***Predictability + Data Assimilation,***
*Nordita*

Stockholm, 25 May 2011

# Empirical Model Reduction and Applications

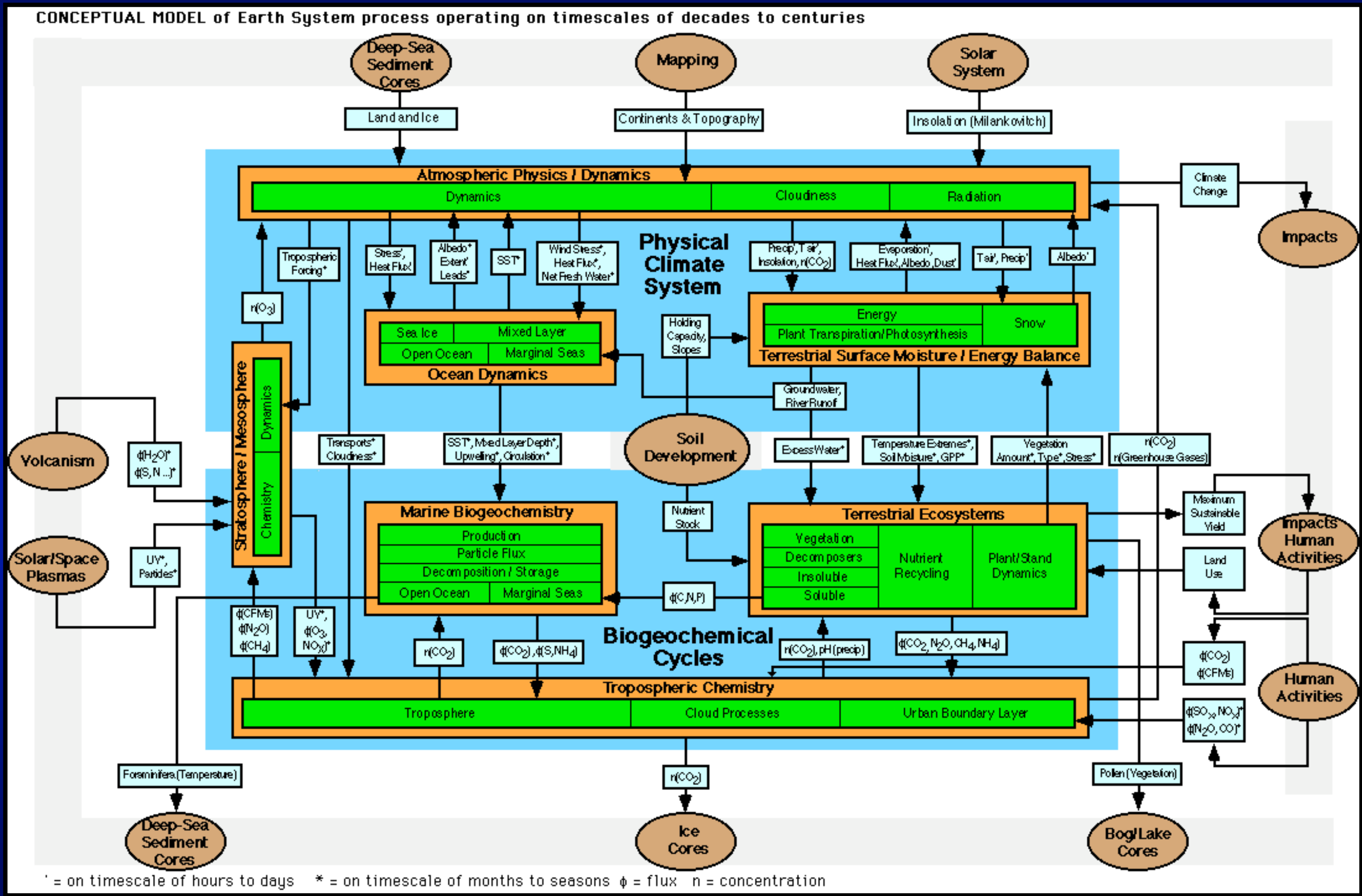

Michael Ghil

Ecole Normale Supérieure, Paris, and
University of California, Los Angeles


Joint work with D. Kondrashov & Y. Shprits (UCLA), S. Kravtsov (U. Wisconsin, Milwaukee), A. W. Robertson (IRI, Columbia U.) and K. Strounine (private sector)

http://www.atmos.ucla.edu/tcd/ and http://www.environnement.ens.fr/

# F. Bretherton's "horrendogram" of Earth System Science



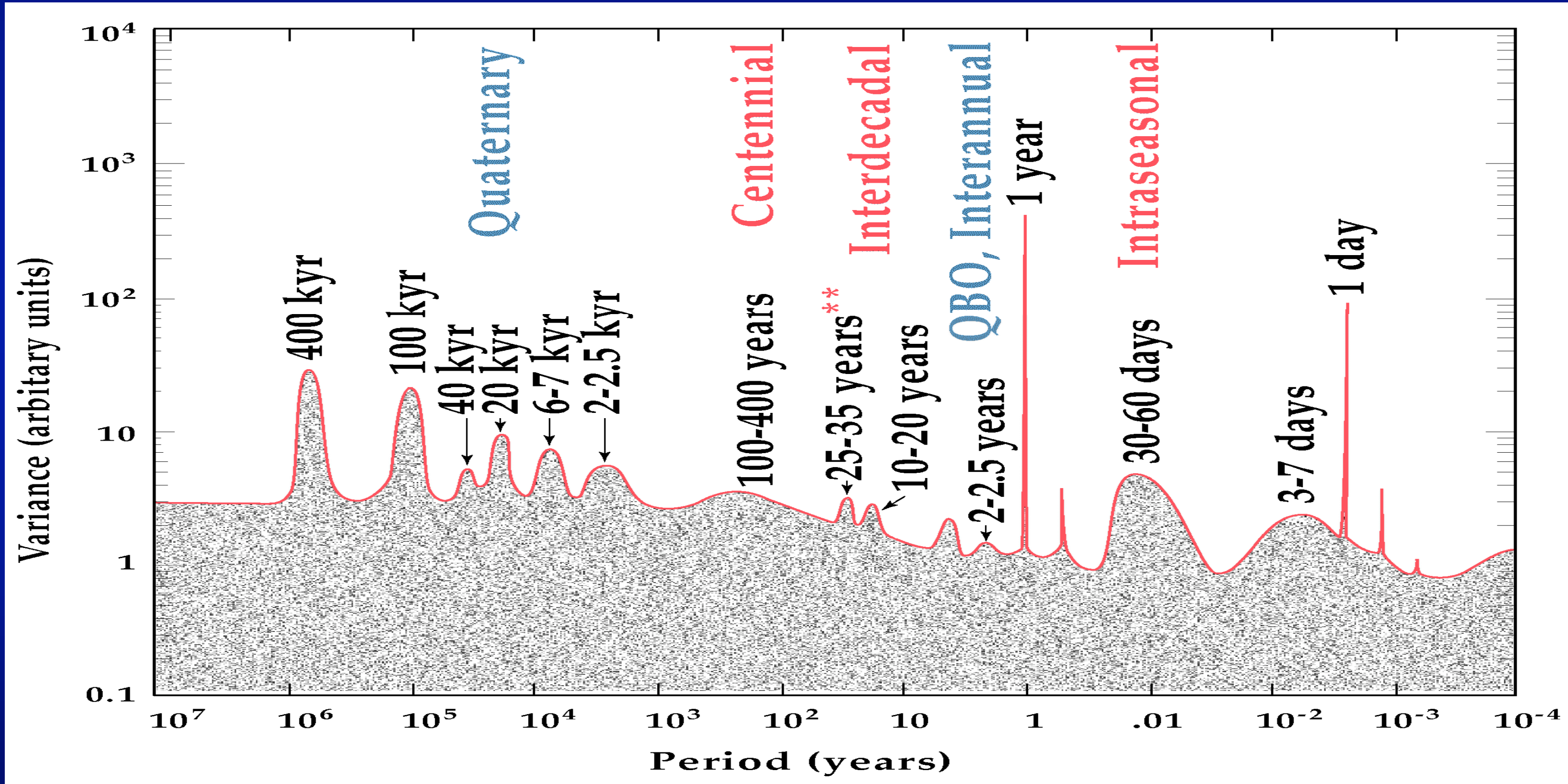

Earth System Science Overview, *NASA Advisory Council*, 1986

# Composite spectrum of climate variability

**Standard treatement of frequency bands:**

*1. High frequencies – white (or "colored") noise*

*2. Low frequencies – slow ("adiabatic") evolution of parameters*

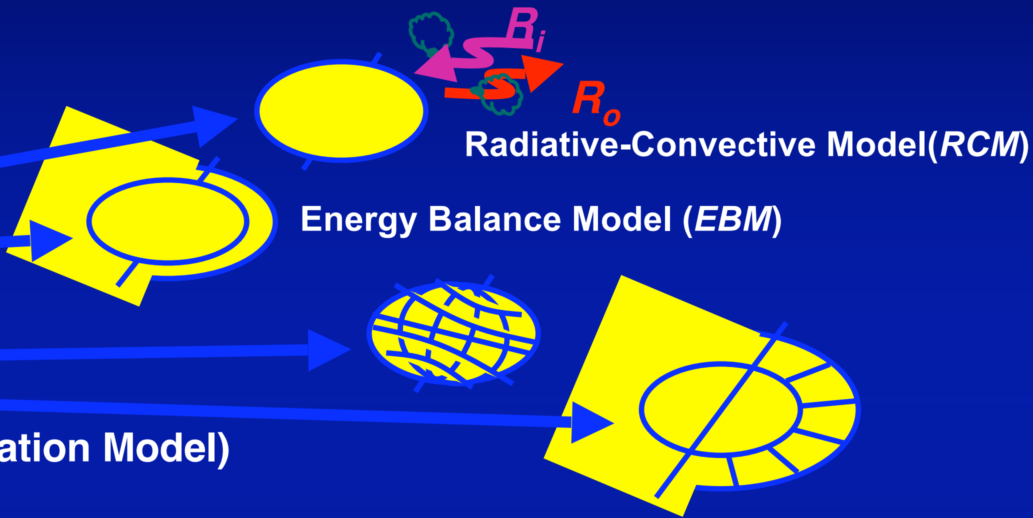



From **Ghil (2001, EGEC)**, after **Mitchell* (1976)**

* **"No known source of deterministic internal variability"**

# Climate models (atmospheric & coupled) : A classification

- ***Temporal***
  - stationary, (quasi-)equilibrium
  - transient, climate variability
- ***Space***
  - 0-D (dimension 0)
  - 1-D
    - vertical
    - latitudinal
  - 2-D
    - horizontal
    - meridional plane
  - 3-D, GCMs (General Circulation Model)
    - horizontal
    - meridional plane
  - Simple and intermediate 2-D & 3-D models




- ***Coupling***
  - Partial
    - unidirectional
    - asynchronous, hybrid
  - Full

**Hierarchy:** from the simplest to the most elaborate,
iterative comparison with the observational data

## Linear inverse model (LIM)

• We aim to use data in order to estimate the two matrices, **B** and **Q**, of the stochastic linear model:

$$dX = BX \cdot dt + d\xi(t), \quad (1)$$

where **B** is the (constant and stable) dynamics matrix, and **Q** is the lag-zero covariance of the vector white-noise process $d\xi(t)$.

• More precisely, the two matrices **B** and **Q** are related by a fluctuation-dissipation relation:

$$\mathbf{BC}(0) + \mathbf{C}(0)\mathbf{B}^t + \mathbf{Q} = 0, \quad (2)$$

where $\mathbf{C}(\tau) = \mathbb{E}\{\mathbf{X}(t+\tau)\mathbf{X}(t)\}$ is the lag-covariance matrix of the process $\mathbf{X}(t)$, and $(\cdot)^t$ indicates the transpose.

• One then proceeds to estimate the Green's function $\mathbf{G}(\tau) = \exp(\tau\mathbf{B})$ at a given lag $\tau_0$ from the sample $\mathbf{C}(\tau)$ by

$$\mathbf{G}(\tau_0) = \mathbf{C}(\tau_0)\mathbf{C}^{-1}(0).$$

## Nonlinear stochastic model (MTV)–I

• Let **z** be a **vector** decomposed into a **slow** (“climate") and a **fast** (“weather") vector of variables, $\mathbf{z} = (\mathbf{x}, \mathbf{y})$.
We model **x** deterministically and **y** stochastically, via the following **quadratic nonlinear dynamics**

$$\frac{d\mathbf{x}}{dt} = L_{11}\mathbf{x} + L_{12}\mathbf{y} + B_{11}^1(\mathbf{x},\mathbf{x}) + B_{12}^1(\mathbf{x},\mathbf{y}) + B_{22}^1(\mathbf{y},\mathbf{y}),$$

$$\frac{d\mathbf{y}}{dt} = L_{21}\mathbf{x} + L_{22}\mathbf{y} + B_{11}^2(\mathbf{x},\mathbf{x}) + B_{12}^2(\mathbf{x},\mathbf{y}) + B_{22}^2(\mathbf{y},\mathbf{y}).$$

• In stochastic modeling, the explicit nonlinear self-interaction for the variable **y**, i.e. $B_{22}^2(\mathbf{y},\mathbf{y})$, is represented by a linear stochastic operator:

$$B_{22}^2(\mathbf{y},\mathbf{y}) \approx -\frac{\Gamma}{\varepsilon}\mathbf{y} + \frac{\sigma}{\sqrt{\varepsilon}}\dot{\mathbf{W}}(t),$$

where $\Gamma$ and $\sigma$ are matrices and $\dot{\mathbf{W}}(t)$ is a **vector-valued** white-noise.

# Nonlinear stochastic model (MTV)–II

• The parameter $\varepsilon$ measures the ratio of the correlation time of the weather and the climate variables, respectively, and $\varepsilon \ll 1$ corresponds to this ratio being very small.
• Using the scaling $t \to \varepsilon t$, we derive the stochastic climate model:

$$\frac{d\mathbf{y}}{dt} = \frac{1}{\varepsilon}(L_{11}\mathbf{x} + L_{12}\mathbf{y} + B^1_{11}(\mathbf{x},\mathbf{x}) + B^1_{12}(\mathbf{x},\mathbf{y})),$$

$$\frac{d\mathbf{y}}{dt} = \frac{1}{\varepsilon}(L_{21}\mathbf{x} + L_{22}\mathbf{y} + B^2_{11}(\mathbf{x},\mathbf{x}) + B^2_{12}(\mathbf{x},\mathbf{y})) - \frac{\Gamma}{\varepsilon^2}\mathbf{y} + \frac{\sigma}{\varepsilon}\dot{\mathbf{W}}(t).$$

• In practice, the climate variables are determined by a variety of procedures, including leading-order empirical orthogonal functions (EOFs), zonal averaging in space, low-pass and high-pass time filtering, or a combination of these procedures.

# References for LIM & MTV

## Linear Inverse Models (LIM)

Penland, C., 1989: Random forcing and forecasting using principal oscillation pattern analysis. *Mon. Wea. Rev.*, **117**, 2165–2185.

Penland, C., 1996: A stochastic model of Indo-Pacific sea-surface temperature anomalies. *Physica D*, **98**, 534–558.

Penland, C., and M. Ghil, 1993: Forecasting Northern Hemisphere 700-mb geopotential height anomalies using empirical normal modes. *Mon. Wea. Rev.*, **121**, 2355–2372.

Penland, C., and L. Matrosova, 1998: Prediction of tropical Atlantic sea-surface temperatures using linear inverse modeling. *J. Climate*, **11**, 483–496.

## Nonlinear reduced models (MTV)

Majda, A. J., I. Timofeyev, and E. Vanden-Eijnden, 1999: Models for stochastic climate prediction. *Proc. Natl. Acad. Sci. USA*, **96**, 14687–14691.

Majda, A. J., I. Timofeyev, and E. Vanden-Eijnden, 2001: A mathematical framework for stochastic climate models. *Commun. Pure Appl. Math.*, **54**, 891–974.

Majda, A. J., I. Timofeyev, and E. Vanden-Eijnden, 2002: A priori test of a stochastic mode reduction strategy. *Physica D*, **170**, 206–252.

Majda, A. J., I. Timofeyev, and E. Vanden-Eijnden, 2003: Systematic strategies for stochastic mode reduction in climate. *J. Atmos. Sci.*, **60**, 1705–1722.

Franzke, C., and Majda, A. J., 2006: Low-order stochastic mode reduction for a prototype atmospheric GCM. *J. Atmos. Sci.*, **63**, 457–479.

# Motivation

- Sometimes we have data but no models.
- Linear inverse models (LIM) are good least-square fits to data, but don't capture all the processes of interest.
- Difficult to separate between the slow and fast dynamics (MTV).
- We want models that are as simple as possible, but not any simpler.

## Criteria for a good data-derived model

- Fit the data, as well or better than LIM.
- Capture interesting dynamics: regimes, nonlinear oscillations.
- Intermediate-order deterministic dynamics.
- Good noise estimates.

# Key ideas

- Nonlinear dynamics: $\dot{\mathbf{x}} = \mathbf{L}\mathbf{x} + \mathbf{N}(\mathbf{x}).$

- Discretized, quadratic:

$$dx_i = (\mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{(0)}\mathbf{x} + c_i^{(0)})\,dt + dr_i^{(0)}; \qquad 1 \le i \le I.$$

- Multi-level modeling of red noise:

$$\begin{aligned}
dx_i &= (\mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{b}_i^{(0)}\mathbf{x} + c_i^{(0)})\,dt + r_i^{(0)}\,dt,\\
dr_i^{(0)} &= \mathbf{b}_i^{(1)}[\mathbf{x}, \mathbf{r}^{(0)}]\,dt + r_i^{(1)}\,dt,\\
dr_i^{(1)} &= \mathbf{b}_i^{(2)}[\mathbf{x}, \mathbf{r}^{(0)}, \mathbf{r}^{(1)}]\,dt + r_i^{(2)}\,dt,\\
&\cdots\\
dr_i^{(L)} &= \mathbf{b}_i^{(L)}[\mathbf{x}, \mathbf{r}^{(0)}, \mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(L)}]\,dt + dr_i^{(L+1)}; \qquad 1 \le i \le I.
\end{aligned}$$

# Nomenclature

*Response* variables:

$$\{y^{(n)}\}\ (1 \le n \le N)\ \equiv\ \{y^{(1)}, \ldots, y^{(N)}\}$$

*Predictor* variables:

$$\{x^{(n)}\}\ (1 \le n \le N)\ \equiv\ \{x^{(1)}, \ldots, x^{(N)}\}$$

- Each $y^{(n)}$ is *normally distributed* about $\widehat{y}^{(n)}$
- Each $x^{(n)}$ is *known exactly. Parameter set* $\{a_p\}$:

$$\widehat{y} = f(x;\, a_1, \ldots, a_P)$$

– known dependence of *f* on $\{x^{(n)}\}$ and $\{a_p\}$.

*REGRESSION*: Find $\{a_p\}\ (1 \le p \le P)$

# LIM extension #1

- Do a least-square fit to a *nonlinear function of the data*:

*J response variables*: $y_i^{(n)} \equiv (x_i^{(n+1)} - x_i^{(n)})/\Delta t$

*Predictor variables* (example – quadratic polynomial of *J* original predictors):

$$\widehat{y}_i = a_{0,\,i} + \sum_{j=1}^{J} a_{j,\,i} x_j + \sum_{j=1}^{J} \sum_{k \geq j} \tilde{a}_{jk,\,i} x_j x_k$$

*Note*: Need to find many more regression coefficients than for LIM; in the example above $P = J + J(J+1)/2 + 1 = O(J^2)$.

# Regularization

- *Caveat*: If the number *P* of regression parameters is comparable to (*i.e.*, it is not much smaller than) the number of data points, then the least-squares problem may become ill-posed and lead to unstable results (overfitting) ==> One needs to transform the predictor variables to *regularize* the regression procedure.

- Regularization involves *rotated predictor variables:* the orthogonal transformation looks for an "optimal" linear combination of variables.

- "Optimal" = (i) rotated predictors are nearly uncorrelated; and
  (ii) they are maximally correlated with the response.

- Canned packages available.

# LIM extension #2

- *Motivation*: Serial correlations in the residual.

Main level, $l = 0$: $$(x^{n+1} - x^n)/\Delta t = a_{x,0} x^n + r_0^n$$

Level $l = 1$: $$(r_0^{n+1} - r_0^n)/\Delta t = a_{x,1} x^n + a_{r_0,1} r_0^n + r_1^n$$

… and so on …

Level $L$: $$r_{L-1}^{n+1} - r_{L-1}^n = \Delta t [a_{x,L} x^n + \ldots] + \Delta r_L$$

- $\Delta r_L$ – Gaussian random deviate with appropriate variance
- If we suppress the dependence on $x$ in levels $l = 1, 2, \ldots L$, then the model above is formally identical to an ARMA model.

# Empirical Orthogonal Functions (EOFs)

- We want models that are as simple as possible, but not any simpler: use leading **empirical orthogonal functions** for data compression and capture as much as possible of the useful (predictable) variance.
- Decompose a spatio-temporal data set **D**($t$,$s$)($t$ = 1,…,$N$; $s$ = 1…,$M$) by using **principal components (PCs**) – $\mathbf{x}_i(t)$ and **empirical orthogonal functions (EOFs)** – $\mathbf{e}_i(s)$: diagonalize the $M$ x $M$ spatial covariance matrix **C** of the field of interest.

$$\mathbf{C} = \frac{1}{N}(\mathbf{D} - < \mathbf{D} >)^{\mathrm{t}}(\mathbf{D} - < \mathbf{D} >)$$
$$\mathbf{C}\lambda_i = \lambda_i e_i, x_i = (\mathbf{D} - < \mathbf{D} >)e_i$$

- EOFs are optimal patterns to capture most of the variance.
- Assumption of robust EOFs.
- EOFs are statistical features, but may describe some dynamical (physical) mode(s).

# Empirical mode reduction (EMR)–I

- *Multiple predictors*: Construct the reduced model using $J$ leading PCs of the field(s) of interest.

- *Response variables*: one-step time differences of predictors; step = sampling interval = $\Delta t$.

- Each response variable is fitted by an *independent* multi-level model:
  The *main level* $l = 0$ is *polynomial* in the predictors;
  all the other levels are linear.

# Empirical mode reduct'n (EMR) – II

- The number *L* of levels is such that each of the last-level residuals (for each channel corresponding to a given response variable) is "white" in time.

- Spatial (cross-channel) correlations of the last-level residuals are retained in subsequent regression-model simulations.

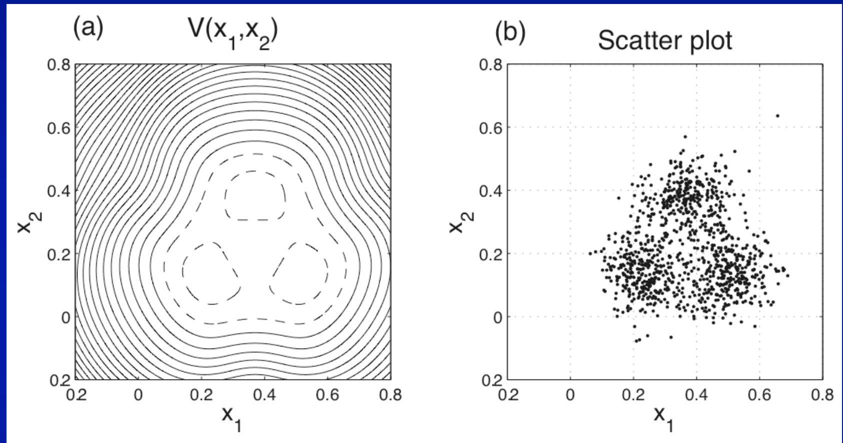
- The number *J* of PCs is chosen so as to optimize the model's performance.

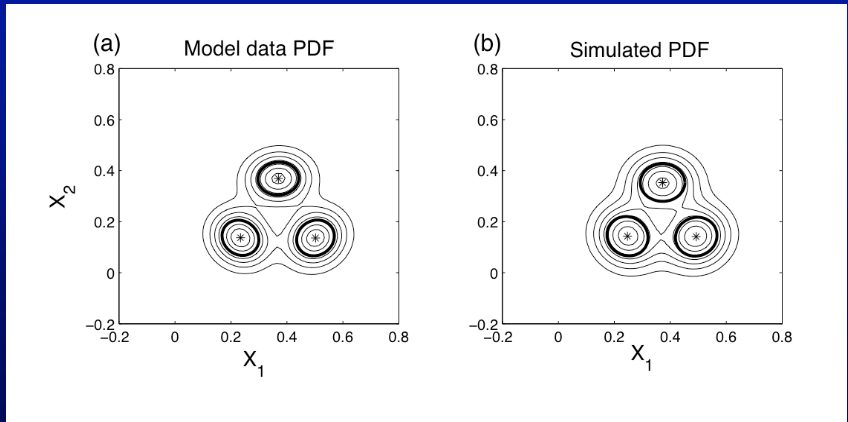
- Regularization is used at the main (nonlinear) level of each channel.

# Illustrative example: Triple well

$$d\mathbf{x}(t) = -\nabla V(\mathbf{x})dt + \sigma\, d\mathbf{b}$$

- *$V(x_1,x_2)$ is not polynomial*!
- Our *polynomial* regression model produces a time series whose statistics are nearly identical to those of the full model!!
- Optimal order is $m = 3$; regularization required for polynomial models of order $m \geq 5$.



# NH LFV in QG3 Model – I

*The QG3 model* (Marshall and Molteni, *JAS*, 1993):

- Global QG, T21, 3 levels, with topography; perpetual-winter forcing; ~1500 degrees of freedom.

- Reasonably realistic NH climate and LFV:
  (i) multiple planetary-flow regimes; and
  (ii) low-frequency oscillations
  (submonthly-to-intraseasonal).

- Extensively studied: A popular "numerical-laboratory" tool to test various ideas and techniques for NH LFV.

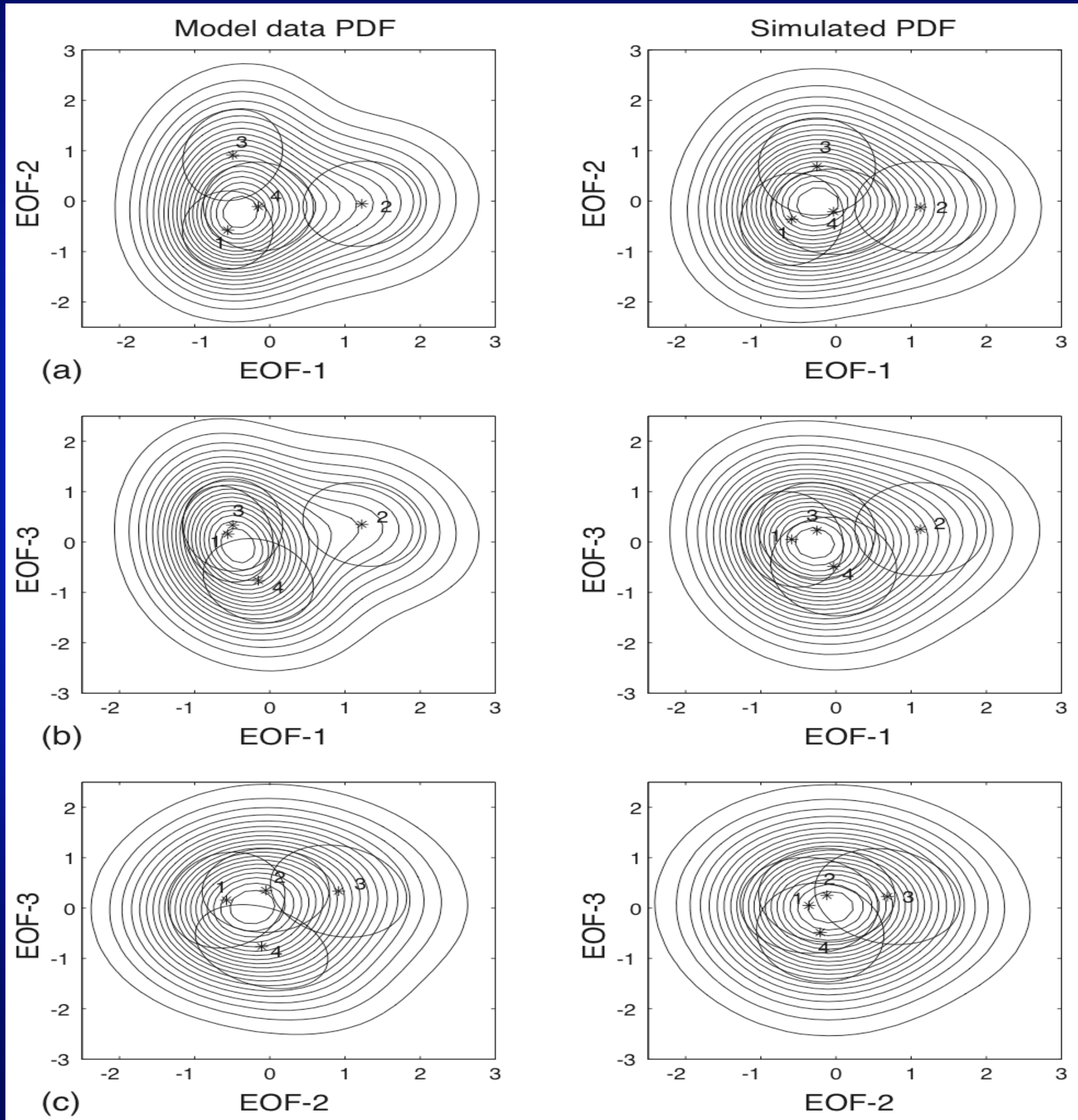
# NH LFV in QG3 Model – II

*Output*: daily streamfunction ($\Psi$) fields (≈ $10^5$ days)

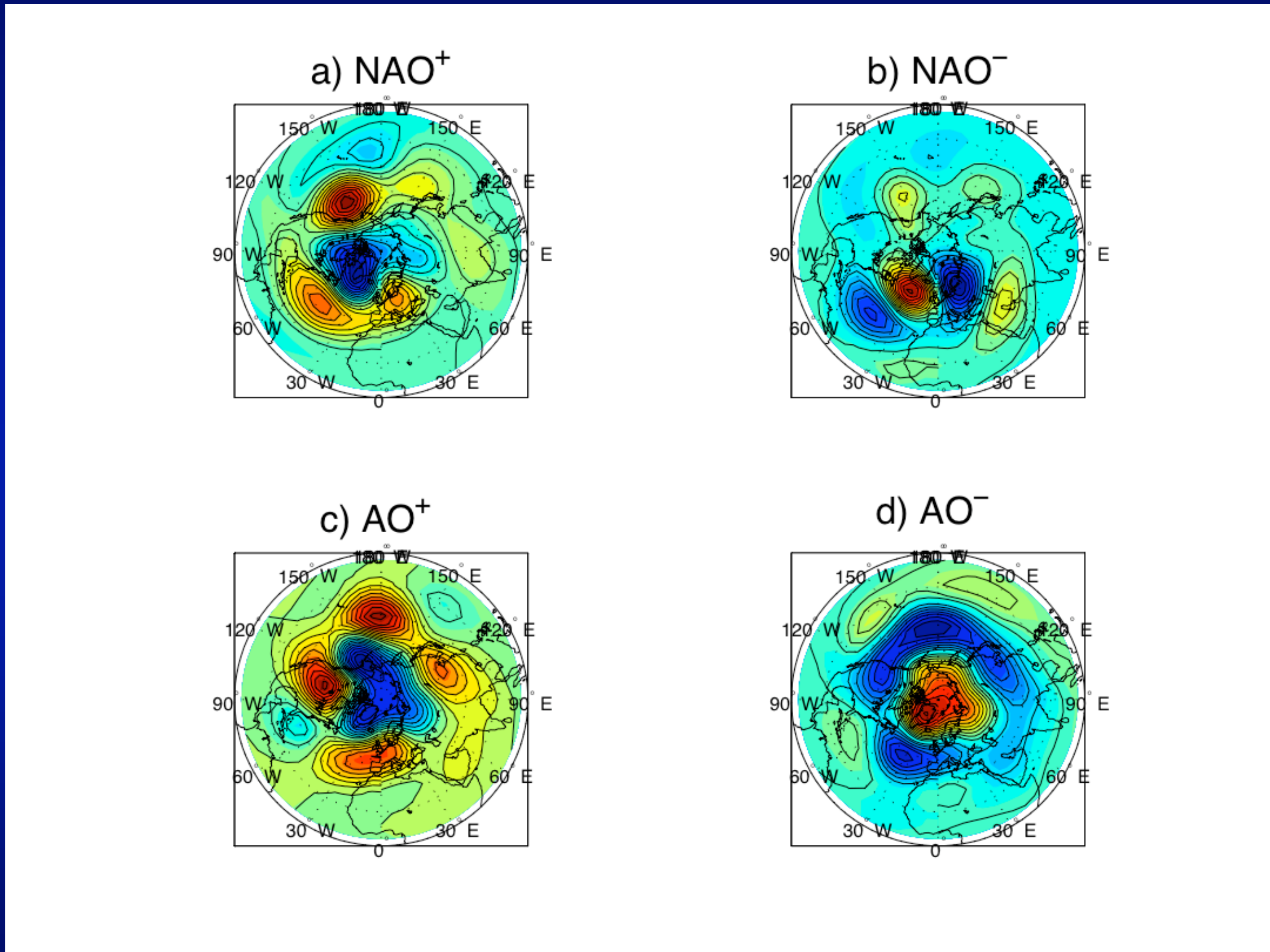
*Regression model*:

- 15 variables, 3 levels ($L$ = 3), quadratic at the main level
- Variables: Leading PCs of the middle-level $\Psi$
- No. of degrees of freedom = 45 (a factor of 40 less than in the QG3 model)
- Number of regression coefficients $P$ = (15+1+15•16/2+30+45)•15 = 3165 (<< $10^5$)
- Regularization via PLS applied at the main level.

# NH LFV in QG3 Model – III



# NH LFV in QG3 Model – IV




The correlation between the QG3 map and the EMR model's map exceeds 0.9 for each cluster centroid.

# Conclusions on QG3 Model

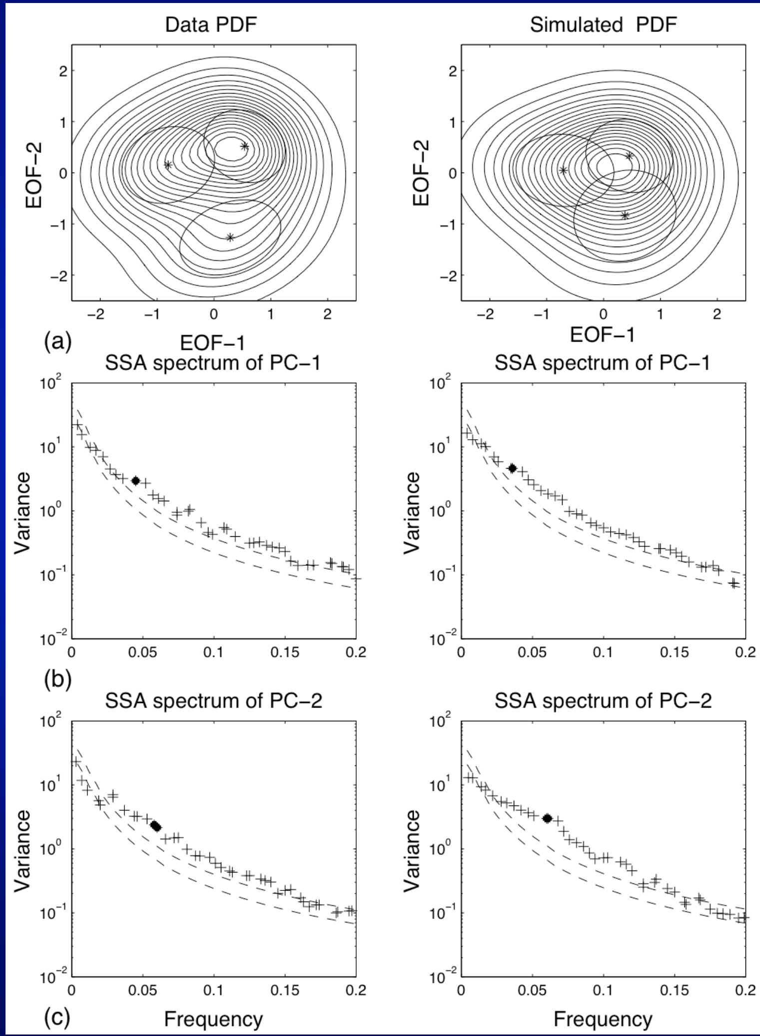
- Our ERM is based on 15 EOFs of the QG3 model and has $L$ = 3 regression levels, *i.e.,* a total of 45 predictors (*).
- The ERM approximates the QG3 model's major statistical features (PDFs, spectra, regimes, transition matrices, etc.) strikingly well.
- The dynamical analysis of the reduced model identifies $AO^{-}$ as the model's unique steady state.
- The 37-day mode is associated, in the reduced model, with the least-damped linear eigenmode.
- The additive noise interacts with the nonlinear dynamics to yield the full ERM's (and QG3's) phase-space PDF.

(*) An ERM model with 4*3 = 12 variables only does not work!

# NH LFV – Observed Heights

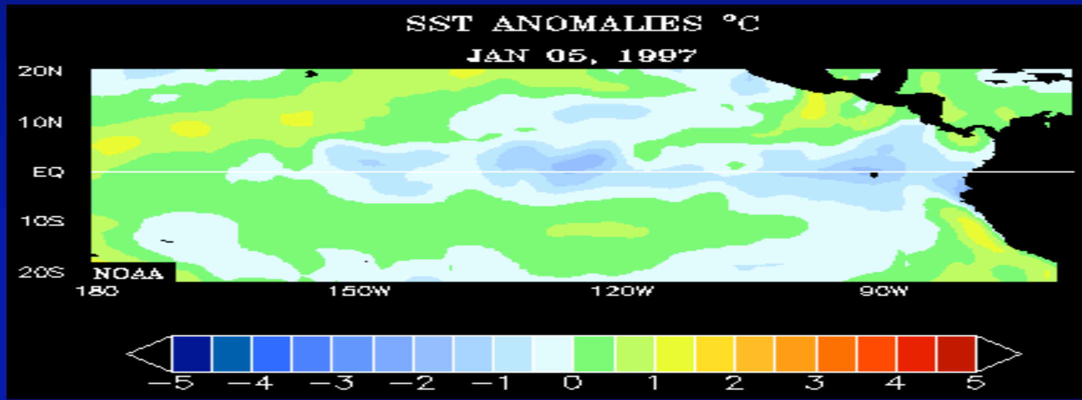
- 44 years of daily 700-mb-height winter data

- 12-variable, 2-level model works OK, but dynamical operator has unstable directions: “sanity checks” required.



# Spatio-temporal evolution of ENSO episode

## 1997-98 El Niño Animation




Anomaly = (Current observation – Corresponding climatological value)
Base period for the climatology is 1950–1979

NOAA-CIRES Climate Diagnostics Center

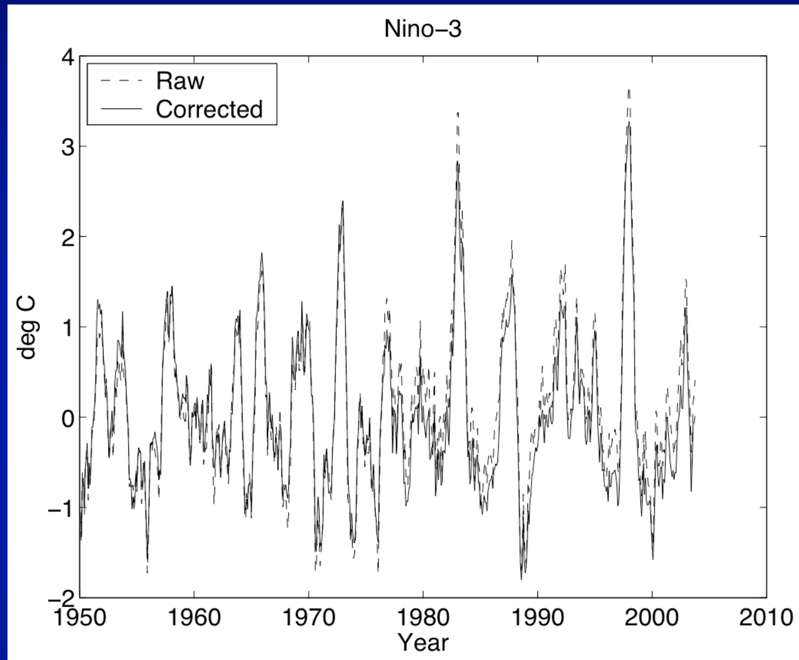
http://www.cdc.noaa.gov/map/clim/sst_olr/old_sst/sst_9798_anim.shtml

Courtesy of NOAA-CIRES Climate Diagnostics Center

# ENSO – I

*Data*:

- Monthly SSTs: 1950–2004, 30 S–60 N, 5x5 grid (Kaplan *et al*., 1998)



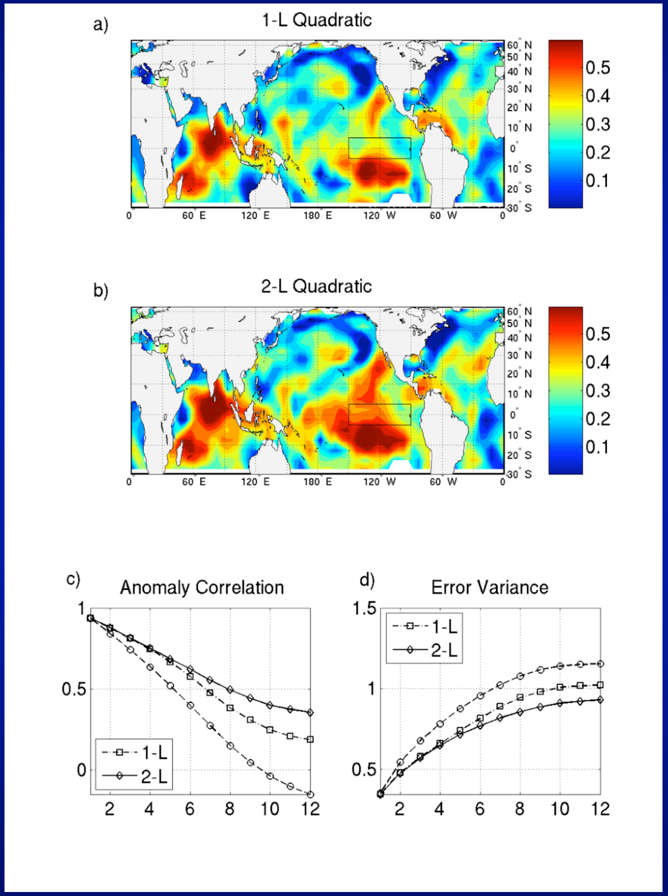

- 1976–1977 shift removed

- Histogram of SST data is skewed (warm events are larger, while cold events are more frequent): Nonlinearity important?

# ENSO – II

*Regression model:*

- *J* = 20 variables (EOFs of SST)
- *L* = 2 levels




- Seasonal variations included in the linear part of the main (quadratic) level.

- Competitive skill: Currently a member of a multi-model prediction scheme of the IRI, see: http://iri.columbia.edu/climate/ENSO/currentinfo/SST_table.html.

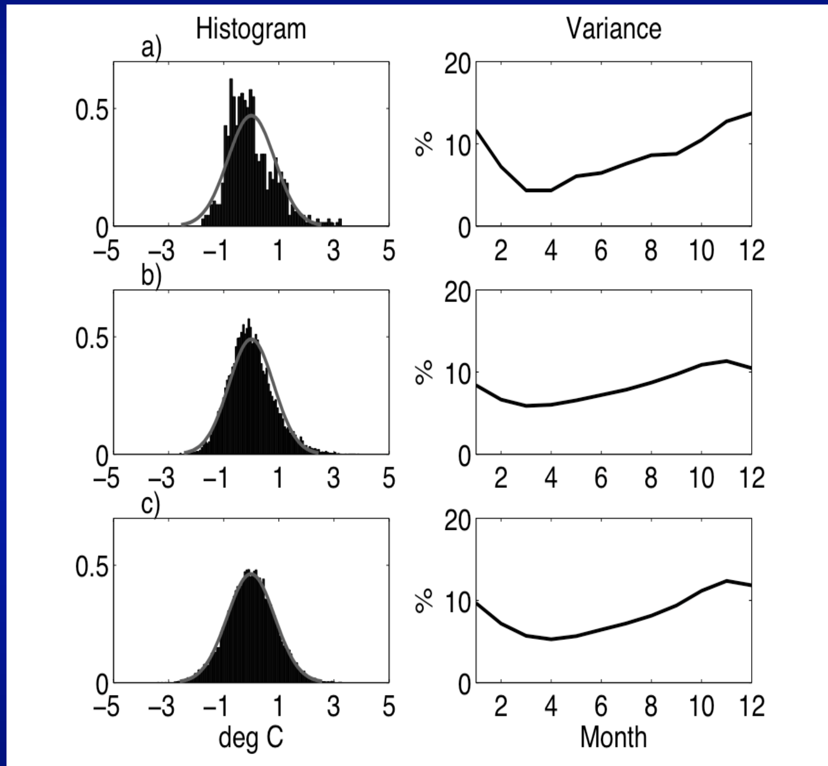
# ENSO – III

## PDF – skewed vs. Gaussian

- Observed

- Quadratic model
  (100-member ensemble)

- Linear model
  (100-member ensemble)

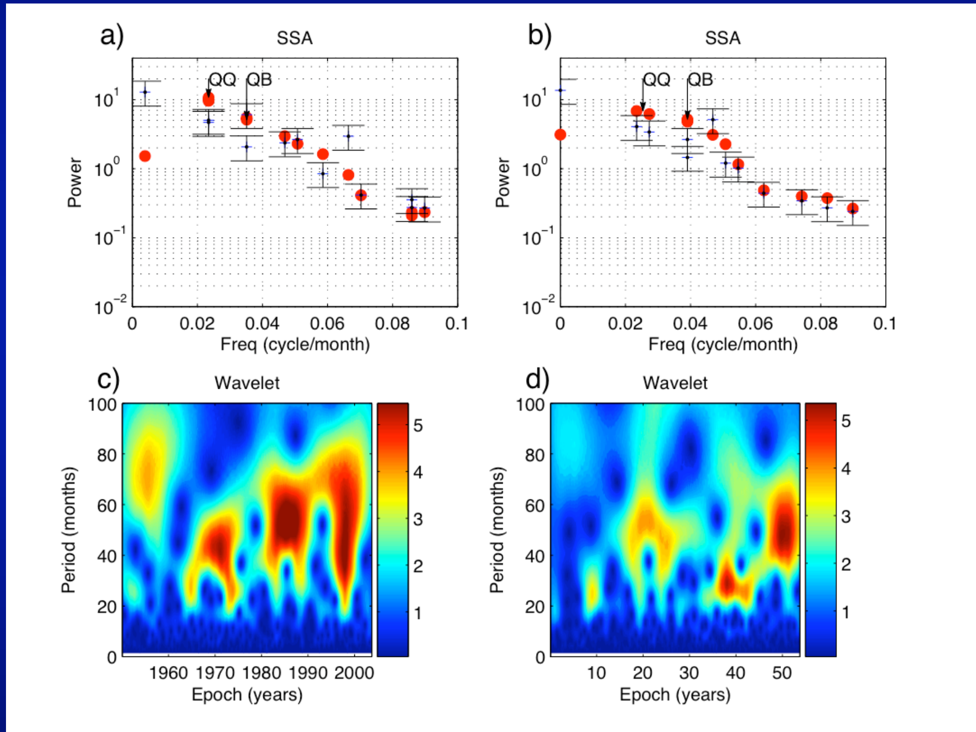



*The quadratic model has a slightly smaller RMS error in its extreme-event forecasts (not shown)*

# ENSO – IV

*Spectra:* Data Model

- SSA
- Wavelet

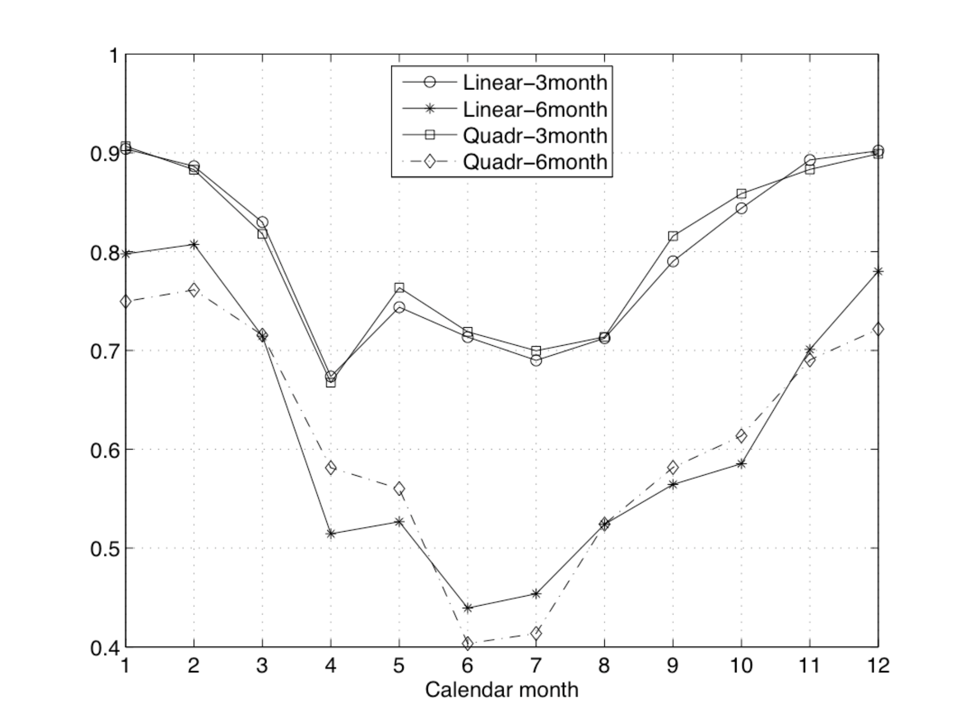



*ENSO's leading oscillatory modes, QQ and QB, are reproduced by the model, thus leading to a skillful forecast.*

# ENSO – V

*"Spring barrier":*

Hindcast skill vs. target month




- SSTs for June are more difficult to predict.

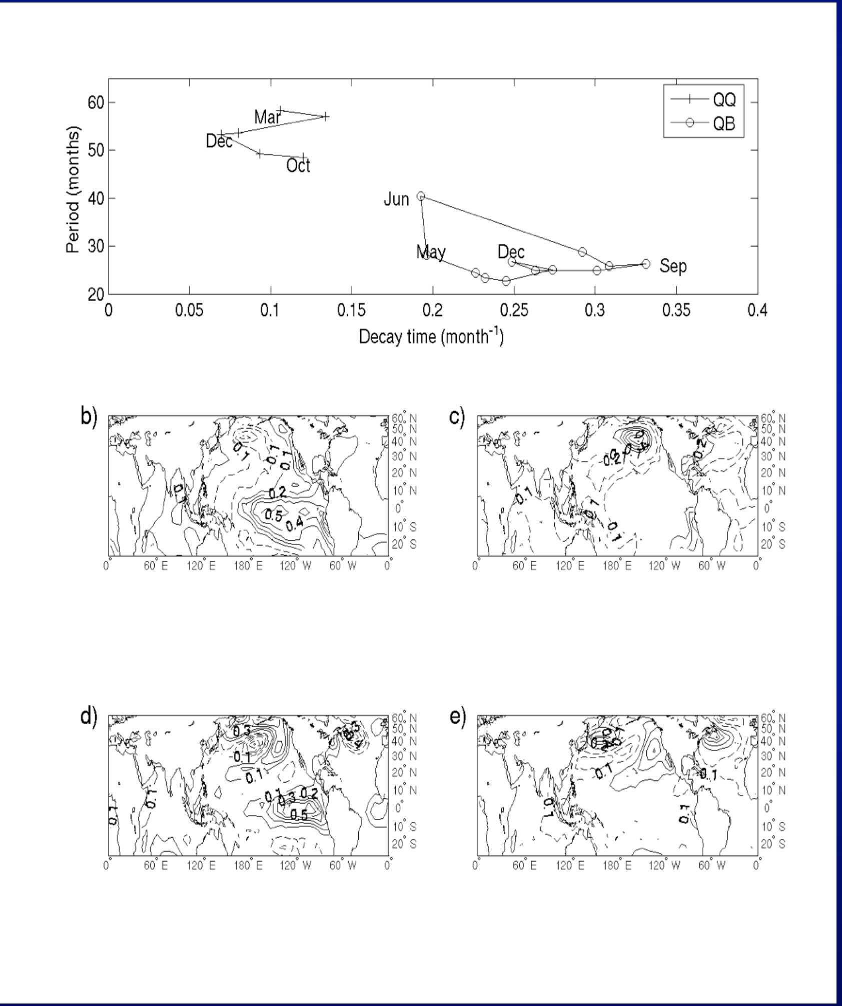
- A feature of virtually all ENSO forecast schemes.

- SST anomalies are weaker in late winter through summer (why?), and signal-to-noise ratio is low.

# ENSO – VI

- Stability analysis, month-by-month, of the linearized regression model identifies weakly damped QQ mode (with a period of 48–60 mo), as well as strongly damped QB mode.

- QQ mode is least damped in December, while it is not identifiable at all in summer!



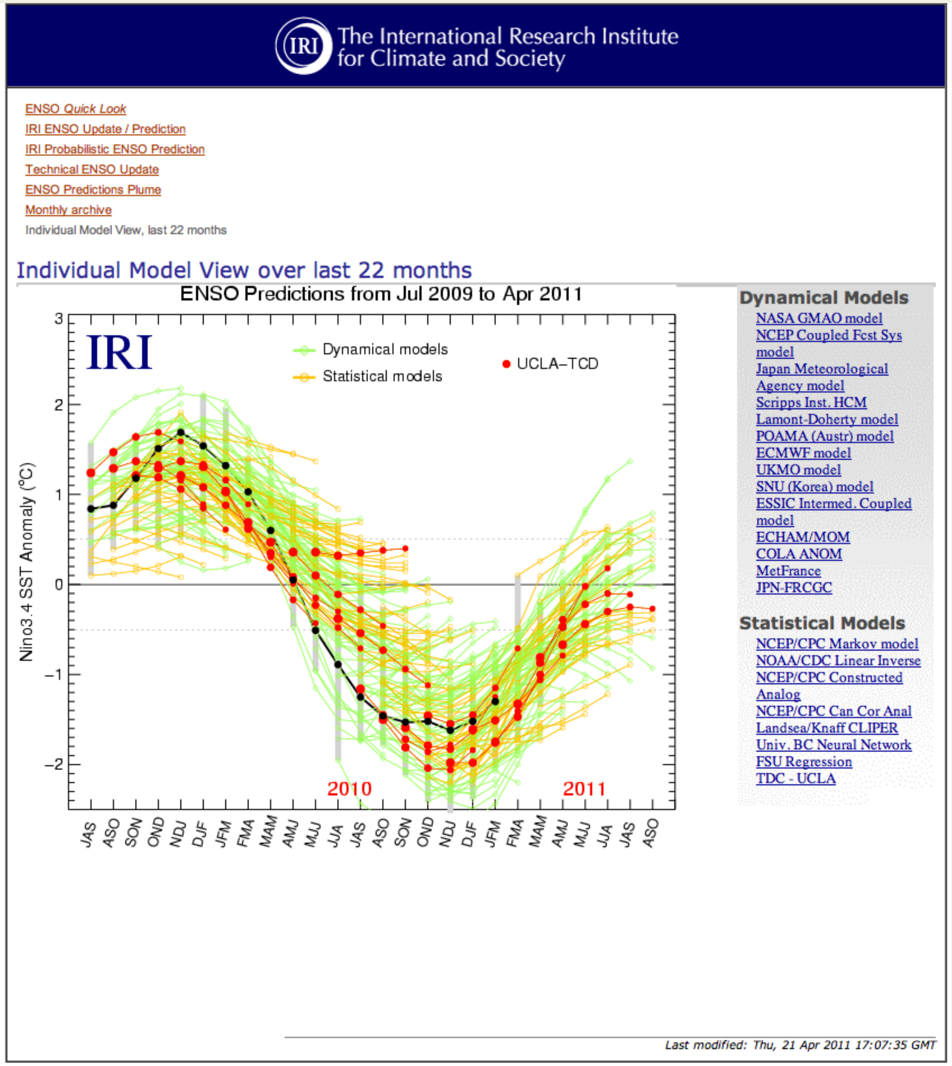
# IRI Forecast Plume for ENSO Prediction

- 14 dynamical + 9 statistical = 23 models

Reasonable skill for 6–8 months ≃ ¼ of QBO period (*)

The new IRI plume will include only the best forecast models so far;

in particular the TCD-UCLA EMR-based model




(*) Large events — warm (El Niño) & cold (La Niña) are "beats" of the QB and QQ periodicities (Ghil & Jiang, 1998, *GRL*)

# Conclusions on ENSO model

- The quadratic, 2-level EMR model has competitive forecast skill.
- Two levels really matter in modeling “noise.”
- EMR model captures well the “linear,” as well as the “nonlinear” phenomenology of ENSO.
- Observed statistical features can be related to the EMR model’s dynamical operator.
- SST-only model: other variables? (A. Clarke)

# Concluding Remarks – I

- The generalized least-squares approach is well suited to derive nonlinear, reduced models (EMR models) of geophysical data sets; regularization techniques such as PCR and PLS are important ingredients to make it work.
- The multi-level structure is convenient to implement and provides a framework for dynamical interpretation in terms of the “eddy–mean flow” feedback (not shown).
- Easy add-ons, such as seasonal cycle (for ENSO, etc.).
- The dynamic analysis of EMR models provides conceptual insight into the mechanisms of the observed statistics.

# Concluding Remarks – II

*Possible pitfalls:*

- The EMR models are maps: need to have an idea about (time & space) scales in the system and sample accordingly.
- Our EMRs are parametric: functional form is pre-specified, but it can be optimized within a given class of models.
- Choice of predictors is subjective, to some extent, but their number can be optimized.
- Quadratic invariants are not preserved (or guaranteed) – spurious nonlinear instabilities may arise.

# References

Kravtsov, S., D. Kondrashov, and M. Ghil, 2005: Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability. *J. Climate*, **18**, 4404–4424.

Kondrashov, D., S. Kravtsov, A. W. Robertson, and M. Ghil, 2005: A hierarchy of data-based ENSO models. *J. Climate*, **18**, 4425–4444.

Kondrashov, D., S. Kravtsov, and M. Ghil, 2006: Empirical mode reduction in a model of extratropical low-frequency variability. *J. Atmos. Sci*., **63**, 1859-1877.

Strounine, K., S. Kravtsov, D. Kondrashov, and **M. Ghil**, 2010: Reduced models of atmospheric low-frequency variability: Parameter estimation and comparative performance, *Physica D*, **239**, 145–166, doi:10.1016/j.physd.2009.10.013.

Kravtsov, S., D. Kondrashov, and **M. Ghil**, 2009: Empirical model reduction and the modeling hierarchy in climate dynamics, in *Stochastic Physics and Climate Modelling*, Eds. T. N. Palmer and P. Williams, Cambridge Univ. Press, pp. 35–72.

Kondrashov, D., S. Kravtsov and **M. Ghil**, 2010: Signatures of nonlinear dynamics in an idealized atmospheric model, *J. Atmos. Sci*., **68**, 3–12, doi: 10.1175/2010JAS3524.1.

http://www.atmos.ucla.edu/tcd/

# Reserve slides