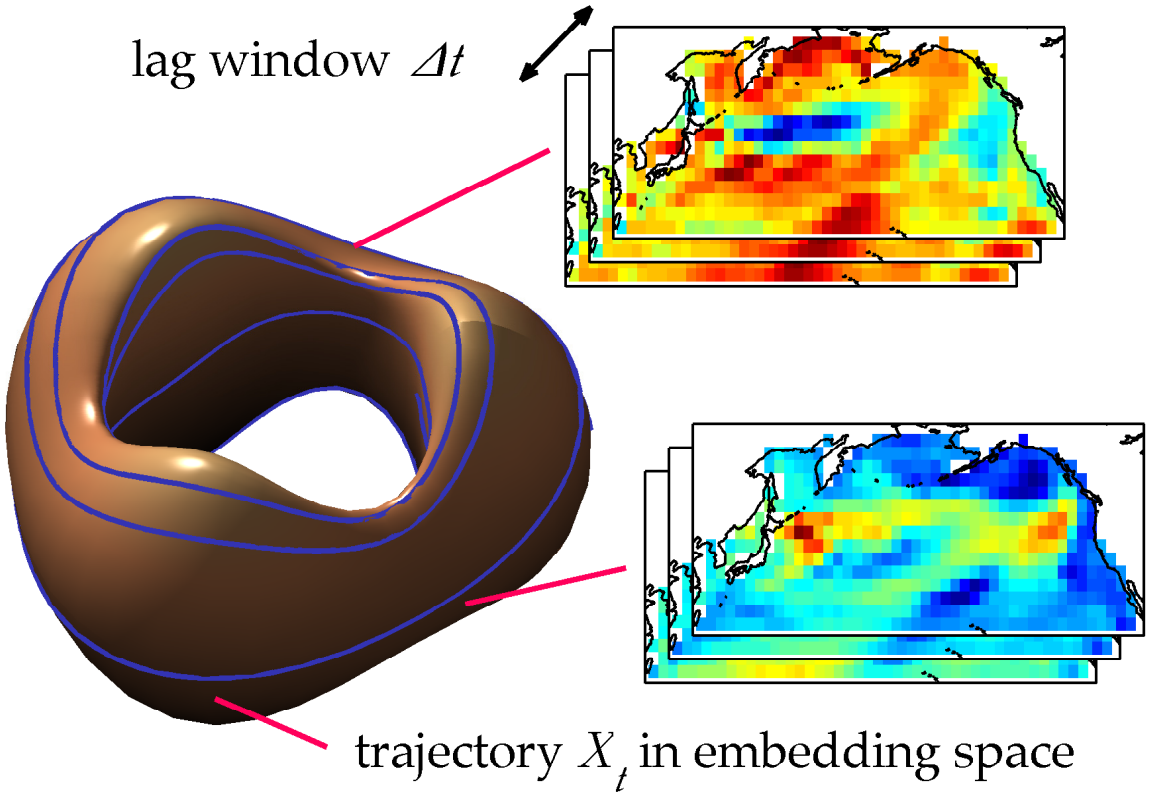
Method blends ideas from

- Qualitative analysis of dynamical systems
- Singular spectrum analysis (SSA)
- Spectral graph theory and nonlinear geometric methods for machine learning

Key ingredients:

- Laplace-Beltrami eigenfunctions as natural orthonormal basis to expand temporal patterns
- Time-adaptive weights to resolve rapid transitions
- Spatial and temporal modes extracted through singular value decomposition (SVD) of linear maps acting on functions on the nonlinear data manifold
- Temporal space dimension estimated via spectral entropy criteria

Time-lagged embedding



- Frequently, the observations x_t are incomplete
- Nearby x_t can actually be far apart in the phase space of the underlying dynamical system

$$x_t \mapsto X_t = (x_t, x_{t-\delta t}, x_{t-2\delta t}, \ldots, x_{t-\Delta t}) \in \mathbb{R}^N$$

- Embedding* "Markovianizes" the data, helping recover information lost by incomplete observations
- Each point in $\mathbb{R}^N$ corresponds to a spatiotemporal process of temporal extent Δt

*Broomhead & King (1986), *Phys. D*, 20, 217; Sauer et al. (1999), *J. Stat. Phys.*, 65, 579

Singular Spectrum Analysis (SSA)*

Data matrix of s observations:

$$X = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ X_0 & X_{\delta t} & \cdots & X_{(s-1)\,\delta t} \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}, \quad X_t \in \mathbb{R}^N$$

Singular value decomposition (SVD):

$$X = U\Sigma V^T, \quad u_i = \sigma_i^{-1} X v_i$$

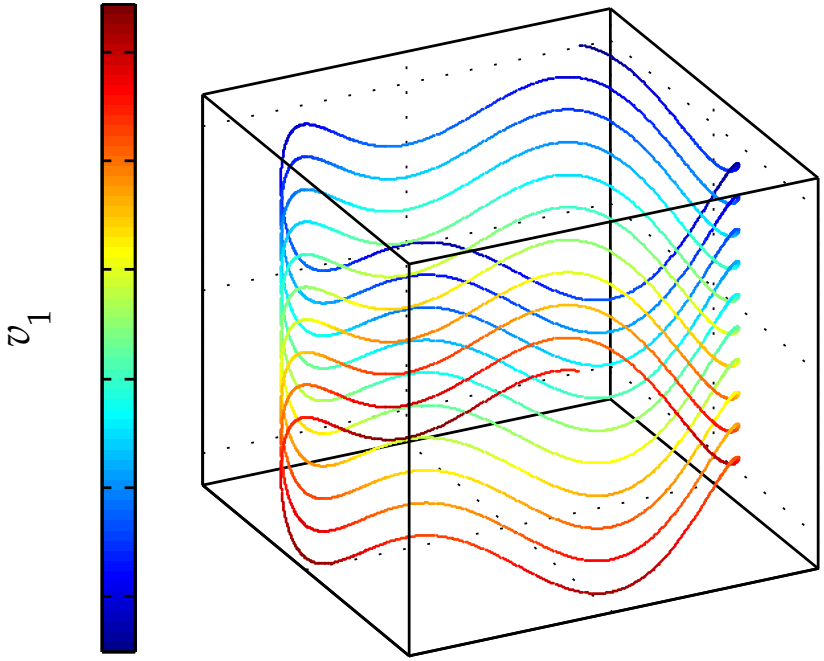
$$U = \begin{pmatrix} \uparrow & & \uparrow \\ u_1 & \cdots & u_n \\ \downarrow & & \downarrow \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_{\min\{n,s\}} \end{pmatrix}, \quad V = \begin{pmatrix} \uparrow & & \uparrow \\ v_1 & \cdots & v_s \\ \downarrow & & \downarrow \end{pmatrix}$$

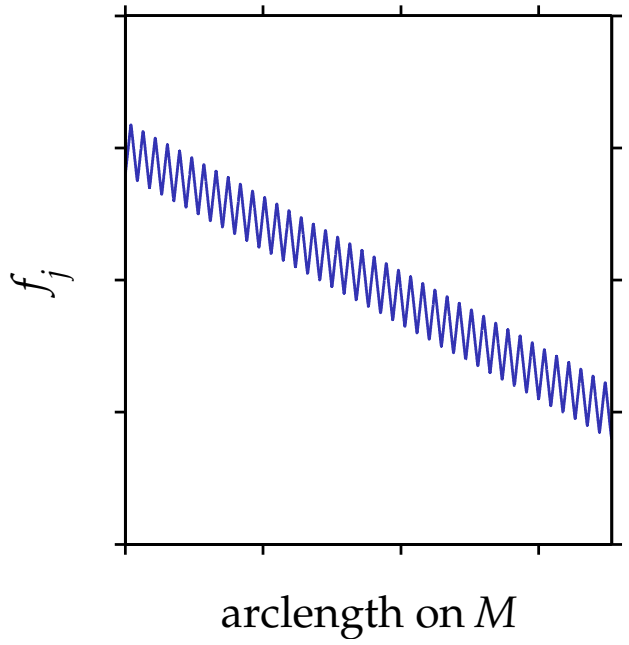
spatial patterns $u_i \in \mathbb{R}^n$ **singular values** $\sigma_i > 0$ **temporal patterns** $v_i \in \mathbb{R}^s$

- u_i are the principal axes of the ellipsoid associated with XX^T
- $\sigma_i v_i$ are linear projections of the data onto those axes

*Aubry et al. (1991) , *J. Stat. Phys.*, 64, 683; Ghil et al. (2002), *Rev. Geophys.*, 40, 1003

Spaces of temporal patterns



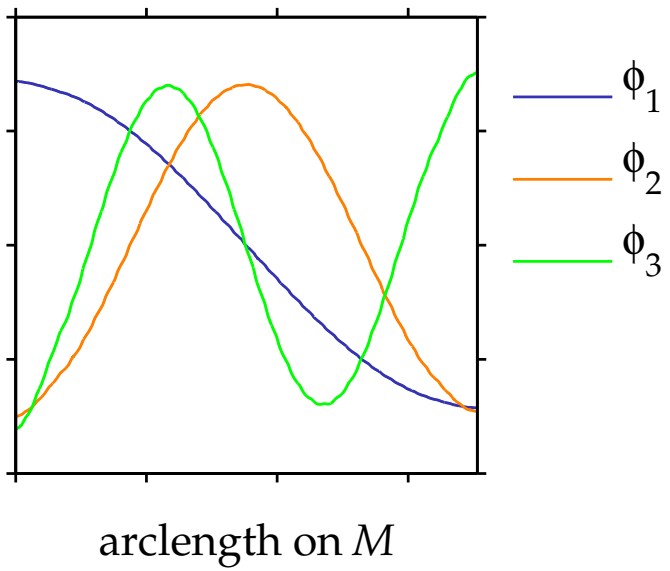


$$v_j = (v_{1j}, \ldots, v_{sj}) \quad \longleftrightarrow \quad f_j(X_{t_i}) = v_{ij} \qquad (*)$$

temporal patterns ⟷ **scalar functions on the nonlinear data manifold** M

- The f_j acquired through linear projections may develop oscillations not related to the intrinsic geometry of M
- But $(*)$ generalizes to spaces of well-behaved functions on M

Spaces of temporal patterns



- NLSA requires that temporal patterns lie in l-dimensional function spaces spanned by natural orthonormal basis functions on M

$$f(X_{t_i}) = \sum_{j=1}^{l} c_j \phi_j(X_{t_i})$$

$$\langle \phi_i, \phi_j \rangle = \sum_{k=1}^{s} \mu_k \phi_i(X_{t_k}) \phi_j(X_{t_k}) = \delta_{ij}$$

- ϕ_j are eigenfunctions of a graph Laplace-Beltrami operator on M

$$\Delta \phi_j(X_{t_i}) = \lambda_j \phi_j(X_{t_i}), \quad 0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots,$$

  and μ the corresponding Riemannian measure

- ϕ_j and μ can be computed efficiently in high-dimensional ambient spaces via algorithms developed in machine learning*

*Belkin & P. Niyogi (2003), *Neural Comput.*, 15, 1373; Coifman & Lafon (2006), *Appl. Comput. Harmon. Anal.*, 21, 5

Time-adaptive weights

- Local phase-space velocity $\xi_i = \|X_{t_i} - X_{t_{i-1}}\|$
- Use ξ_i to tune the edge weights for the LB eigenfunctions

$$W_{ij} = \exp\left(-\frac{\|X_{t_i} - X_{t_j}\|^2}{\epsilon \xi_i \xi_j}\right), \quad P_{ij} = \frac{W_{ij}}{\sum_{k=1}^{s} W_{ik}}$$

- Rapid transitions with large ξ_i acquire larger Riemannian measure than in the uniform ξ case

$$\mu P = \mu, \quad P\phi_i = \lambda_i \phi_i$$

$$\mu_i = \sum_{j=1}^{s} W_{ij}$$

*Such states carry low variance (i.e., might not be detectable through classical PCA), yet may play an important dynamical role, e.g., in metastable regime transitions**

* Aubry et al. (1993), *SIAM J. Sci. Comput.*, 14, 483; Crommelin & Majda (2004), *J. Atmos. Sci.* 61, 2206

Singular value decomposition

- Family of linear maps $A^l : V_l \mapsto \mathbb{R}^N$ with

$$A^l_{ik} = \sum_{j=1}^{s} \mu_j X_{i,t_j} \phi_k(X_{t_j})$$

s.t. $y = A^l(f)$ with

$$y = (y_1, \ldots, y_N), \quad y_i = \sum_{k=1}^{l} A^l_{ik} c_k, \quad f(X_{t_j}) = \sum_{k=1}^{l} c_k \phi_k(X_{t_j}),$$

[recall $y_i = \sum_j X_{i,t_j} f(X_{t_j})$ in SSA]

- Singular value decomposition

$$A^l_{ik} = \sum_{r=1}^{l} u_{ir} \sigma^l_r v_{kr}$$

[A^l_{ik} forms an $N \times l$ matrix; cf. $N \times s$ in SSA]

Data reconstruction

- Spatiotemporal patterns in embedding space

$$\tilde{X}_{t_j} = \sum_{k=1}^{l} X^k_{t_j}, \quad X^k_{t_j} = u_k \sigma^l_k \sum_{r=1}^{k} v_{rk} \phi_r(X_{t_j})$$

- $\tilde{X}$ is identical to input dataset X for $l = s$. In applications, $l \ll s$
- Uniform-weight projection to physical space

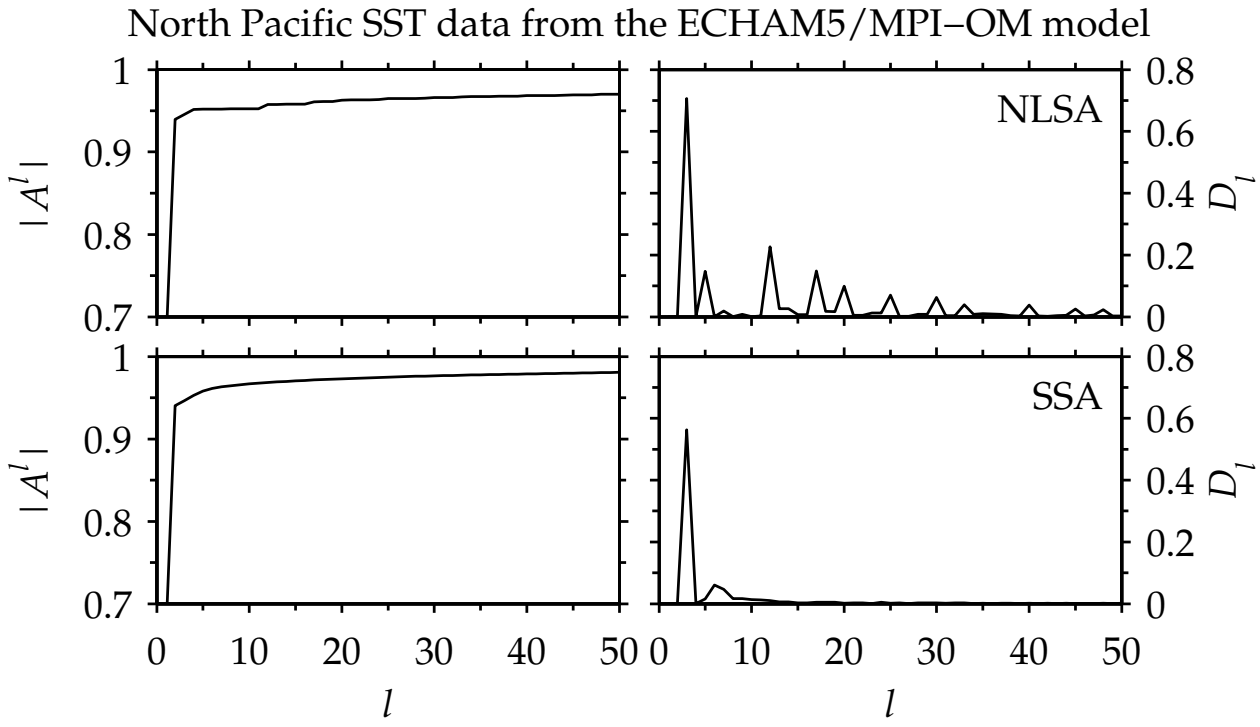
$$X^k = \begin{pmatrix} \uparrow & & \uparrow \\ X_{t_1} & \cdots & X_{t_s} \\ \downarrow & & \downarrow \end{pmatrix} = \begin{pmatrix} \hat{x}_{1,1} & \cdots & \hat{x}_{1,s} \\ \vdots & \ddots & \vdots \\ \hat{x}_{q,1} & \cdots & \hat{x}_{q,s} \end{pmatrix}$$

$$X^k_{t_j} \mapsto x^k_{t_j} = \sum_{i,i' : t_i - t_{i'} = t_j} \hat{x}_{i,i'} / q.$$

- Decomposition of the input signal:

$$\tilde{x}_t = \sum_{k=1}^{l} x^k_t$$

Setting the temporal space dimension



$$p_i^l = \frac{(\sigma_i^l)^2}{\sum_{j=1}^{l}(\sigma_j^l)^2}, \quad \pi_i^{l+1} = \frac{\hat{\sigma}_i^2}{\sum_{j=1}^{l}\hat{\sigma}_j^2}, \quad \hat{\sigma}_i = \begin{cases} \sigma_i^l, & i \le l \\ \sigma_{i-1}^l, & i = l+1 \end{cases}$$

spectral entropy measure $D_l = \sum_{i=1}^{l} p_i^{l+1} \log \frac{p_i^{l+1}}{\pi_i^{l+1}}$

- **Practical criterion:** Stop increasing $l = \dim(V_l)$ when D_l becomes small

2. 6D chaotic ODE system

Low-order model for atmospheric regime behavior

$$\dot{x} = \mathcal{F}(x) = F + A(x) + B(x, x)$$

- Model derived by Galerkin projection of the barotropic vorticity equation*
- x: vector of leading Fourier components of the streamfunction
- F: forcing
- A: linear operator
- B: energy- and enstrophy-preserving quadratic nonlinearity

$$\dot{x}_1 = \tilde{\gamma}_1 x_3 - C(x_1 - x_1^*)$$
$$\dot{x}_2 = -(\alpha_1 x_1 - \beta_1)x_3 - Cx_2 - \delta_1 x_4 x_6$$
$$\dot{x}_3 = (\alpha_1 x_1 - \beta_1)x_2 - \gamma_1 x_1 - Cx_3 + \delta_1 x_4 x_5$$
$$\dot{x}_4 = \tilde{\gamma}_2 x_6 - C(x_4 - x_4^*) + \epsilon(x_2 x_6 - x_3 x_5)$$
$$\dot{x}_5 = -(\alpha_2 x_1 - \beta_2)x_6 - Cx_5 - \delta_2 x_4 x_3$$
$$\dot{x}_6 = (\alpha_2 x_1 - \beta_2)x_5 - \gamma_2 x_4 - Cx_6 + \delta_2 x_4 x_2$$

x^*	zonal forcing
C	damping
α	advection
β	Coriolis β effect
γ, $\tilde{\gamma}$	topographic height
δ, ϵ	nonlinear triad interaction

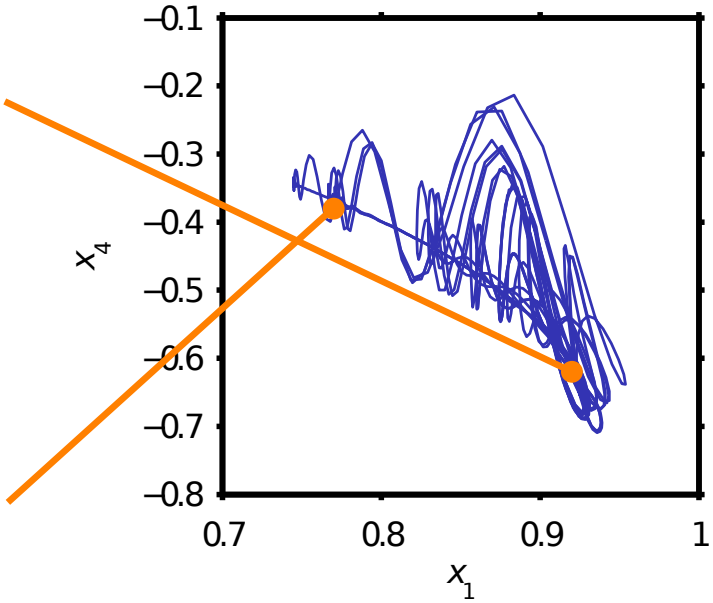
*De Swart (1989), *Phys. D*, 36, 222; Crommelin et al. (2004), *J. Atmos. Sci.*, 61, 1406; Crommelin & Majda (2004), *J. Atmos. Sci.*, 61, 2206

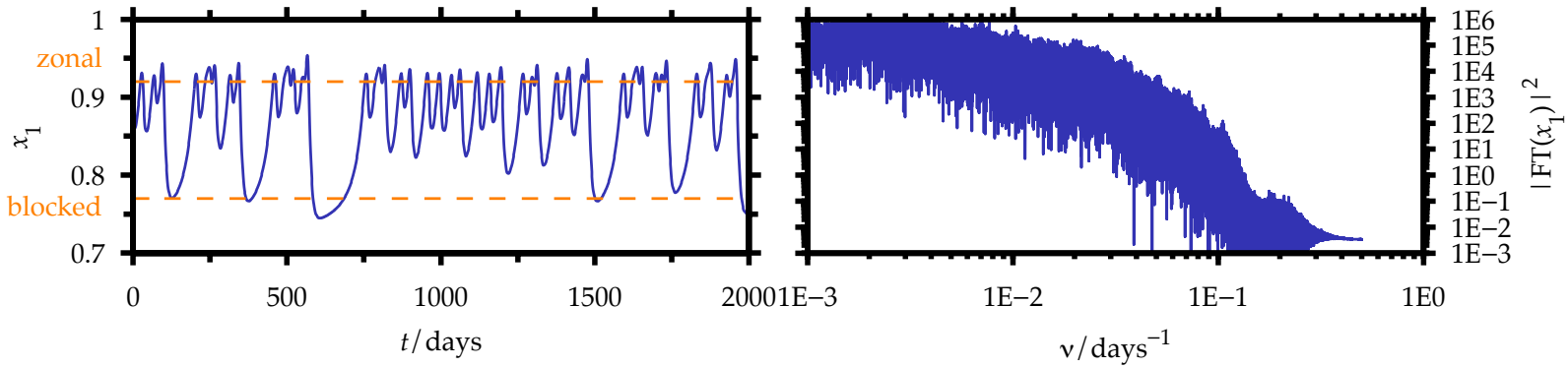
Atmospheric regimes

zonal

Crommelin et al. (2004), J. Atmos. Sci., 61, 1406.

blocked





Chaotic itineracy: Irregular transitions between *zonal* and *blocked* regimes due to topographic and barotropic instability

Dimensional reduction

Full model

$$\dot{x} = \mathcal{F}(x),$$
$$x = (x_1, \ldots, x_6)$$

Objective

Find spatial patterns u_i such that the reduced model exhibits *chaotic regime behavior*

- Not possible using PCA or Optimal Persistence Patterns*

Reduced model

Orthogonal spatial patterns:

$$U = \begin{pmatrix} \uparrow & & \uparrow \\ u_1 & \cdots & u_K \\ \downarrow & & \downarrow \end{pmatrix},$$

$$u_i \in \mathbb{R}^6, \quad K < 6, \quad U^T U = I_{K\times K}$$

Bare truncation:

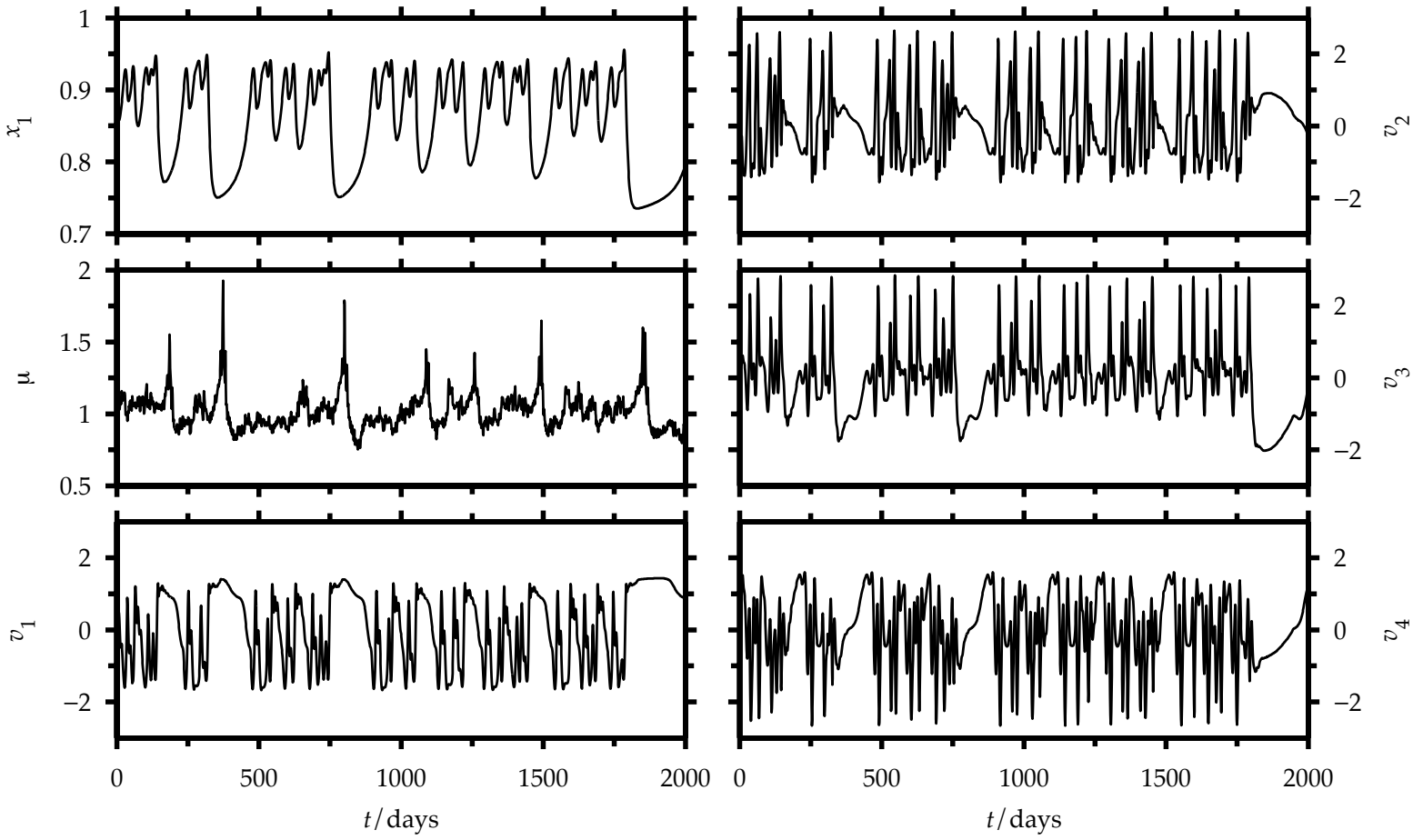
$$x_i = \sum_{j=1}^{K} U_{ij} z_j$$

Governing equations:

$$\dot{z} = U^T \mathcal{F}(zU)$$

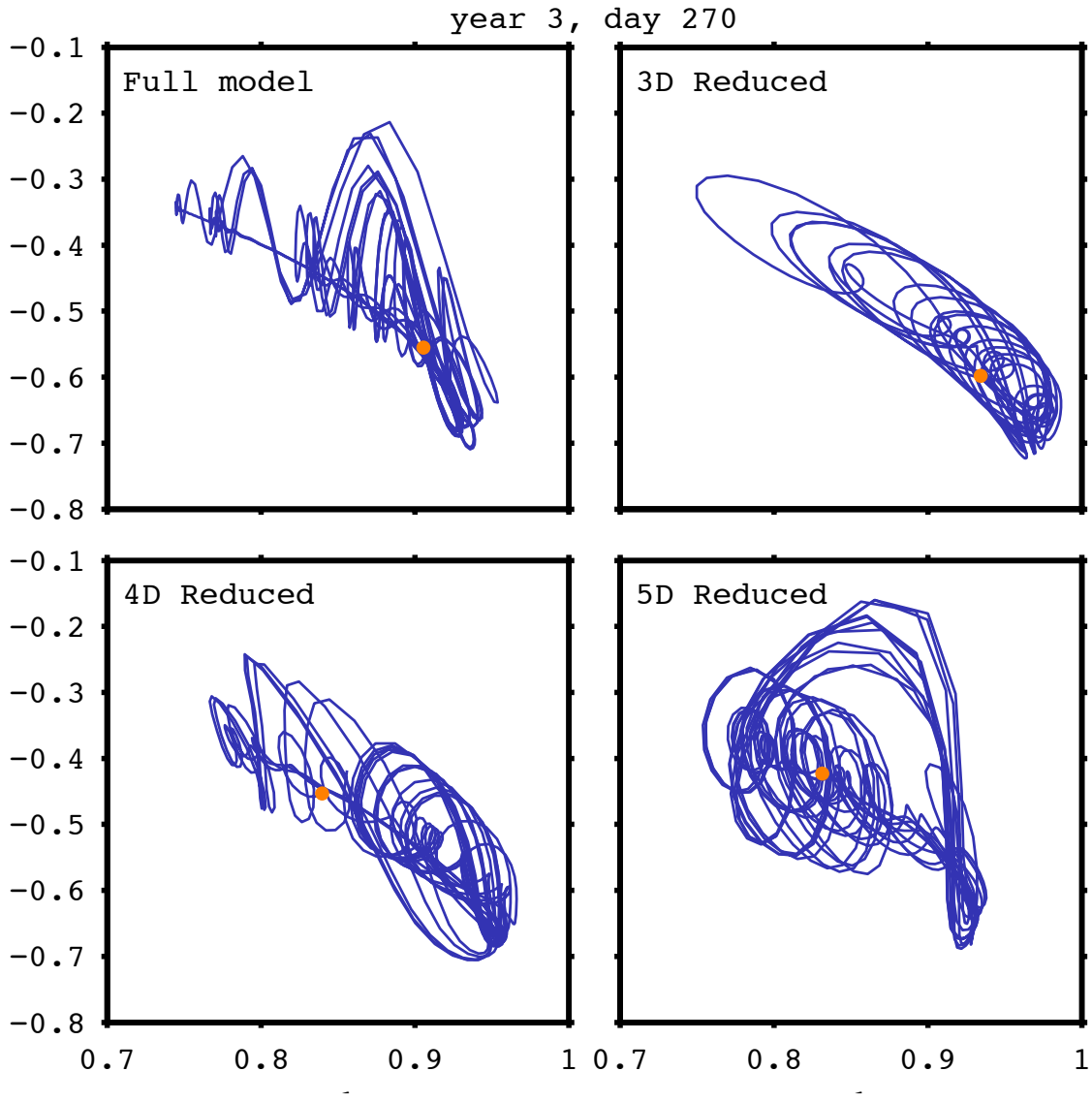
*Crommelin & Majda (2004), *J. Atmos. Sci.*, 61, 2206

Temporal patterns & the Riemannian measure



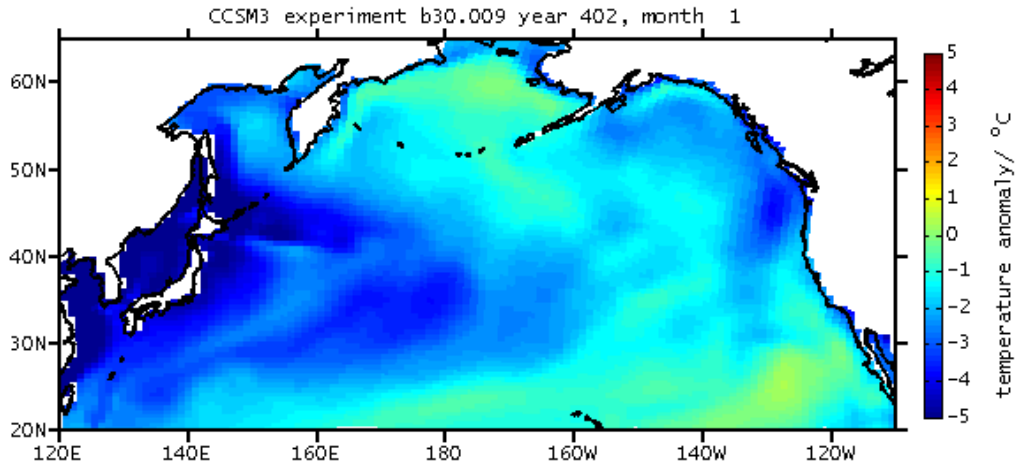
The Riemannian measure μ (volume) is large during transitions between metastable regimes

- Enhanced capability of capturing rare events


year 3, day 270
Full model
3D Reduced
4D Reduced
5D Reduced
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
0.7
0.8
0.9
1

3. North Pacific SST variability in comprehensive climate models

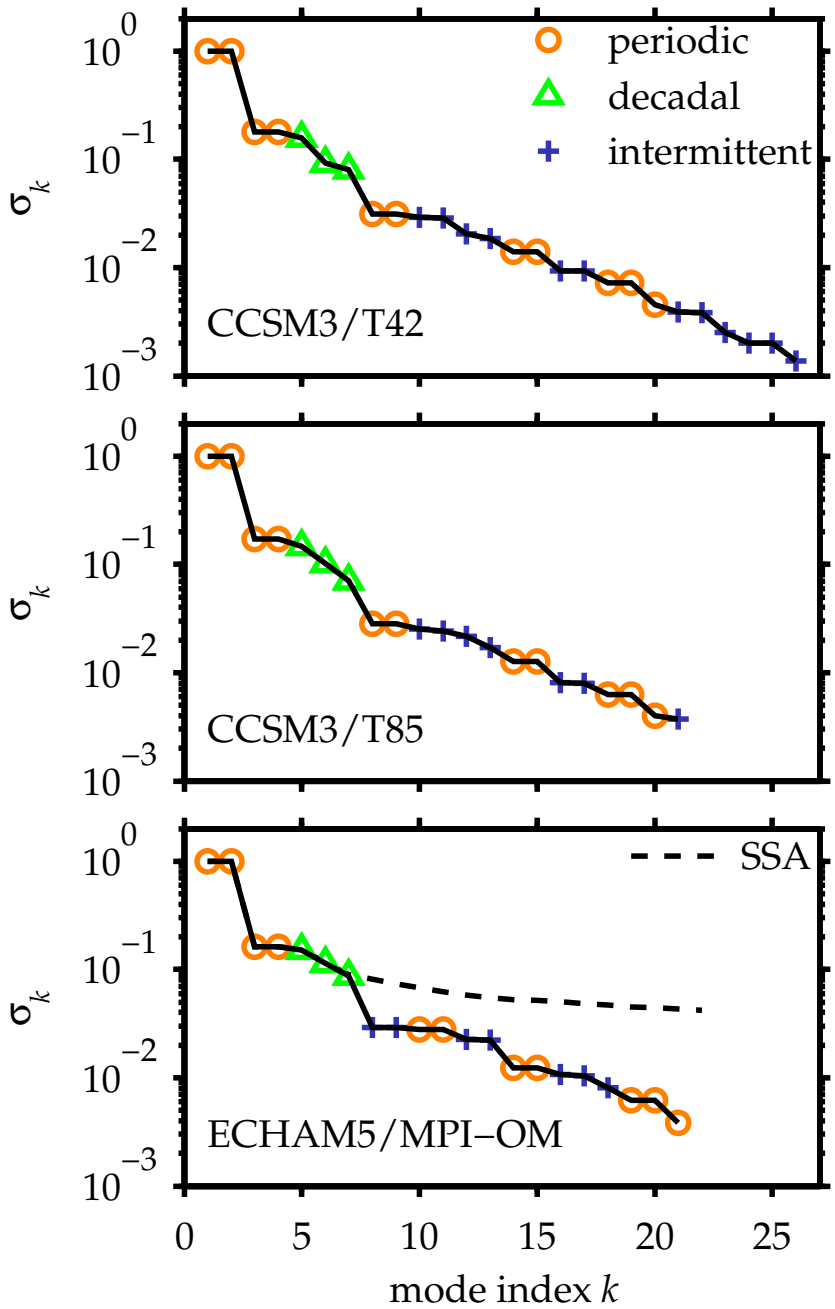
North Pacific SST data sets



- Monthly-averaged sea surface temperature (SST) field from extended control runs in IPCC4 AR4 with fixed greenhouse forcings
- Two-year lagged embedding window used throughout

Model	CCSM3	CCSM3	ECHAM5/ MPI-OM
Atmosphere grid	T42	T85	T63
Ocean grid resoln.	1°	1°	1.5°
Temporal extent (yr)	900	450	500
# gridpoints in data set	6671	6671	1204
Embedding space dim.	160,104	160,104	28,896
NLSA temporal space dim. l	26	21	21

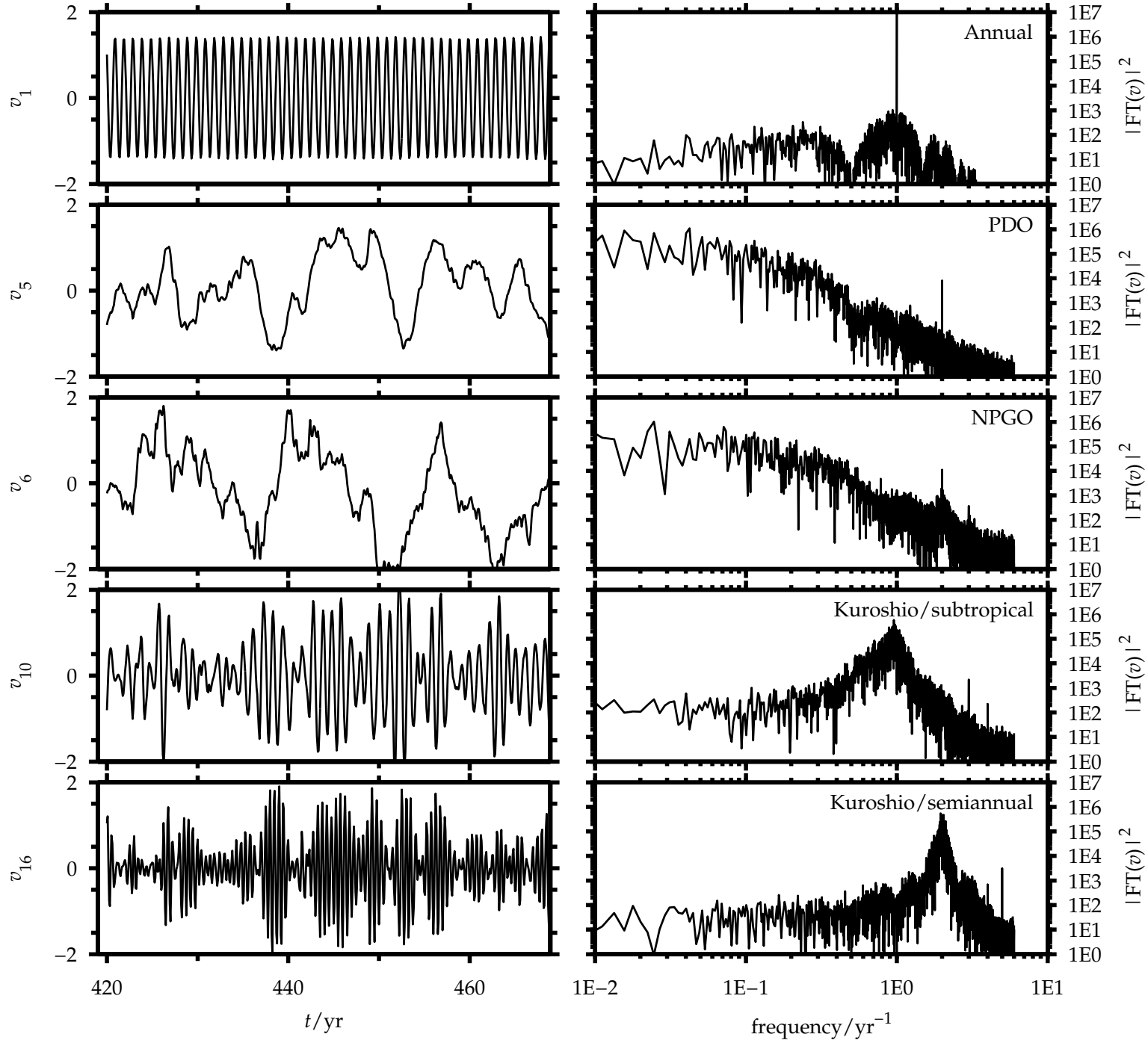
Families of spatiotemporal modes

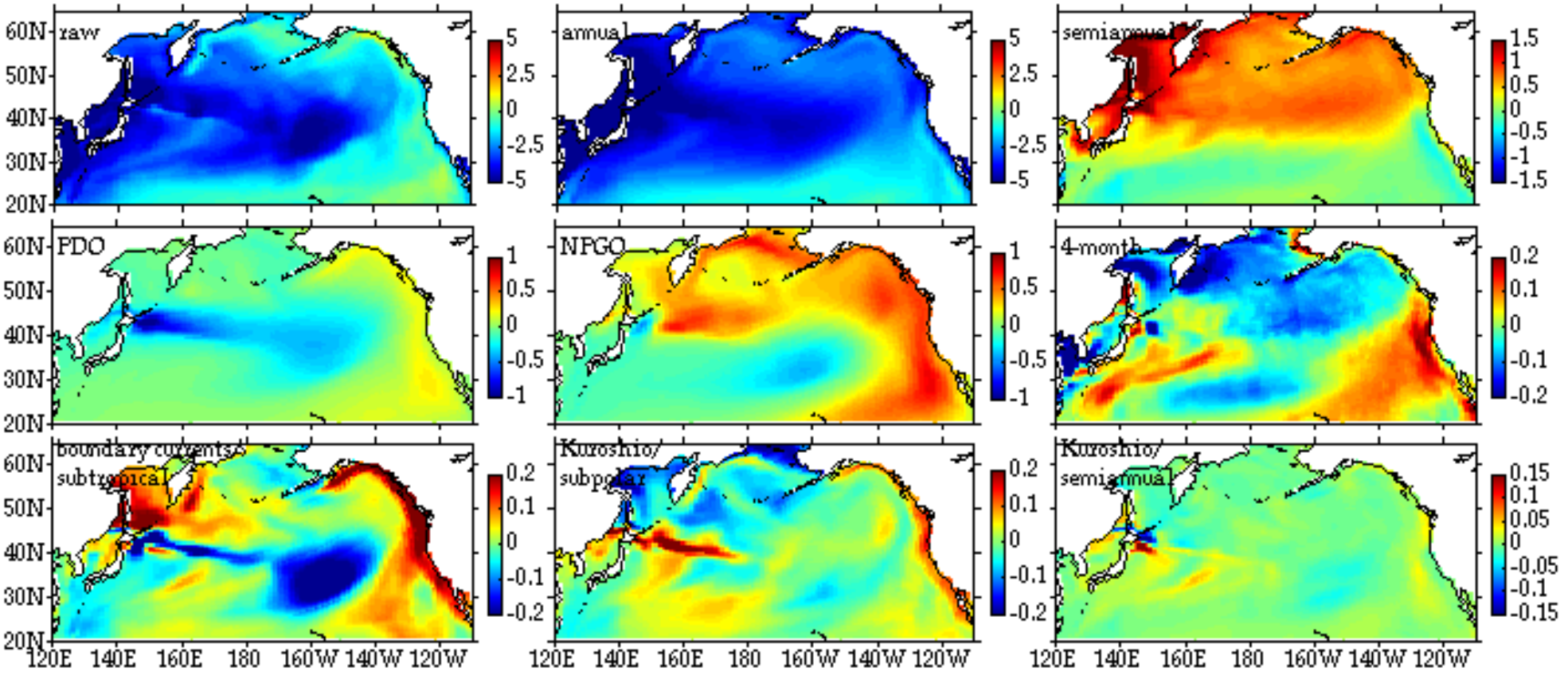


- **Periodic**
  - Annual, semiannual, 4-month, etc
- **Low-frequency**
  - Pacific Decadal Oscillation*
  - North Pacific Gyre Oscillation†
  - El Niño signal in the North Pacific
- **Intermittent**
  - Boundary currents (Kuroshio, Alaska, California)
  - Subtropical gyre

*Mantua & Hare (2002), *J. Oceanogr.*, 58, 35

†Di Lorenzo et al. (2008), *Geophys. Res. Lett.*, 35, L08, 607


Annual
PDO
NPGO
Kuroshio/subtropical
Kuroshio/semiannual
v_1
v_5
v_6
v_{10}
v_{16}
2
0
−2
1E7
1E6
1E5
1E4
1E3
1E2
1E1
1E0
$|\mathrm{FT}(v)|^2$
420
440
460
t/yr
1E−2
1E−1
1E0
1E1
frequency/yr^{-1}


raw
annual
semiannual
PDO
NPGO
4-month
boundary currents/
subtropical
Kuroshio/
subpolar
Kuroshio/
semiannual
60N
50N
40N
30N
20N
120E
140E
160E
180
160W
140W
120W
5
2.5
0
-2.5
-5
1.5
1
0.5
-0.5
-1
-1.5
0.2
0.1
-0.1
-0.2
0.15
0.05
-0.05

Regression modeling with external factors

Objective

- Given the values of the temporal modes $v_i(t)$ at time t, predict the values at time $t + \delta t$:

$$v_i(t + \delta t) = f_i(\{v_j(t)\}) + \sigma_i \epsilon_i(t)$$

Model selection

- How to choose the model structure f_i?
- Nonlinearities in f_i can lead to blowups

Limits of predictability

- Identify subsets of modes $\{v_j(t)\}$ which, if prescribed, explain the time-evolution of $v_i(t)$ with high fidelity

NLSA modes used for regression modeling

- **Periodic modes**
  - Seasonal cycle, $(P_1, P_2) \leftrightarrow (v_1, v_2)$
- **Low-frequency modes**
  - Pacific Decadal Oscillation, $L_1 \leftrightarrow v_5$
  - North Pacific Gyre Oscillation, $L_2 \leftrightarrow v_6$
  - ENSO signal in the North Pacific, $L_3 \leftrightarrow v_7$
- **Intermittent modes**
  - Boundary currents & subtropical gyre, $(I_1, I_2) \leftrightarrow (v_{10}, v_{11})$
  - Boundary currents & subpolar gyre, $(I_3, I_4) \leftrightarrow (v_{12}, v_{13})$

Model structure

Low-frequency modes

$$L_i(t+\delta t) = \mathbf{A}_i^T \mathbf{L}_i(t) + \mathbf{B}_i^T [\mathbf{P} * \mathbf{I}]_i(t) + \mathbf{C}_i^T [\mathbf{L} * \mathbf{L} * \mathbf{L}]_i(t) + \sigma_i \epsilon_i(t)$$

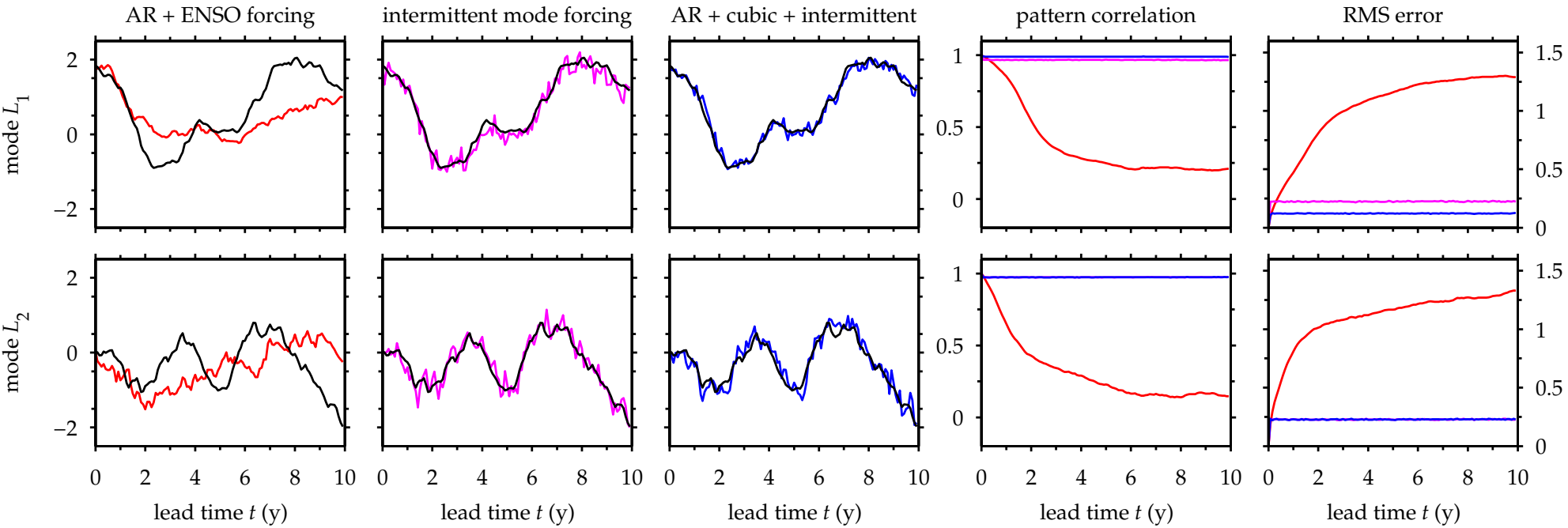
- Periodic and intermittent modes prescribed as external factors
- Products $[\mathbf{P} * \mathbf{I}]_i(t)$ demodulate the intermittent modes to produce low-frequency forcings

Intermittent modes

$$I_i(t+\delta t) = \mathbf{A}_i^T \mathbf{I}_i(t) + \mathbf{B}_i^T [\mathbf{P} * \mathbf{L}]_i(t) + \mathbf{C}_i^T [\mathbf{I} * \mathbf{I} * \mathbf{I}]_i(t) + \sigma_i \epsilon_i(t)$$

- Periodic and low-frequency modes prescribed as external factors
- Products $[\mathbf{P} * \mathbf{L}]_i(t)$ modulate the periodic modes to produce intermittent-type forcings

Hindcast skill – low-frequency modes

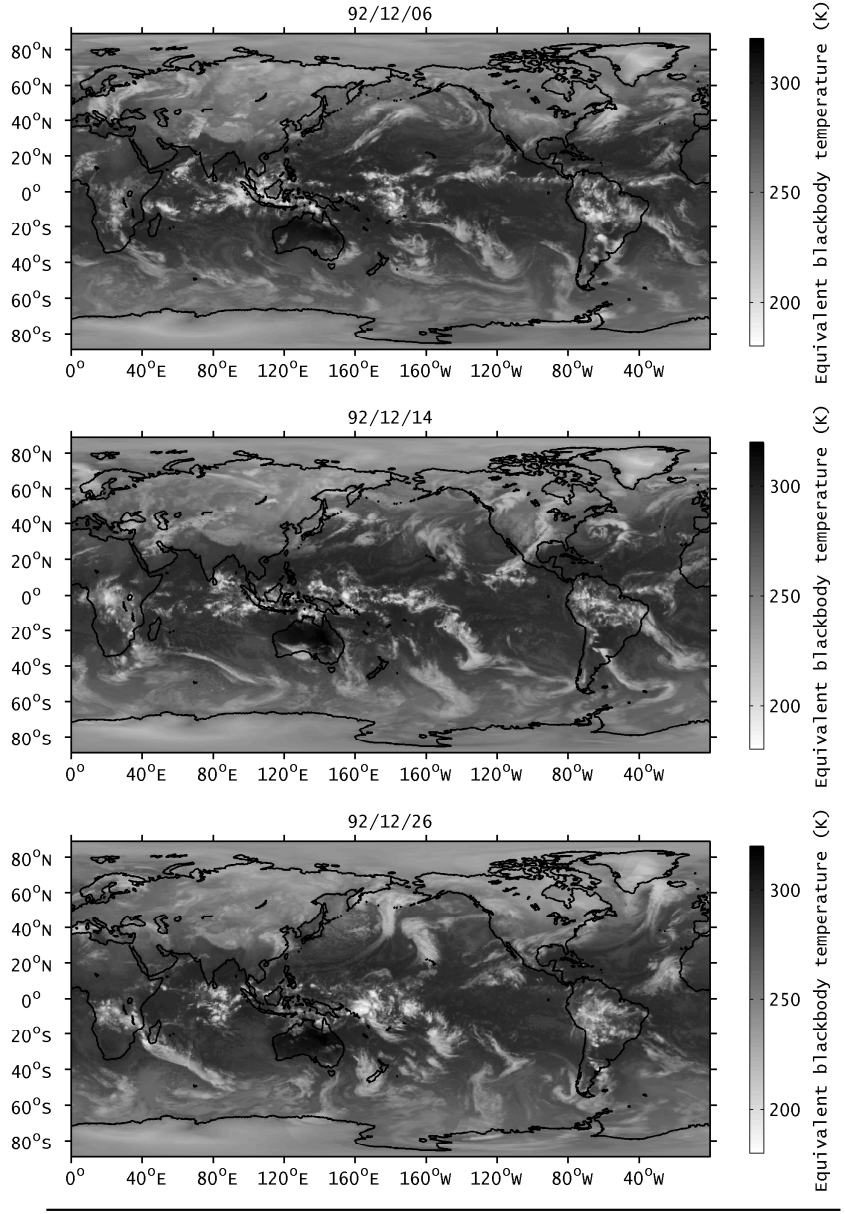


- Predictability of large-scale modes like the PDO and NPGO is limited by features of the North Pacific associated with the intermittent modes (e.g., the Kuroshio current), without direct reference to ENSO signals
- Cf. ENSO-forced autoregressive models for low-frequency variability of SST*

*Newman et al. (1993), *J. Clim.*, 16, 3853

5. Analysis of infrared brightness temperature satellite data

Infrared brightness temperature satellite data

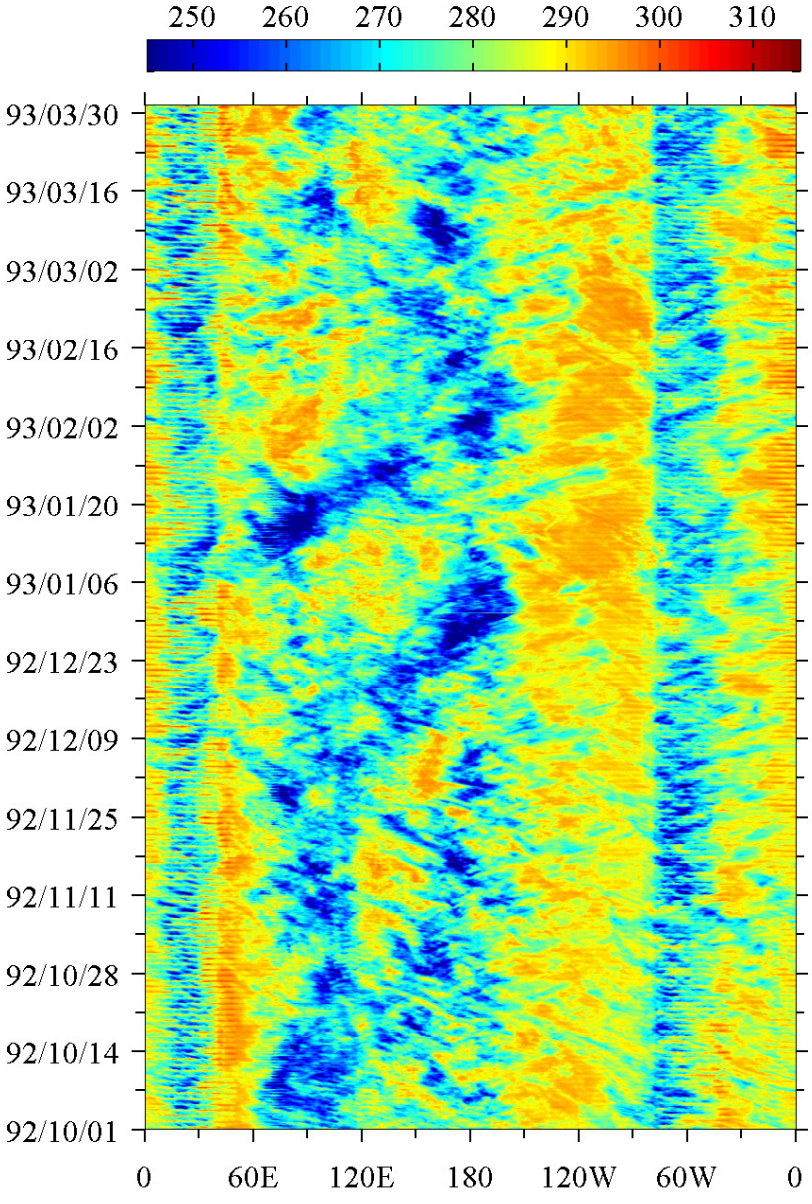


- **C**loud **A**rchive **U**ser **S**ervice (CLAUS) multi-satellite infrared brightness temperature (*Tb*)*
- High (low) *Tb* corresponds to decreased (increased) cloudiness and convection
- Highly multiscaled spatiotemporal signal†
  - Madden-Julian Oscillation (MJO)
  - El Niño, Monsoon
  - Diurnal cycle
  - Convectively-coupled waves
  - Mesoscale convective systems

*Hodges et al. (2000), *J. Atmos. Ocean Tech.*, 17, 1296

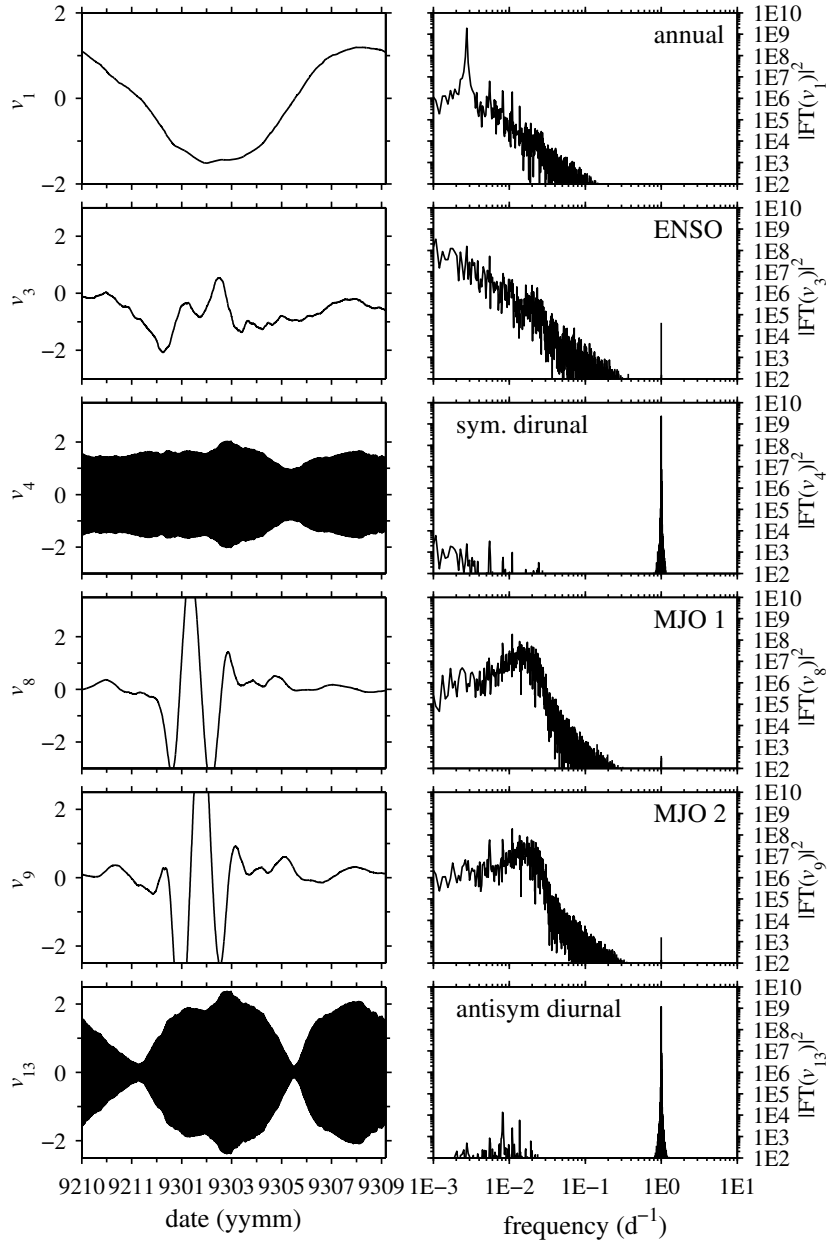
†Kiladis et al. (2009), *Rev. Geophys.*, 2008RG000266

Overview of analysis via NLSA

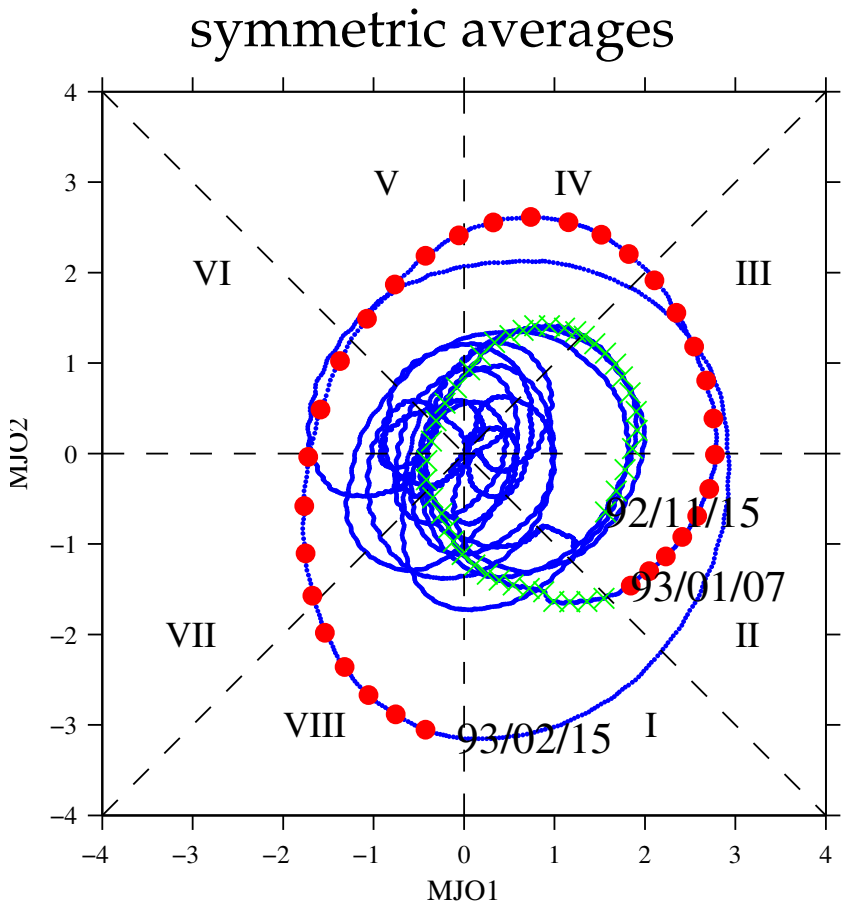


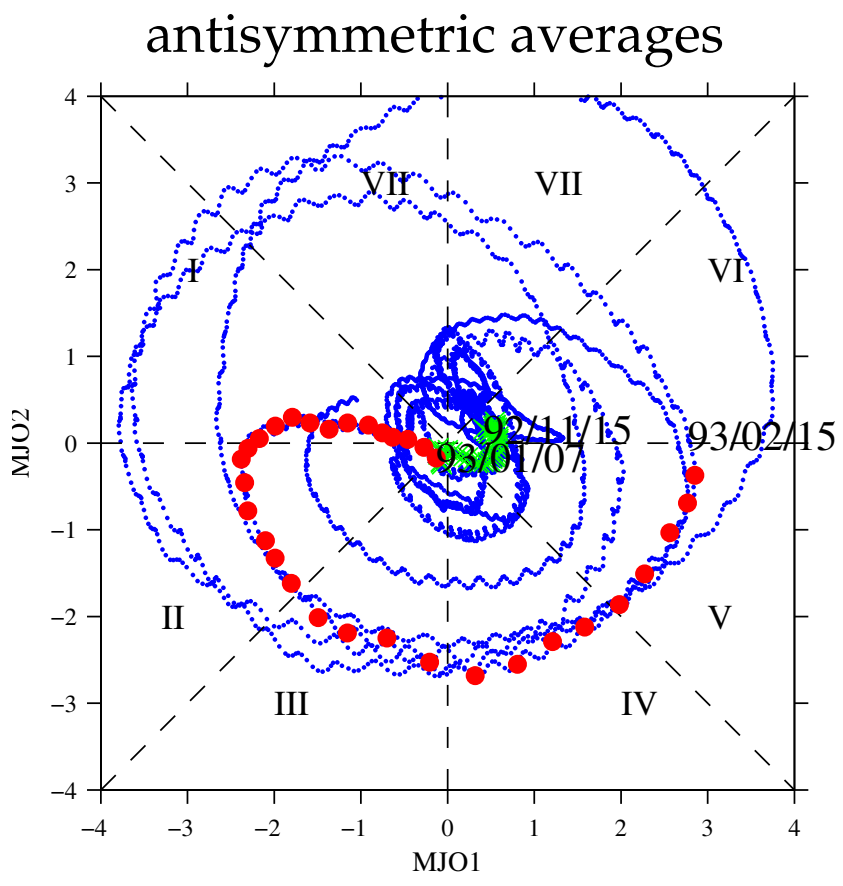
- 0.5° spatial resolution
- 3 h sampling interval
- $\sim$ 67,000 samples from 1983–2006
- Intraseasonal lagged-embedding window ($\Delta t = 64$ d)
- Embedding space dimension $\sim 2 \times 10^7$

Hierarchy of temporal patterns (right singular vectors)



MJO phase identification





■: November 1992 event ★: January 1993 event

Summary and outlook

- NLSA algorithms offer a way of capturing low-variance intermittent processes in high-dimensional datasets, which may be of high dynamical significance
- The intermittent modes of North Pacific SST variability, suitably modulated by multiplicative couplings to the annual cycle, can explain most of the variability of the prominent low-frequency North Pacific SST patterns despite carrying low variance
- Hierarchy of spatiotemporal patterns of tropical variability extracted from *Tb* satellite data

Ongoing & future work:

- Map previously unseen samples to the NLSA coordinates (initialization)
- Multi-component datasets (SST & sea ice)
- MJO predictability and initiation

References

- D. Giannakis, A. J. Majda (2012), Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability, *Proc. Natl. Acad. Sci.*, 109(7), 2222

- D. Giannakis and A. J. Majda (2012), Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear Laplacian spectral analysis, *Stat. Anal. Data. Min.*, in press

- D. Giannakis and A. J. Majda (2012), Limits of predictability in the North Pacific sector of a comprehensive climate model, *Geophys. Res. Lett.*, 39, L24602

- D. Giannakis W.-w. Tung, and A. J. Majda (2012), Hierarchical Structure of the Madden-Julian Oscillation in Infrared Brightness Temperature Revealed through Nonlinear Laplacian Spectral Analysis, 2012 NASA Conference on Intelligent Data Understanding