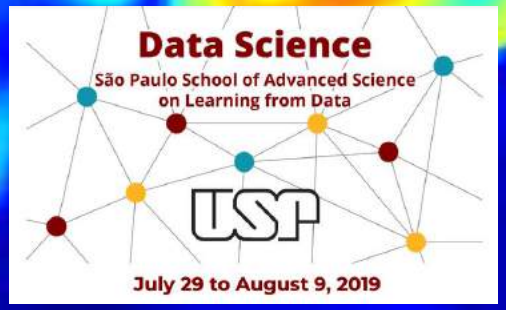
Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas

How "Big Data" impacted Atmospheric Sciences?

Pedro Leite da Silva Dias
Institute of Astronomy, Geophysics and Atmospheric Sciences
University of São Paulo



Premises

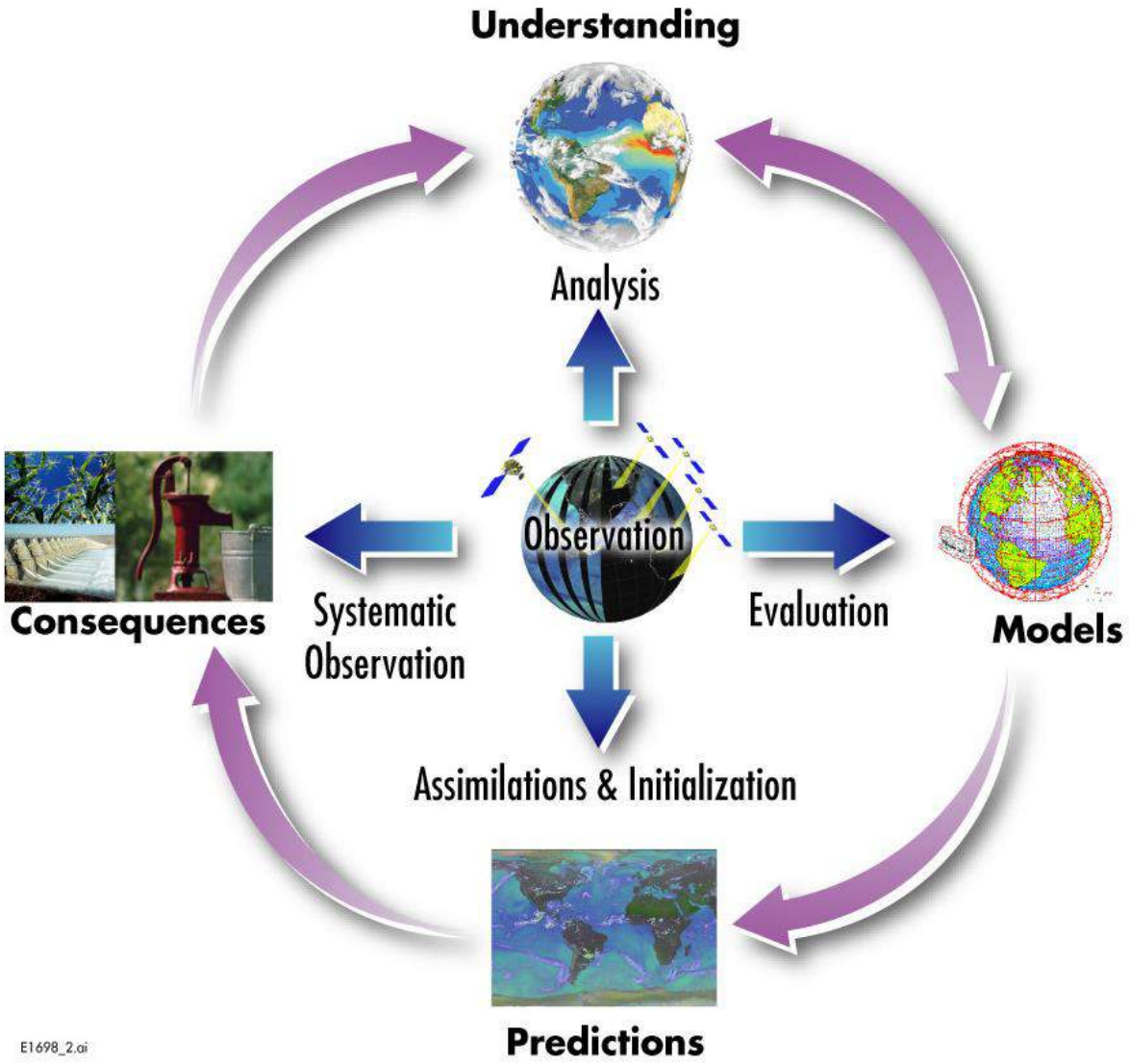
- **Large impacts of knowledge-discovery through Data Mining in health informatics, marketing, business intelligence, and smart city, where big data science contributed to several of the most recent breakthroughs.**
- **However, in Weather and Climate Prediction to date, this potential has not been fully realized, in spite of the exponential growth of climate data.**
- **Success in nowcasting – short time (e.g based on Deep Neural Network – DNN) and huge amount of data from remote sensing (radar + satellite + surface based data)**
- **This disparity stems from the complexity and variety of climate data, as well as the scientific questions climate science brings forth.**

Future

- **Some challenging problems; ensemble forecasting, data assimilation, theoretical studies on causality …**
- **Required steps to improve training in Data Science Applications in Atmospheric Sciences**

Examples on identification of climatic patterns with large socio-economical impacts:

- **Climate Patterns - El Nino/La Nina (Southern Oscillation)**
- **Pacific Decadal Oscillation**
- **Atlantic Meridional Mode**
- **Indian Ocean Dipole**
- **PNA, PSA, EU, etc……**


Understanding
Analysis
Observation
Models
Evaluation
Assimilations & Initialization
Predictions
Systematic
Observation
Consequences
E1698_2.ai

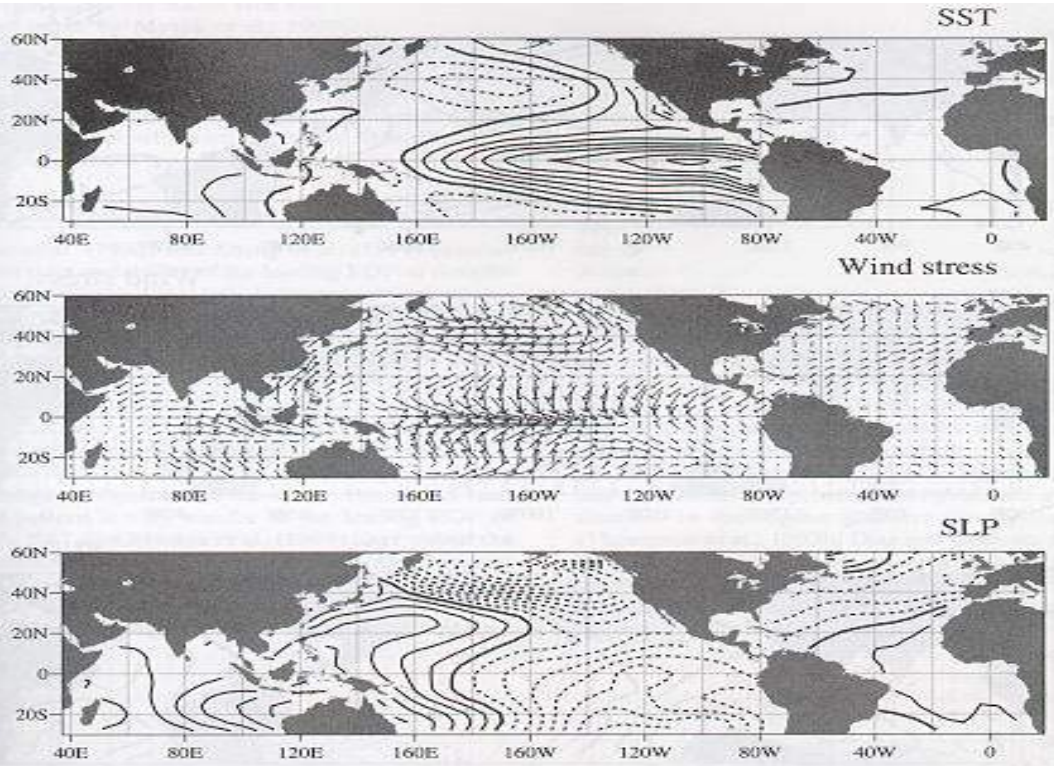


FIGURE 3-6 Global SST pattern, wind stress, and sea-level pressure that is related to the interannual variability associated with ENSO, based on linear regression between a high-pass-filtered "cold-tongue index" (CT) and global SST. (From Zhang et al., 1997; reprinted with permission of the American Meteorological Society.)

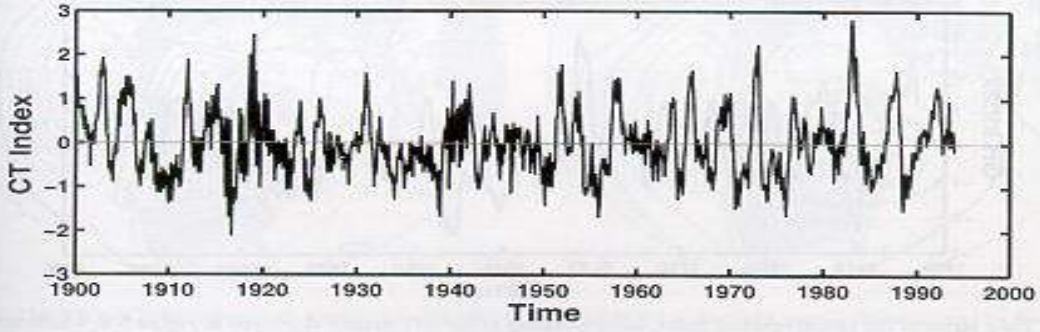
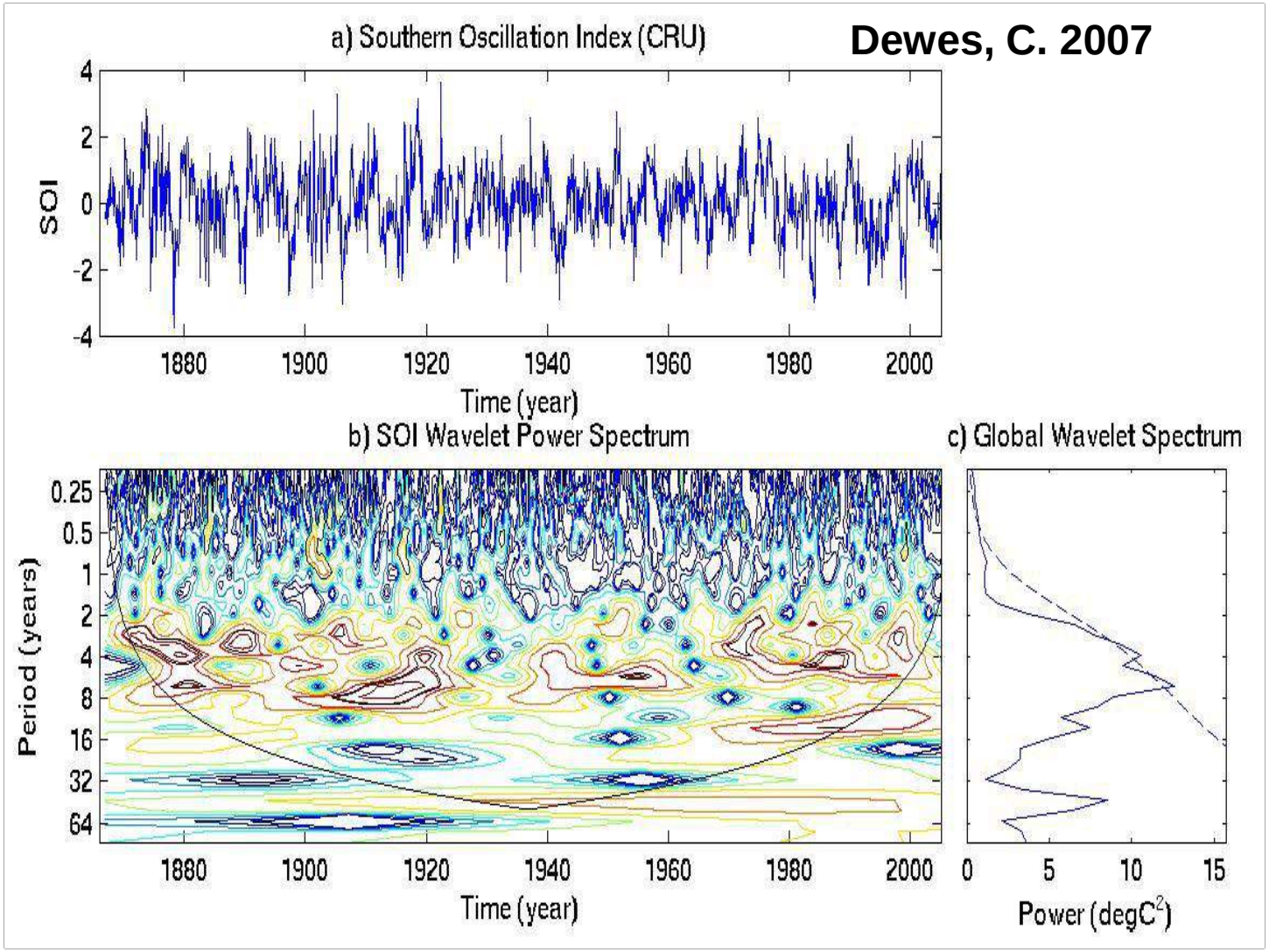


FIGURE 3-7 Time series of a cold-tongue index, corresponding to the SST pattern displayed in Figure 3-6. (From Zhang et al., 1997; reprinted with permission of the American Meteorological Society.)

El Nino/Southern Oscillation pattern

Interanual

Dewes, C. 2007


a) Southern Oscillation Index (CRU)
SOI
4
2
0
-2
-4
1880
1900
1920
1940
1960
1980
2000
Time (year)
b) SOI Wavelet Power Spectrum
Period (years)
0.25
0.5
1
2
4
8
16
32
64
1880
1900
1920
1940
1960
1980
2000
Time (year)
c) Global Wavelet Spectrum
0
5
10
15
Power (degC^2)

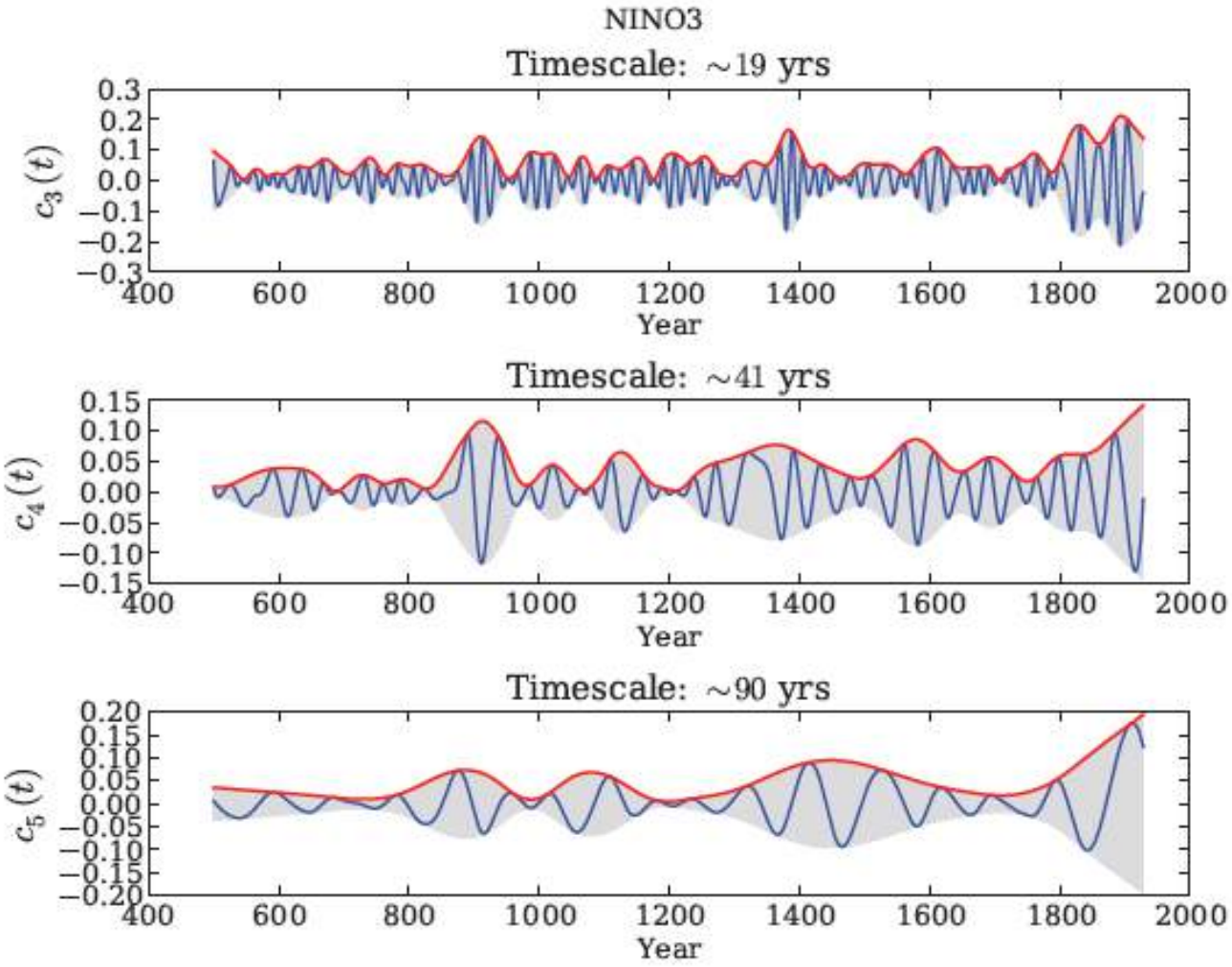


Figure: Empirical Mode Decomposition: NINO3

Paleoclimate

Reconstruction of ENSO events in the last 1.600 yrs

Welter, G. 2014

Decadal

Pattern –

Pacific Decadal

Oscillation

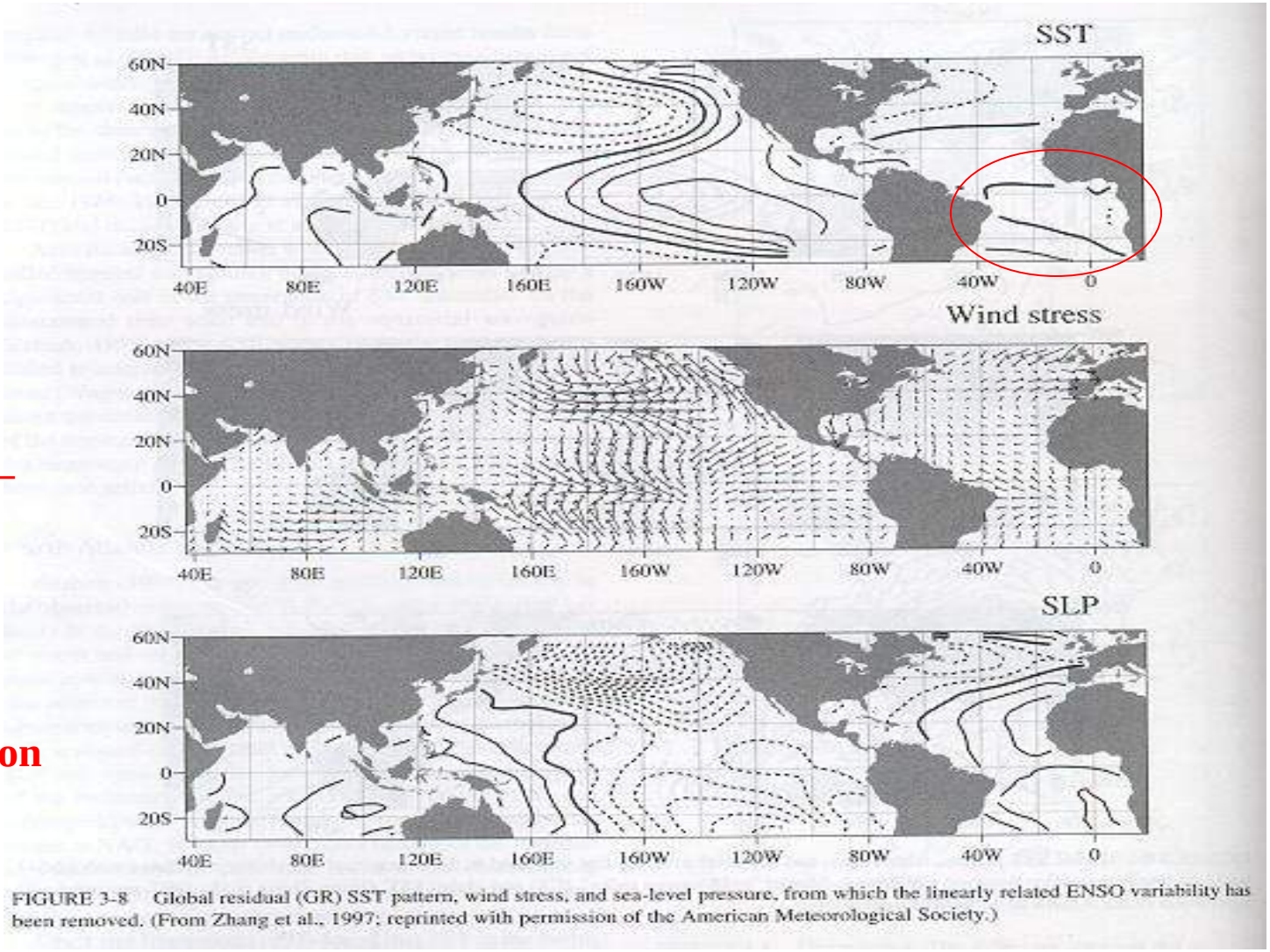
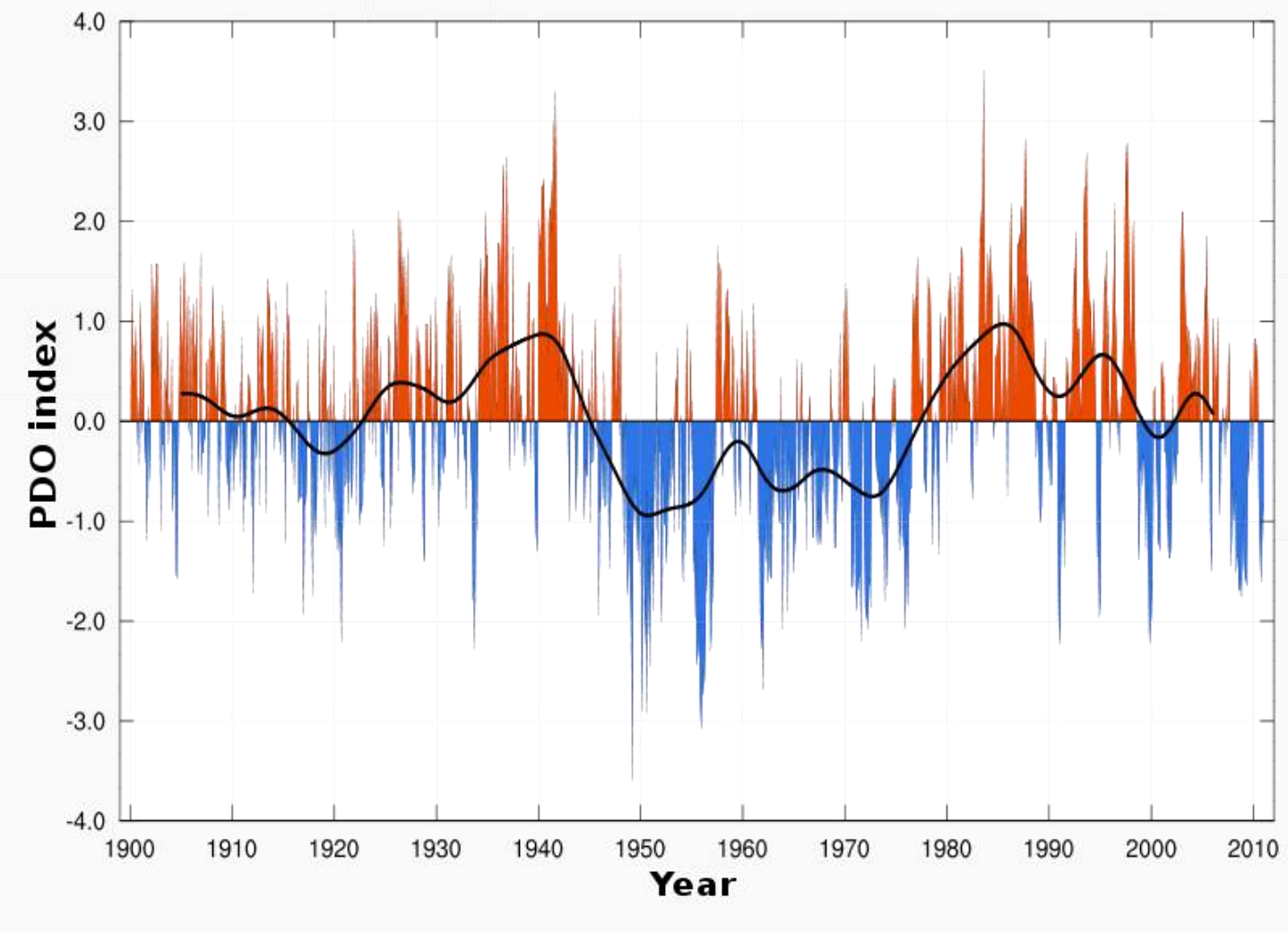
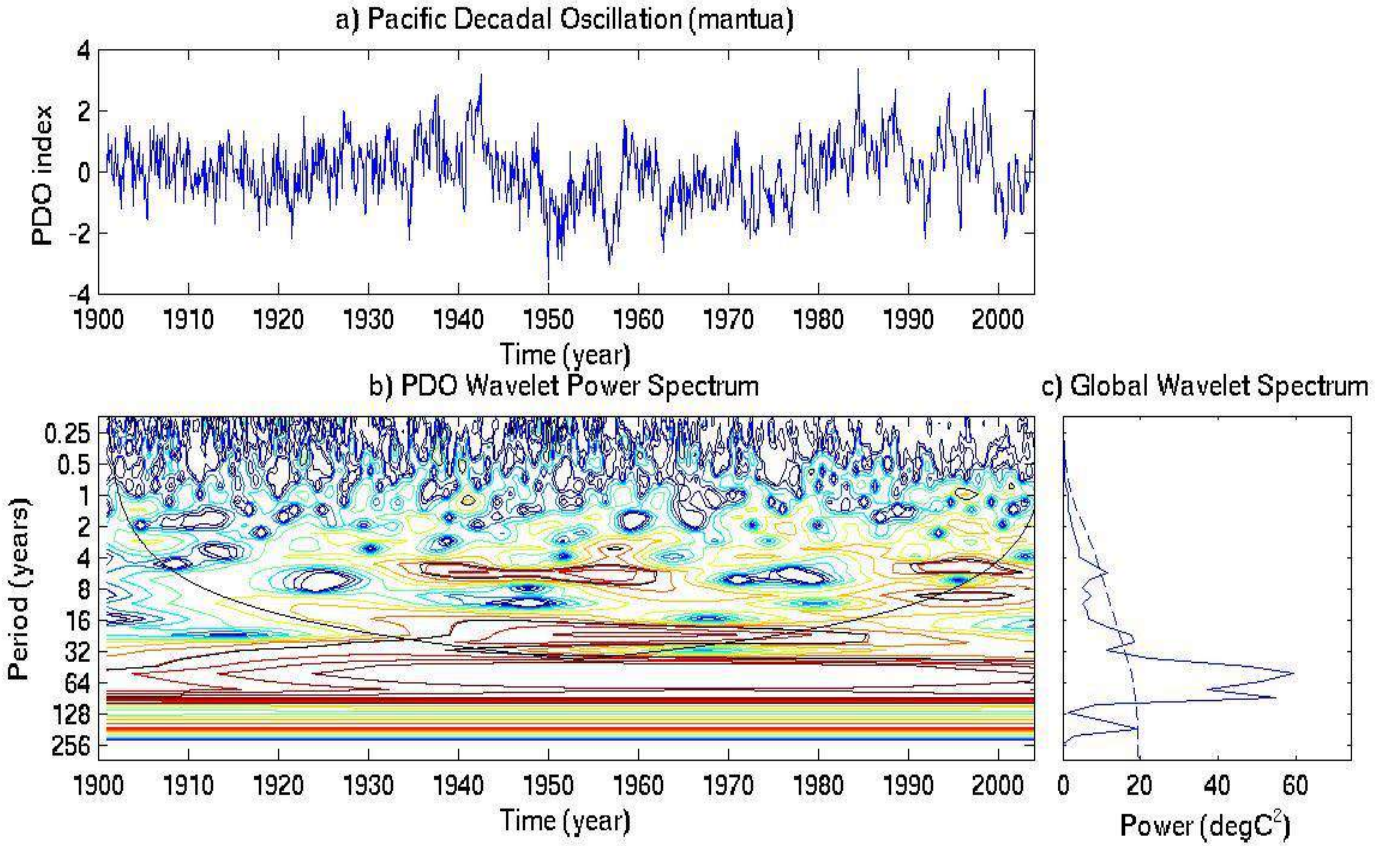


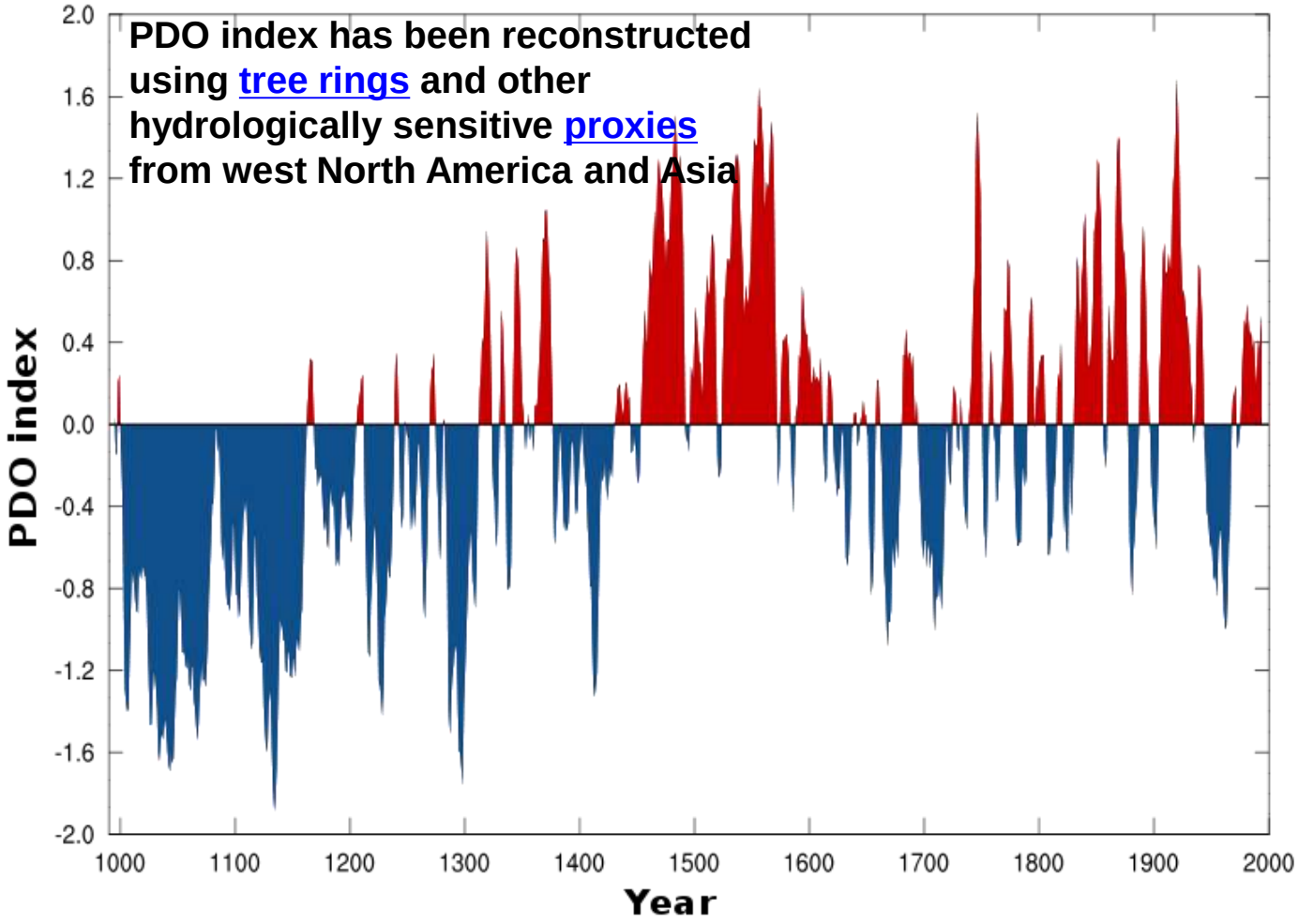
FIGURE 3-8 Global residual (GR) SST pattern, wind stress, and sea-level pressure, from which the linearly related ENSO variability has been removed. (From Zhang et al., 1997; reprinted with permission of the American Meteorological Society.)



Instrumental period


a) Pacific Decadal Oscillation (mantua)
PDO index
4
2
0
-2
-4
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Time (year)
b) PDO Wavelet Power Spectrum
Period (years)
0.25
0.5
1
2
4
8
16
32
64
128
256
Time (year)
c) Global Wavelet Spectrum
0
20
40
60
Power (degC^2)

Dewes, C. 2007


PDO index has been reconstructed using tree rings and other hydrologically sensitive proxies from west North America and Asia
PDO index
2.0
1.6
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
-1.6
-2.0
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Year

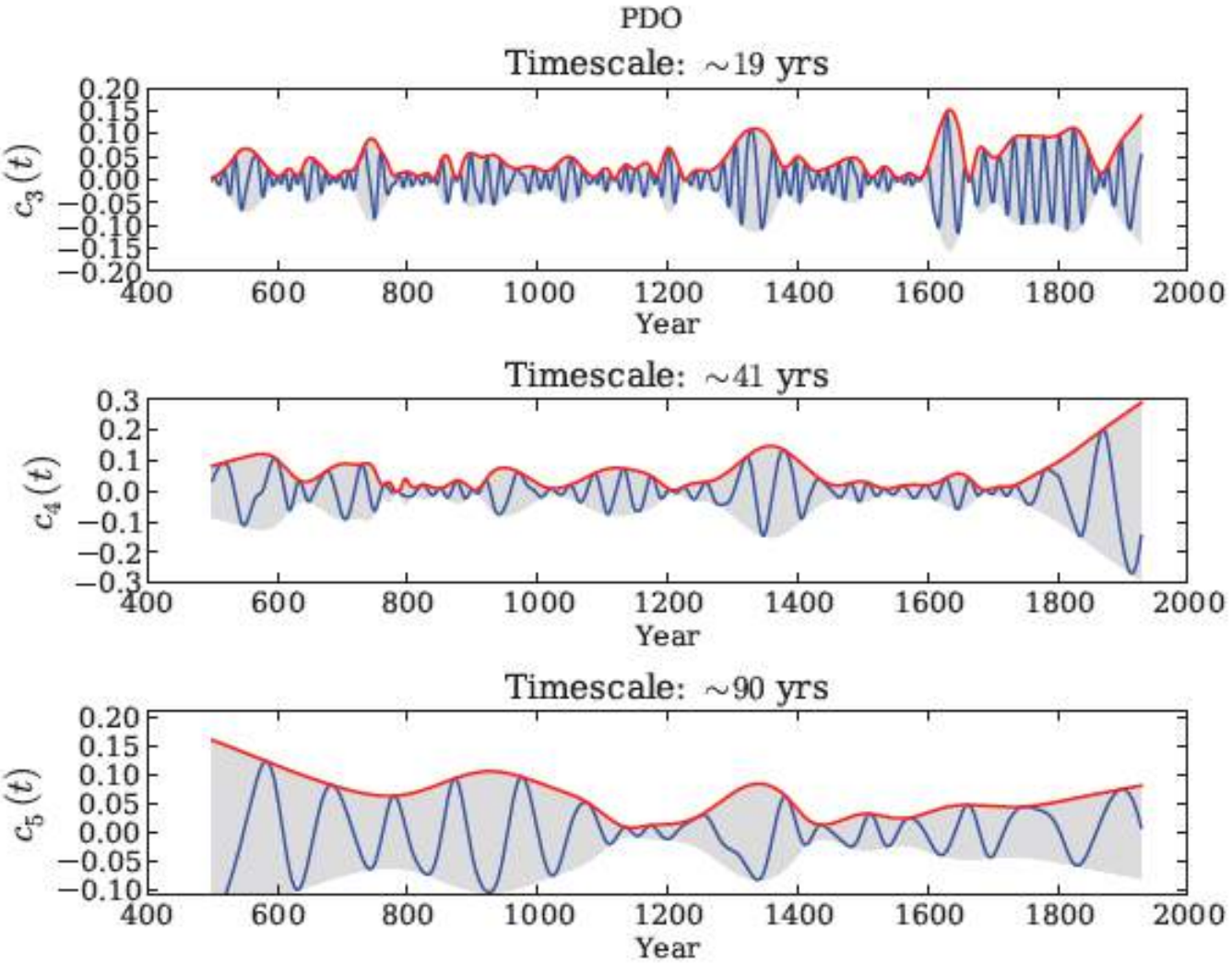
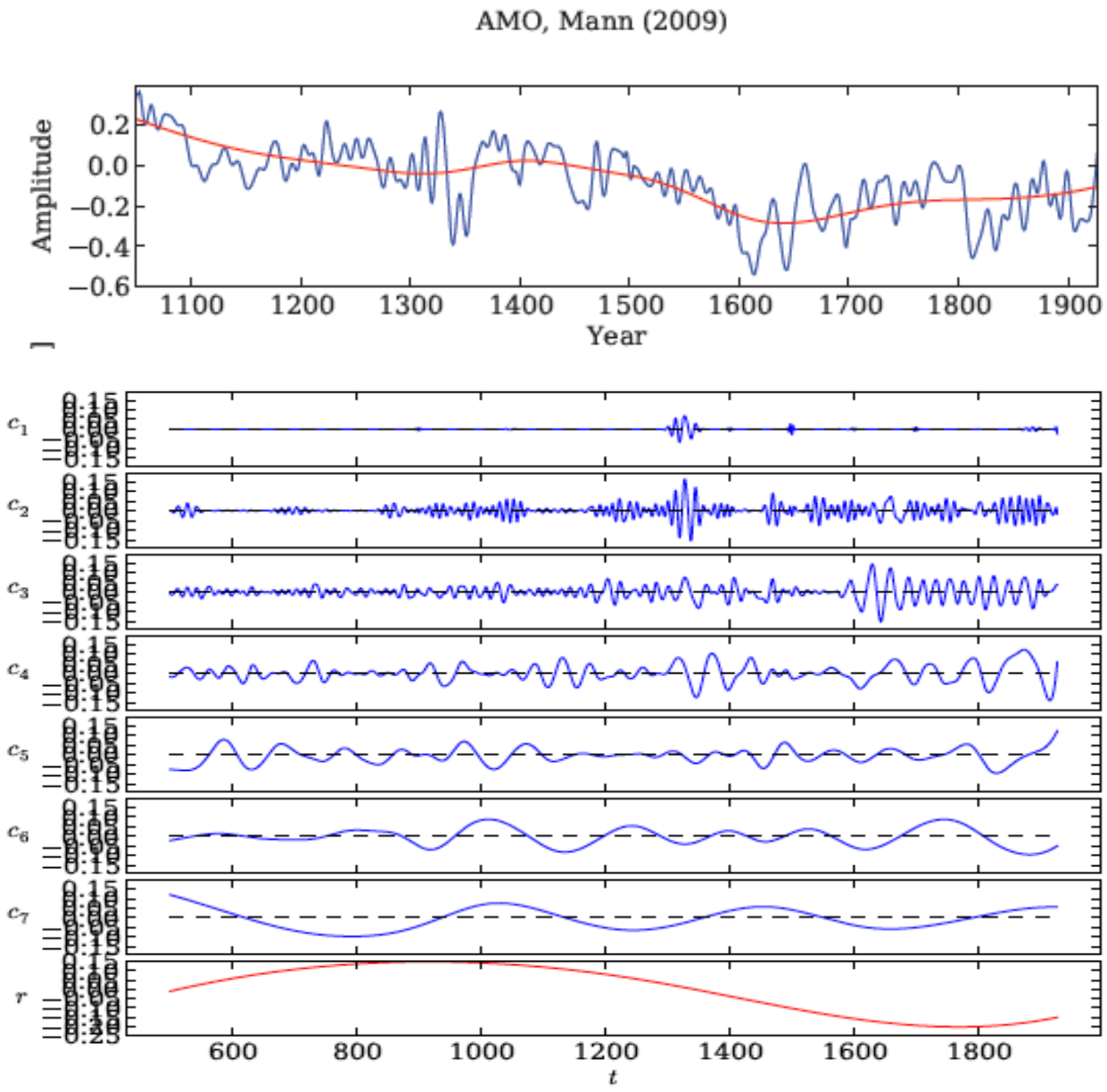


Figure: Empirical Mode Decomposition: PDO

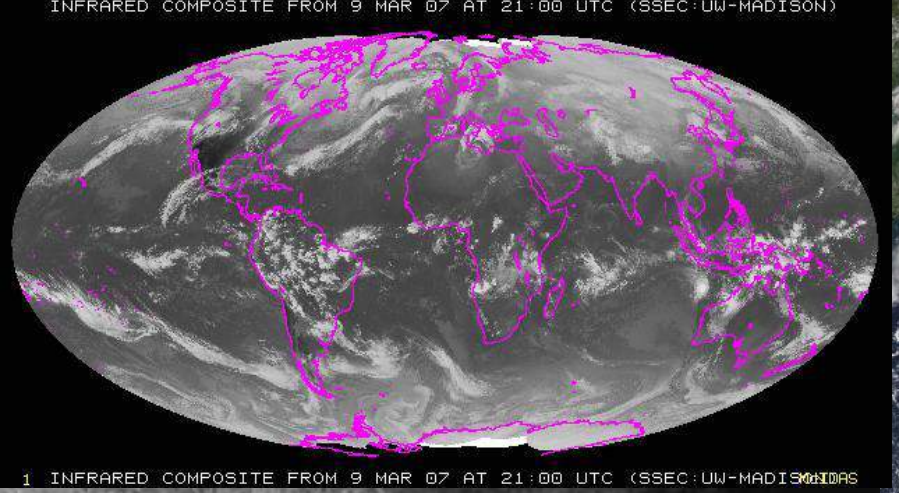
Welter, G. 2014

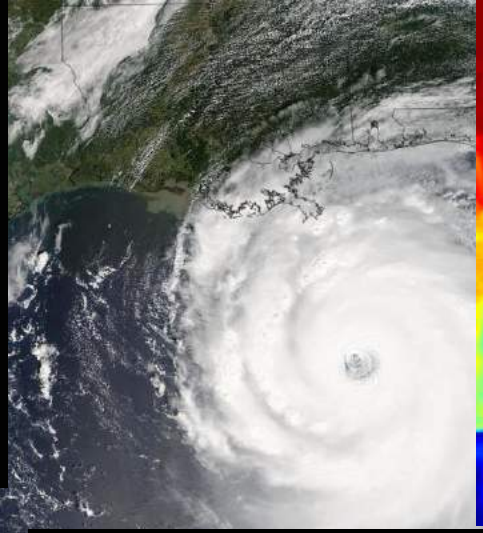


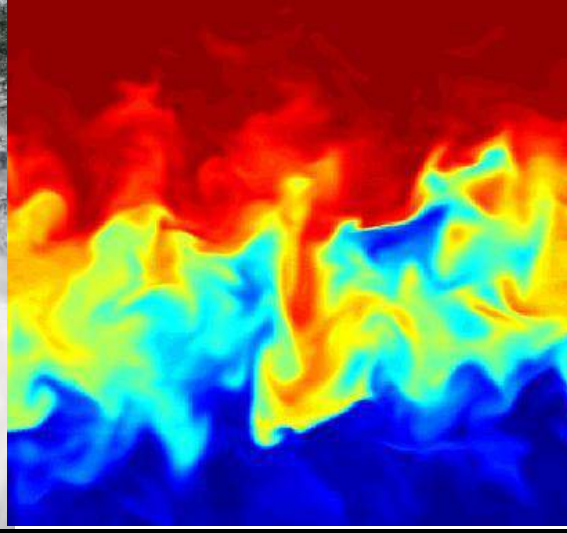
Indicators of the Atlantic Meridional Overtunrning circulation

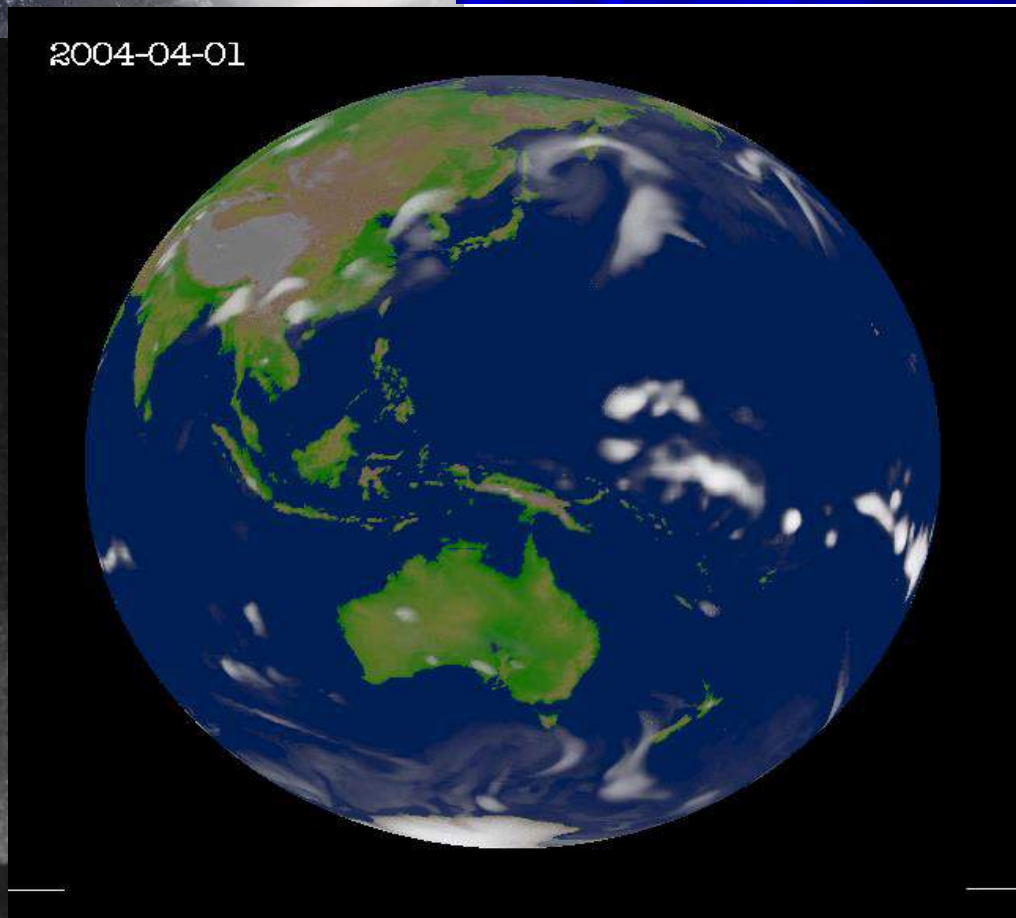
- **Lots of papers on pattern identification of climatic modes: EOF's, SOM, DNN (space/time)**
    - **Also on possible role of solar cycles**
- **Causality studies - networking**
- **However, limited success in predicting temporal evolution …. Transitions ….of climatic patterns and also weather patterns on time scale beyond a few days.**

Why?

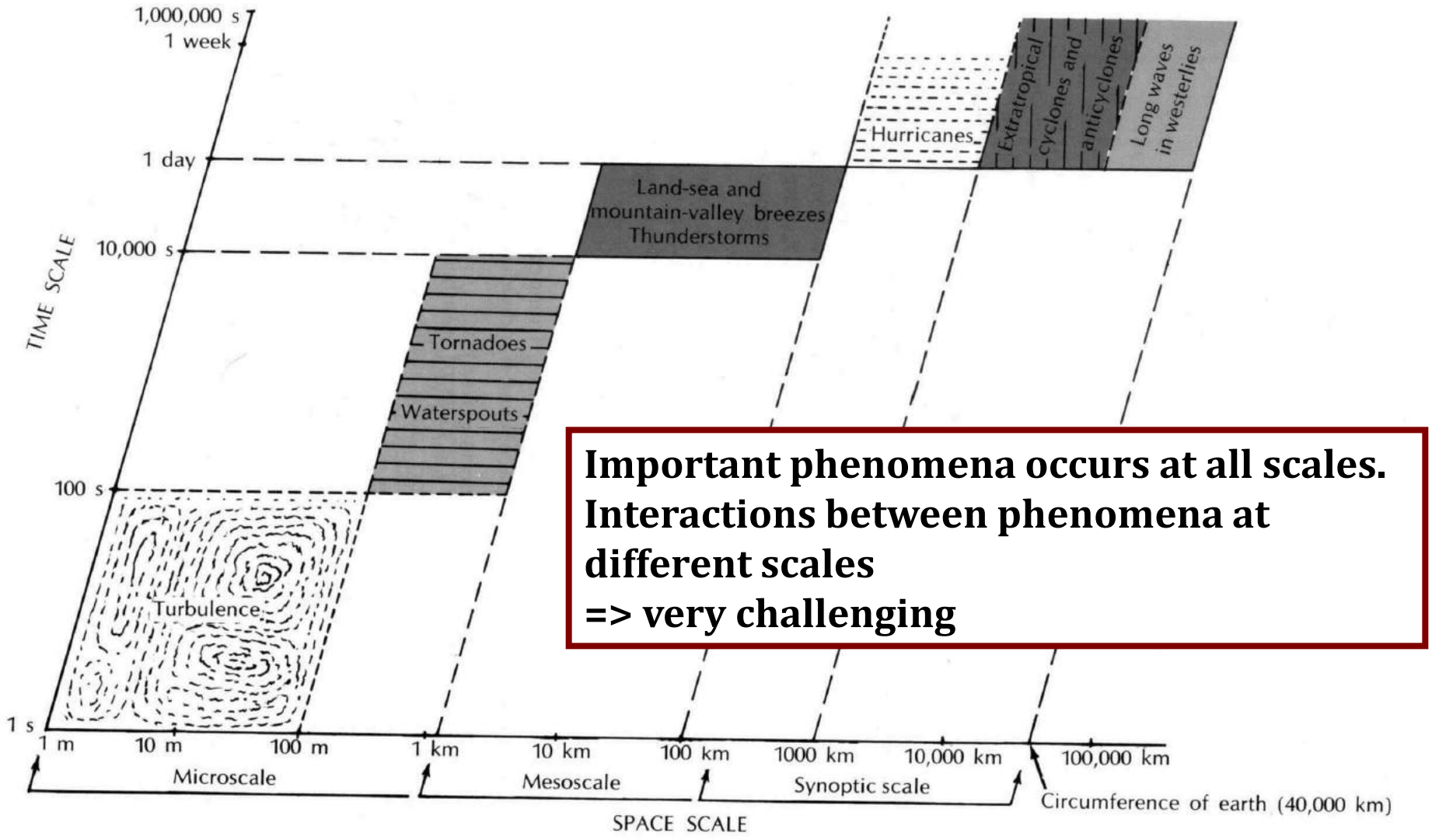

INFRARED COMPOSITE FROM 9 MAR 07 AT 21:00 UTC (SSEC:UW-MADISON)
1 INFRARED COMPOSITE FROM 9 MAR 07 AT 21:00 UTC (SSEC:UW-MADIS





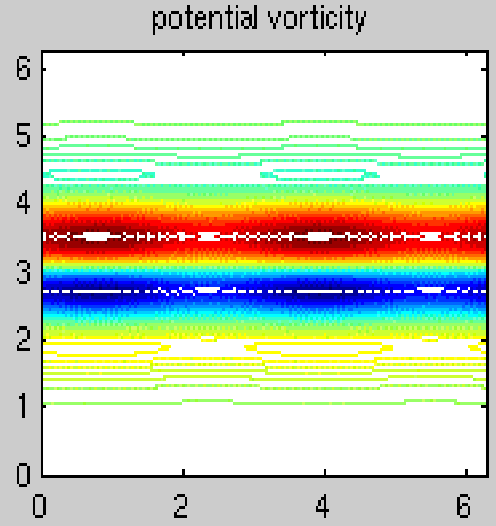

2004-04-01

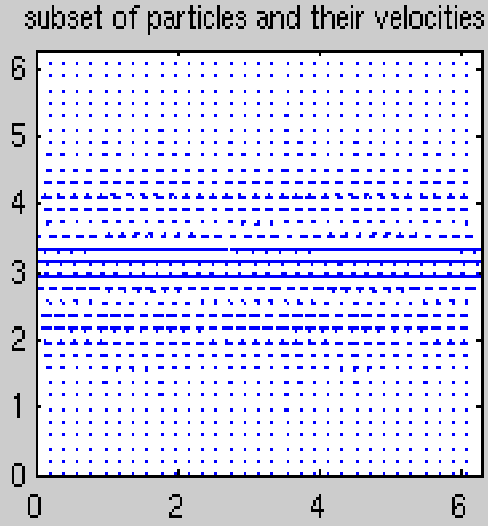
Scales of Atmospheric Processes

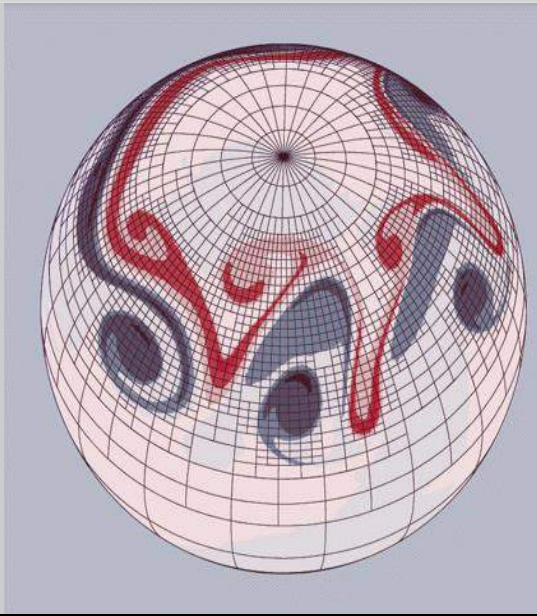


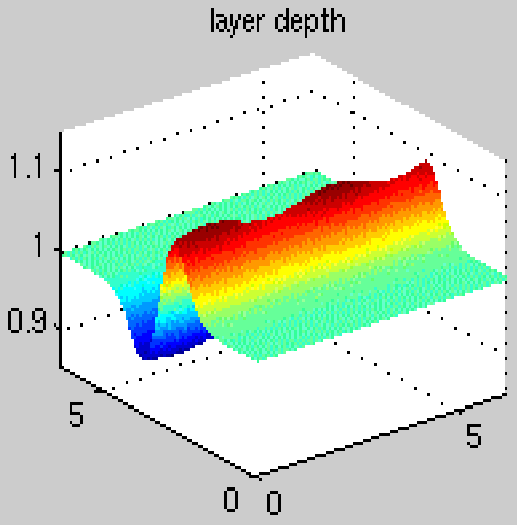
Important phenomena occurs at all scales.
Interactions between phenomena at different scales
=> very challenging

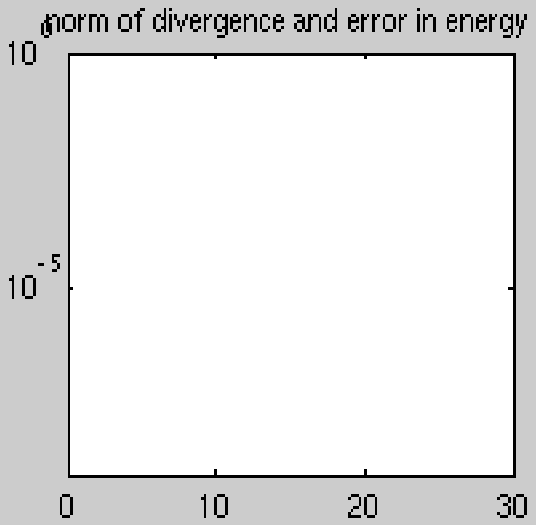
Changing scales – physical instability and non-linearity

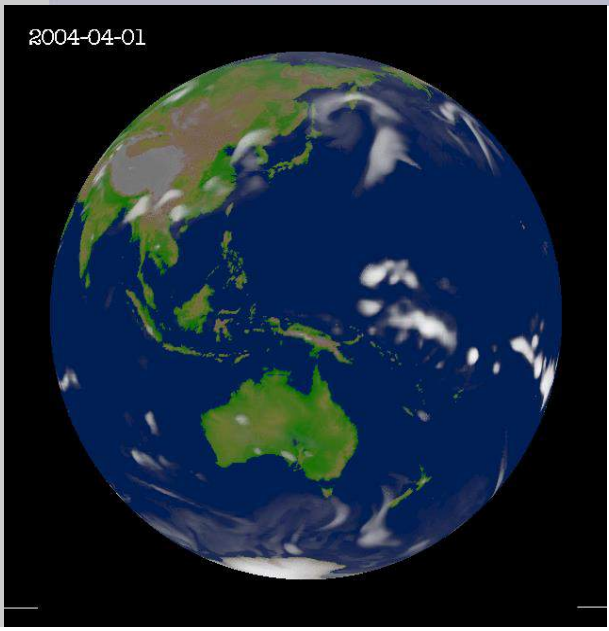




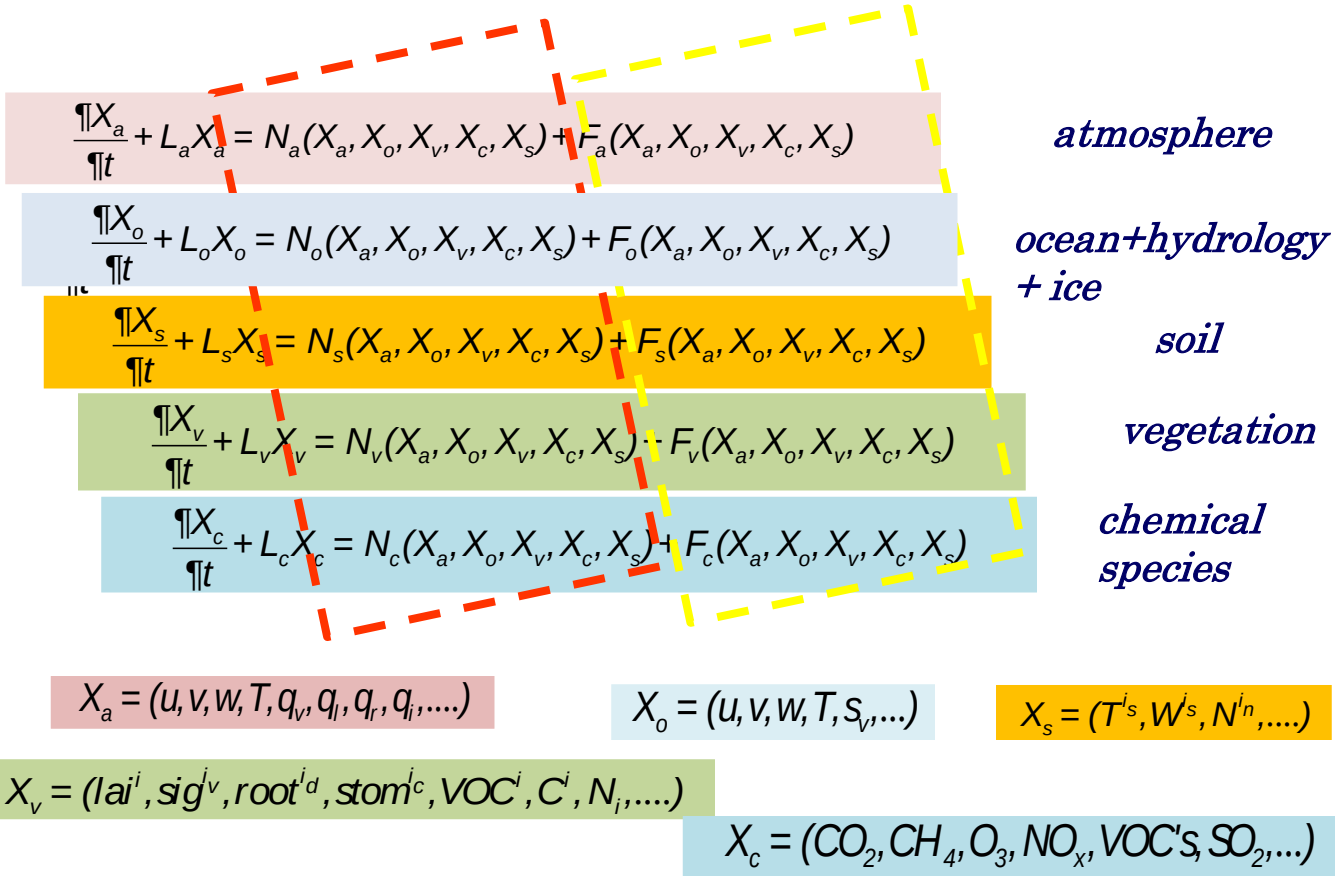








Modeling the Earth Atmosphere System



Non-linearity of the weather/climate prediction problem

$$\frac{\partial \xi}{\partial t} + \Omega \xi = N \qquad (1)$$

Eigenfunctions of the linear operator form an orthogonal and complete set

$$\xi(x,y,t) = \sum_a c_a(t)\xi_a(y)e^{ik_a x} \qquad (2)$$

$c_a(t)$ → **Spectral amplitudes**

Conjugate modes (a,a*) ⇒ a = (k_a, ω_a, n_a,r_a)

a* = (-k_a, -ω_a, n_a,r_a)

Two conjugate modes are mathematically independent and physically equivalent ⇒ $c_{a^*}(t) = -c_a^*(t)$

Substituting (2) in (1) – vector form of the model, multiplication by the conjugate of the eigenfunctions $\xi_a(y)$ and global integration lead to:

$$\frac{dc_a(t)}{dt} = i\omega_a c_a(t) + \sum_b \sum_c \sigma_a^{bc} c_b^*(t) c_c^*(t) \qquad (3)$$

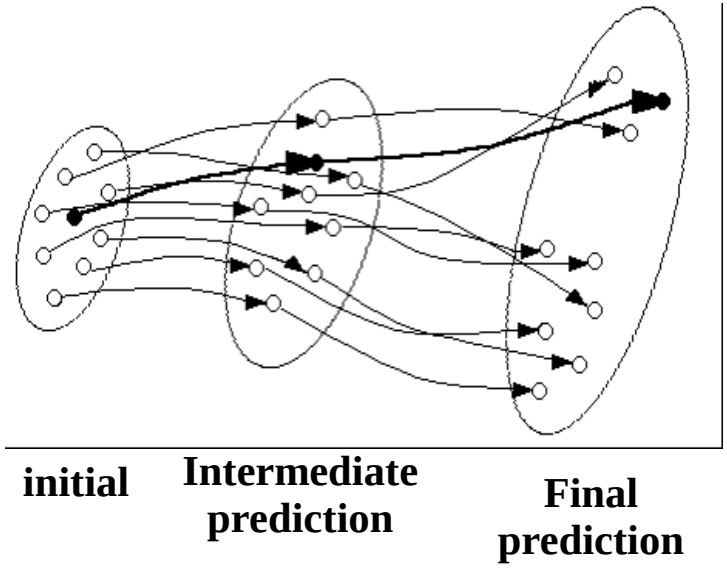
$$\sigma_a^{bc} = -\delta_{abc}\left\langle \left\{ u_b ik_c \xi_c + \mathrm{v}_b \frac{d\xi_c}{dy} + [0,0,1]^T \phi_b \left(ik_c u_c + \frac{d\mathrm{v}_c}{dy} \right) + CP \right\} \bullet \xi_a^* \right\rangle$$

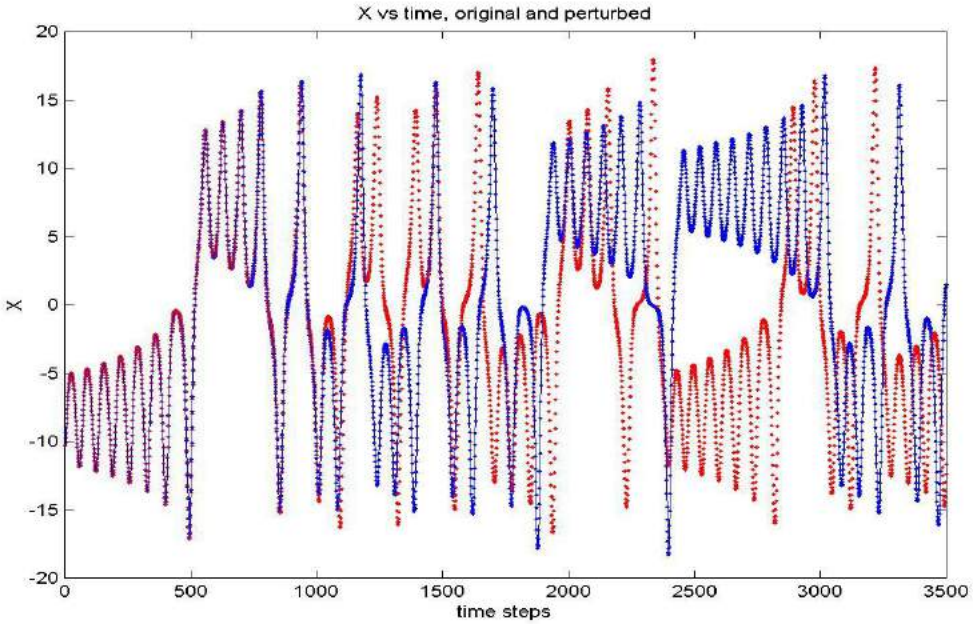
Coupling coefficient among the modes a, b and c. These coefficients are real and invariant under permutations of the superscripts bc.

Typical Dimension in Operational Models: 10^{10-11}

Lorenz (1963, 1965, 1969):

- *Governing equations show strong dependency on the initial conditions;*
- *Slightly different initial conditions lead to significant changes in the forecast after a few days;*
- *Lorenz hypothesizes that 10 days would be the predictability limit…*

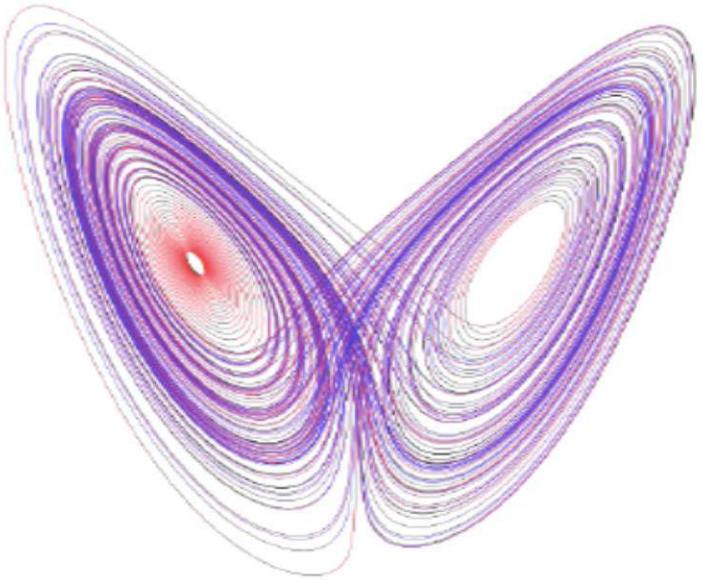




How can we predict weather beyond 10 days?
- effect of slow forcing in Lorenz's model

'Predictability in the midst of chaos'

If we can't predict the weather beyond the next week or so, why is it possible to make seasonal forecasts?



What is the impact of boundary forcing, *f*, on Lorenz Attractor?

$$\dot{X} = -\sigma X + \sigma Y + f$$

$$\dot{Y} = -XZ + rX - Y + f$$

$$\dot{Z} = XY - bZ$$

Birth of Seasonal Forecasting

'Predictability of Monsoons'

J. G. Charney and J. Shukla, 1981

Nonlinearity => energy transfer among scales

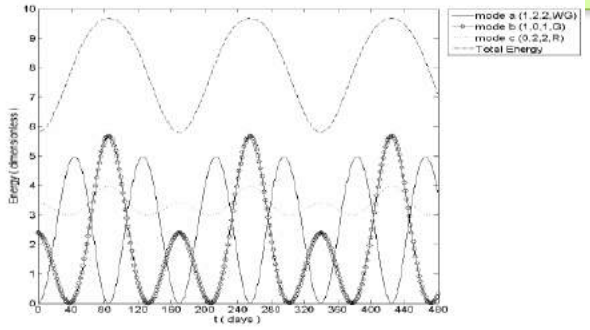
Dynamics of resonant interaction

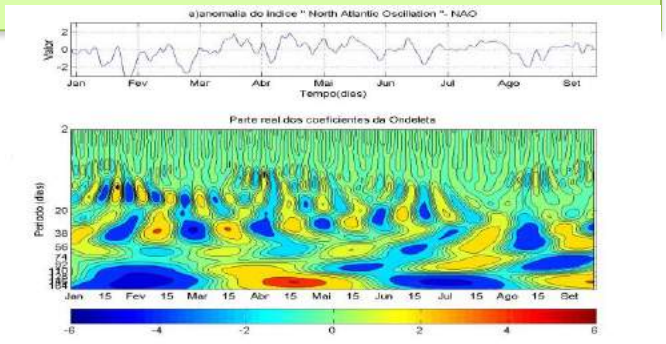
- ***Examples:***
  - ***Interaction between slow (O(5-7days)) and fast modes (O(1 day or less)) => intraseasonal scales (20-60 days) Raupp et al. (2008)***
  - ***Importance of diurnal variation leading to energy in intrasesonal time scales (Raupp and Silva Dias, 2009)***
  - ***Coupled ocean/atmosphere simplified models: interaction between intraseasonal scale (O(20-60d)) with interannual (El Nino/La Nina) - O (2-3 yr) => decadal/multidecadal time scales (Enver et al. 2017)***

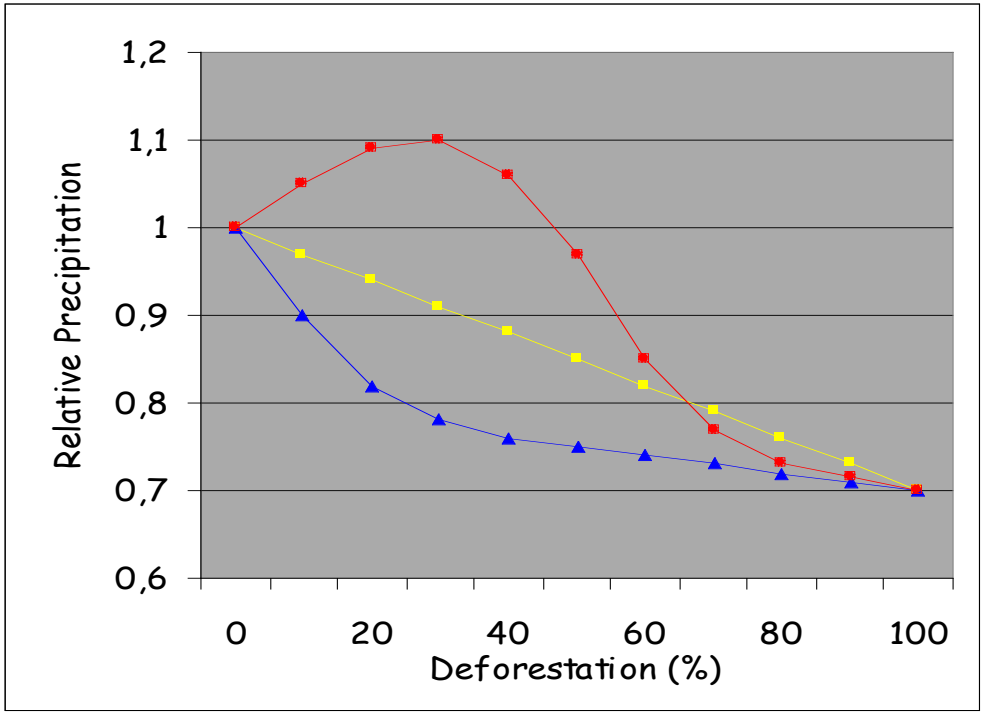
$$c_a^{\,2}\frac{dA_a}{d\tau} = A_b A_c \eta_a^{bc} + q_a \delta\left(\frac{\varpi_a \pm \nu_j}{\varepsilon}\right)$$

$$c_b^{\,2}\frac{dA_b}{d\tau} = A_a A_c^{*} \eta_b^{ac} + q_b \delta\left(\frac{\varpi_b \pm \nu_i}{\varepsilon}\right)$$

$$c_c^{\,2}\frac{dA_c}{d\tau} = A_a A_b^{*} \eta_c^{ab} + q_c \delta\left(\frac{\varpi_a \pm \nu_l}{\varepsilon}\right)$$

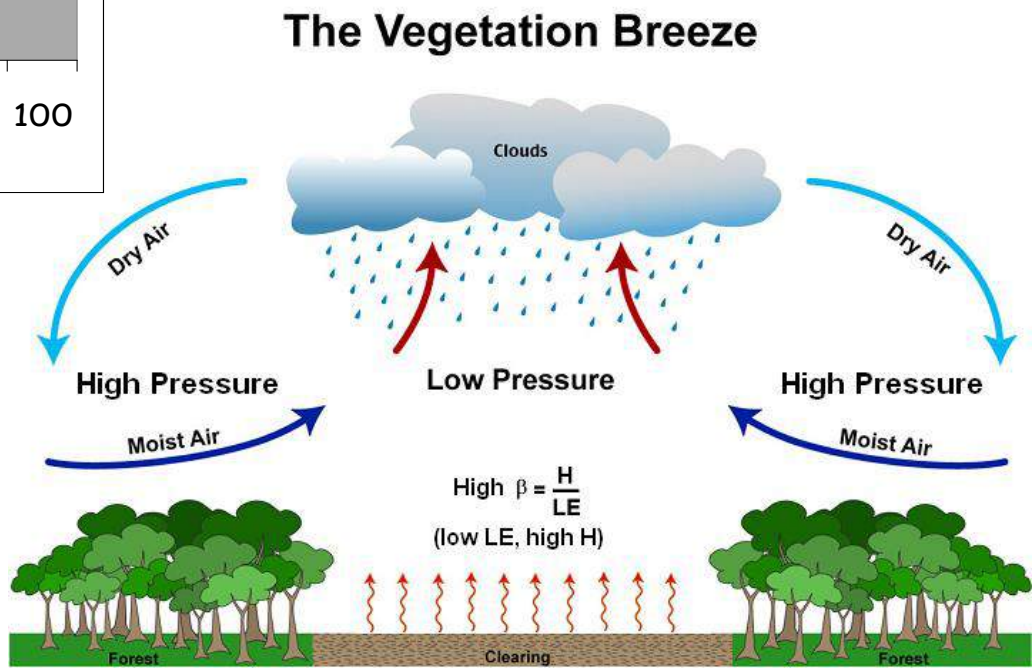




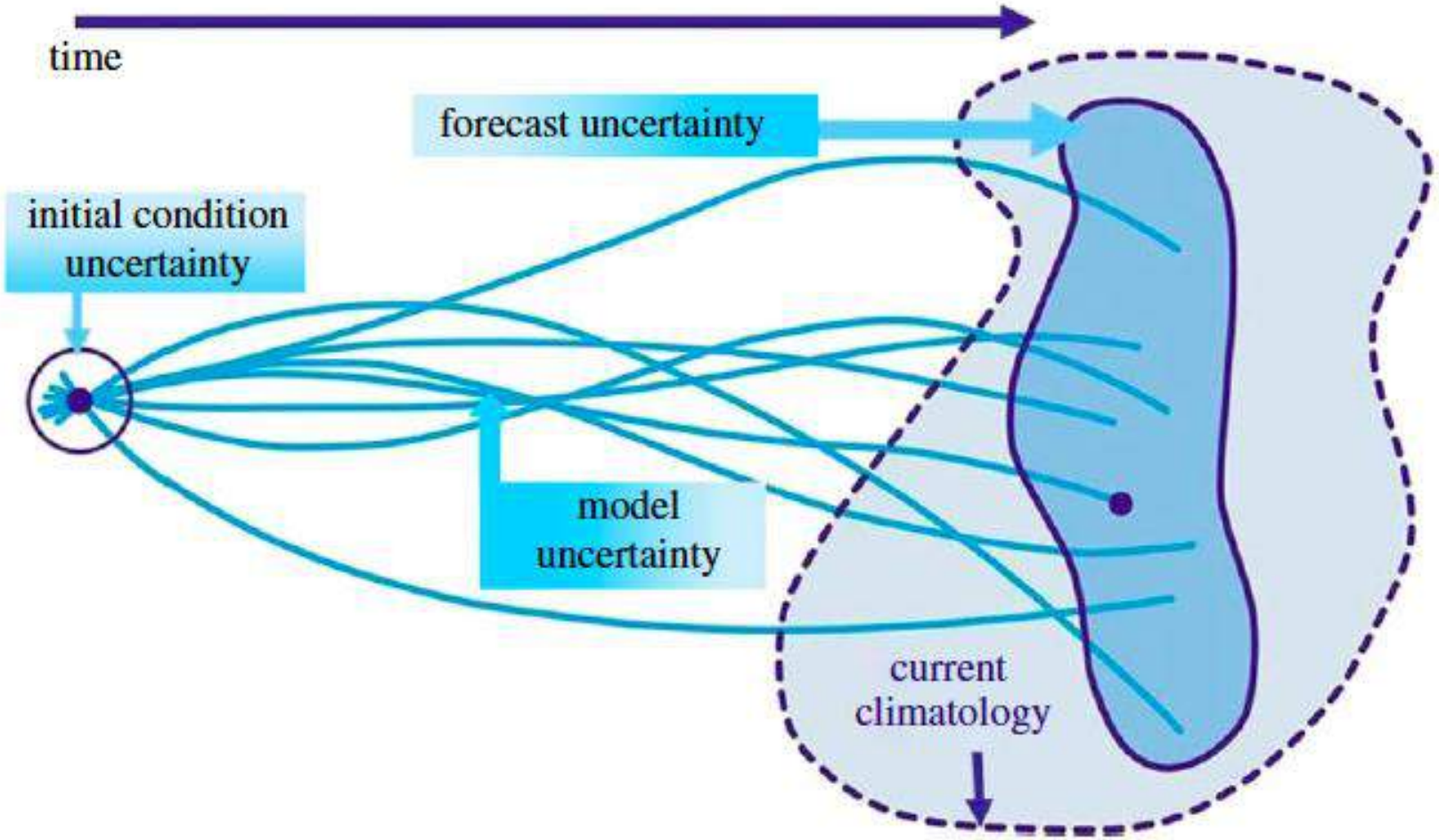


Conceptual impact of deforestation on relative precipitation. The three curves indicate different models, among many other possible ones (Avissar, P. Silva Dias, M. Silva Dias and Nobre et al 2002 – JGR – LBA special issue)

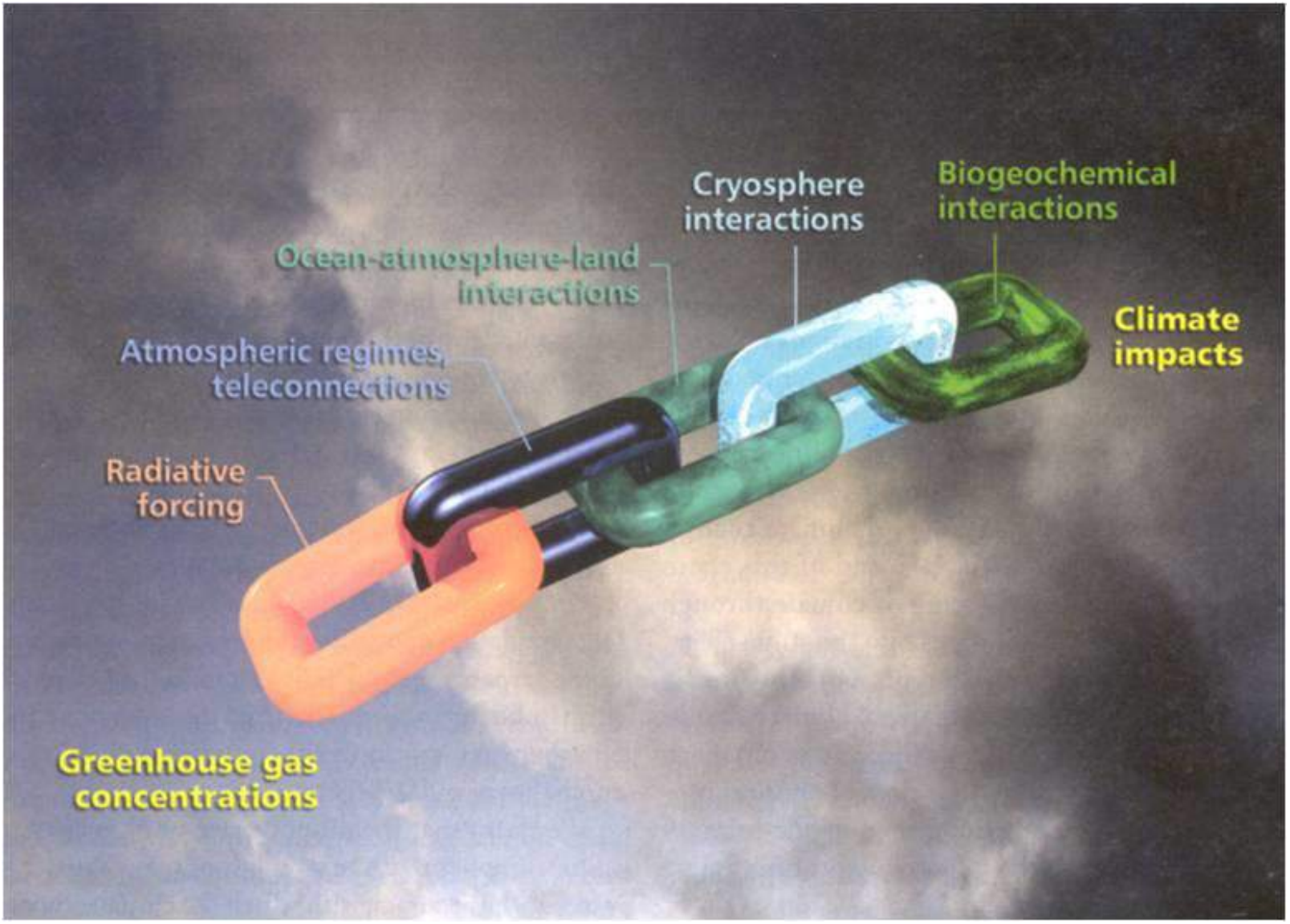
Precipitation change as a function of deforestation



Message: realist deforestation patterns may lead to precipitation increase. Total substitution by pasture leads to drying. Need further studies on the impact of palm trees and other biofuel crops.


time
forecast uncertainty
initial condition uncertainty
model uncertainty
current climatology

Climate is a complex system: interaction among several components



Bull. American Met. Soc. - 2008

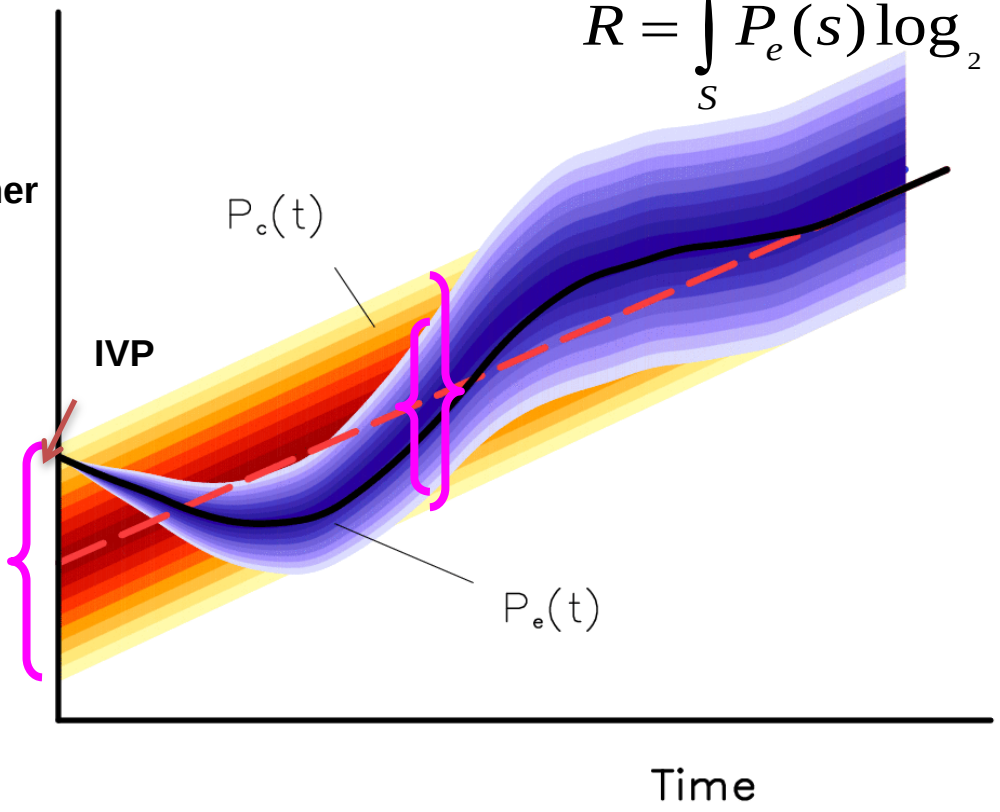
A HARD PROBLEM

Schematic Prediction (absent climate drift):

Initial value, Climate change & Total Predictability

$$R = \int_S P_e(s) \log_2 \left[\frac{P_e(s)}{P_C(s)} \right] ds \frac{1}{n}$$

WORKS FOR ANY 'Climate' our Weather PREDICTION PROBLEM (NON STATIONARY)



Add information from initial state distribution (uncertain) + model uncertainty

How long is this information retained? (dynamics of the problem)

Predictability question

"Mean" and "Spread" contributions

to predictive 'information'

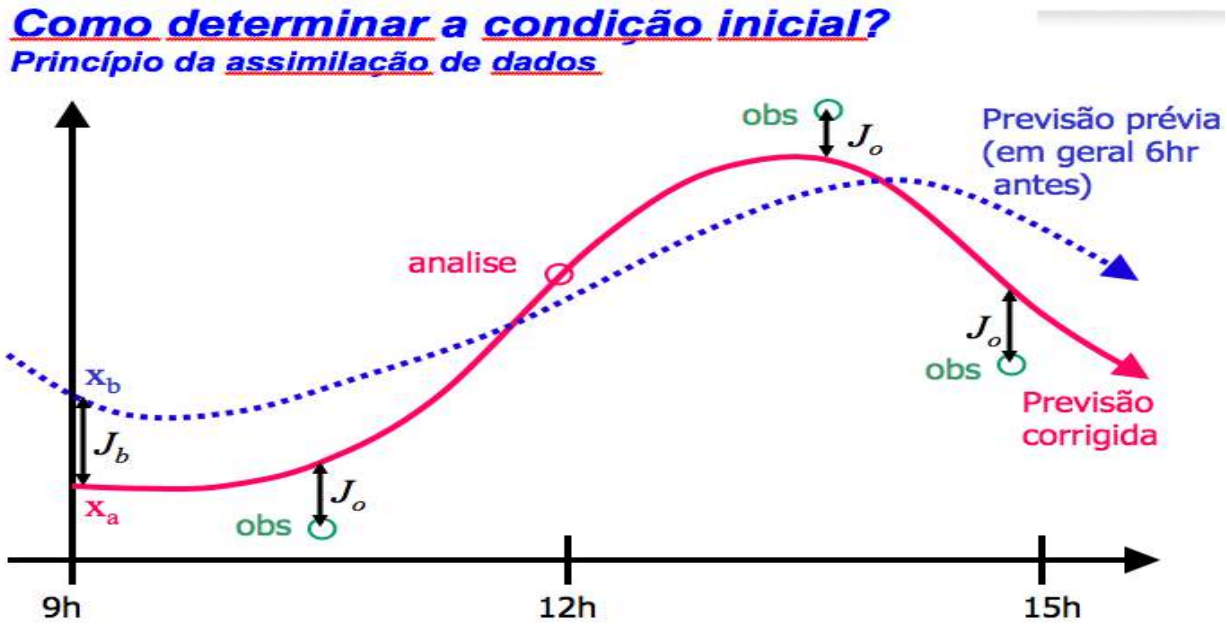
G. Branstator
Haiyan Teng

Dealing with Uncertainties -1 :

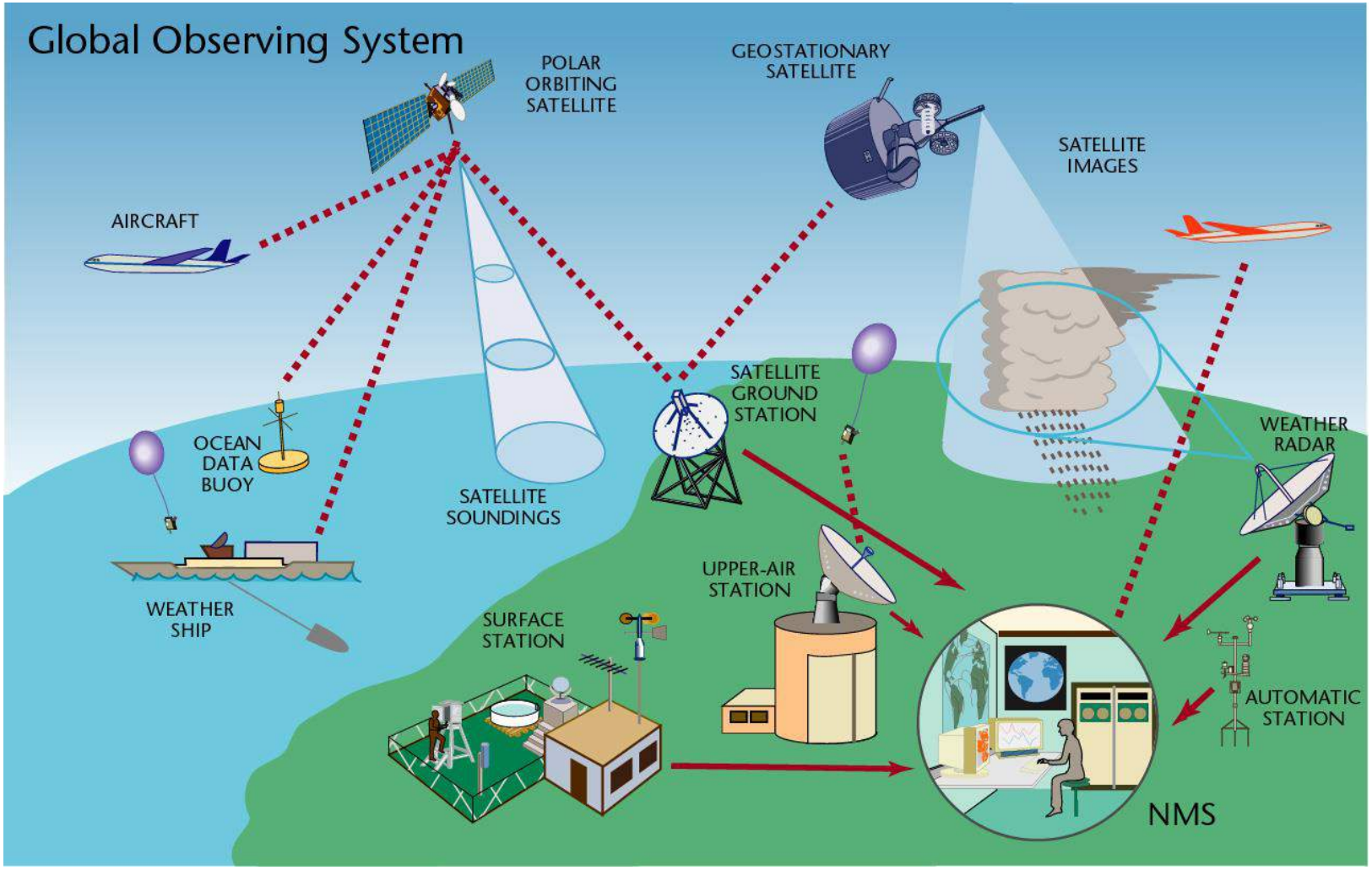
1.Initial Value Problem: what is the initial state of the atmosphere?

$DX/Dt = f(X,t)$ $\quad X(t=0) = Xo \quad \Rightarrow$

Observations + initial guess (short forecast)



World Meteorological Organization data types

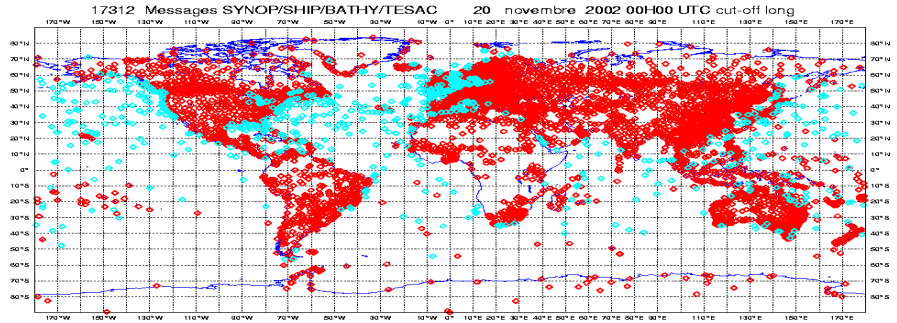


http://www.wmo.ch/web/www/OSY/GOS.html

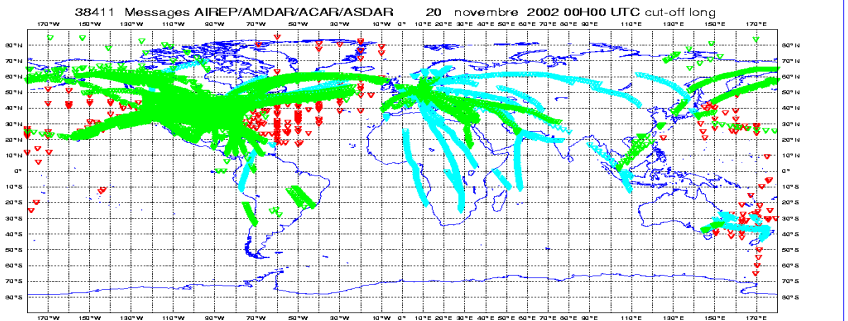
Major Categories of data sources in Atmospheric Sciences

- **Satellite Remote Sensing Data**
- **Assimilated Datasets (Validation Data)**
- **Model Output:**
  - **Weather Forecasts;**
  - **Climate Prediction. (S2S Subseasonal to seasonal – interannual – decadal);**
  - **Climate Projections for the future (Climate Change due Human influence).**

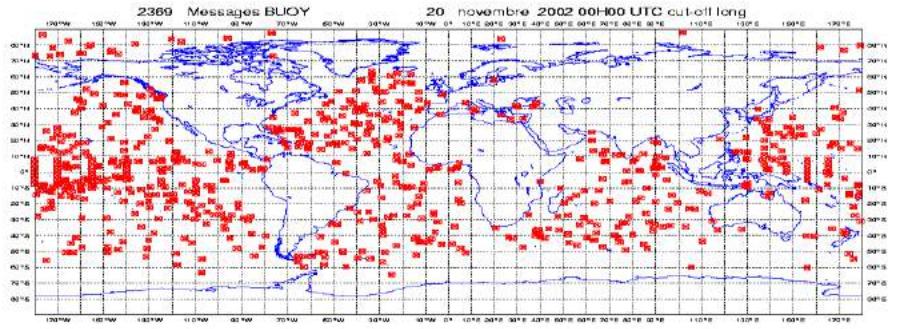
SYNOP et SHIP surface P,T,HU,wind



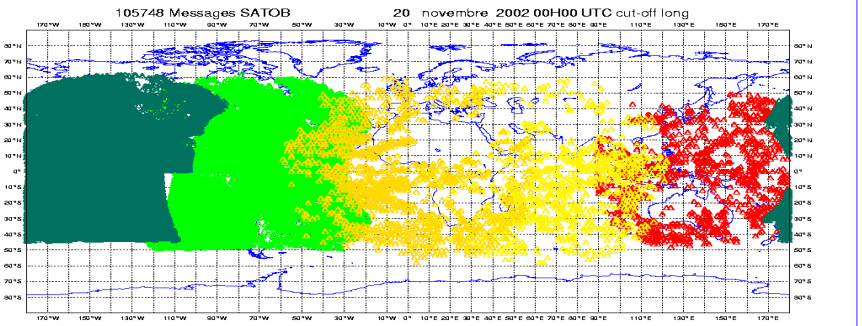
aircraft T,wind



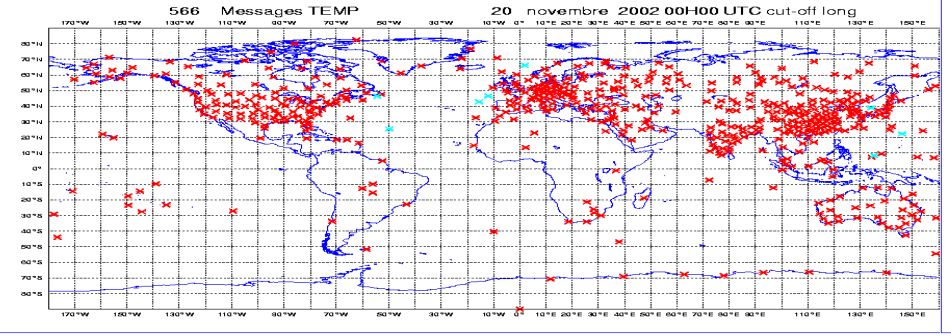
buoys surface P,T,HU,wind



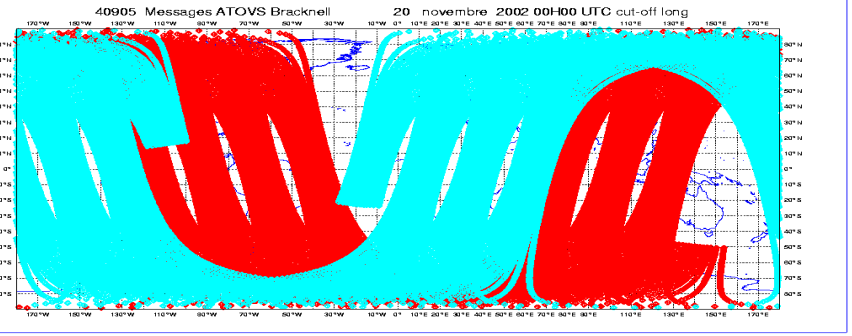
Geostationary satellite winds



radiosoundings P,T,Hu,wind



Radiances ATOVS NOAA



Under-determination (ex. CMC)

X = model state vector

Model	Lat x long x lev x variables
CMC global oper.	800x600x58x4 $=1x10^8$
CMC meso-strato	800x600x80x4 $=1.5x10^8$

Z = observation vector

Data	Reports x items x levels
Sondes,pibal	720x5x27
AMSU-A,B	14000x12
SM, ships, buoys	7000x5
aircraft	19000x3x18
GOES	5000x1
Scatterometer	7000x2
Sat. winds	21000x2
TOTAL	$1.3x10^6$

$$\frac{N_X}{N_Z} \approx 75$$

- Cannot do X=f(Y), must do Y=f(X)
- Problem is underdetermined, always will be
- Need more information: prior knowledge, time evolution, nonlinear coupling

Optimal Interpolation

N×1 N×1 N×M M×1 M×N N×1

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}\left(\mathbf{z} - H(\mathbf{x}^b)\right)$$

Analysis vector

Background or model forecast

Weight matrix

Observation vector

Observation operator

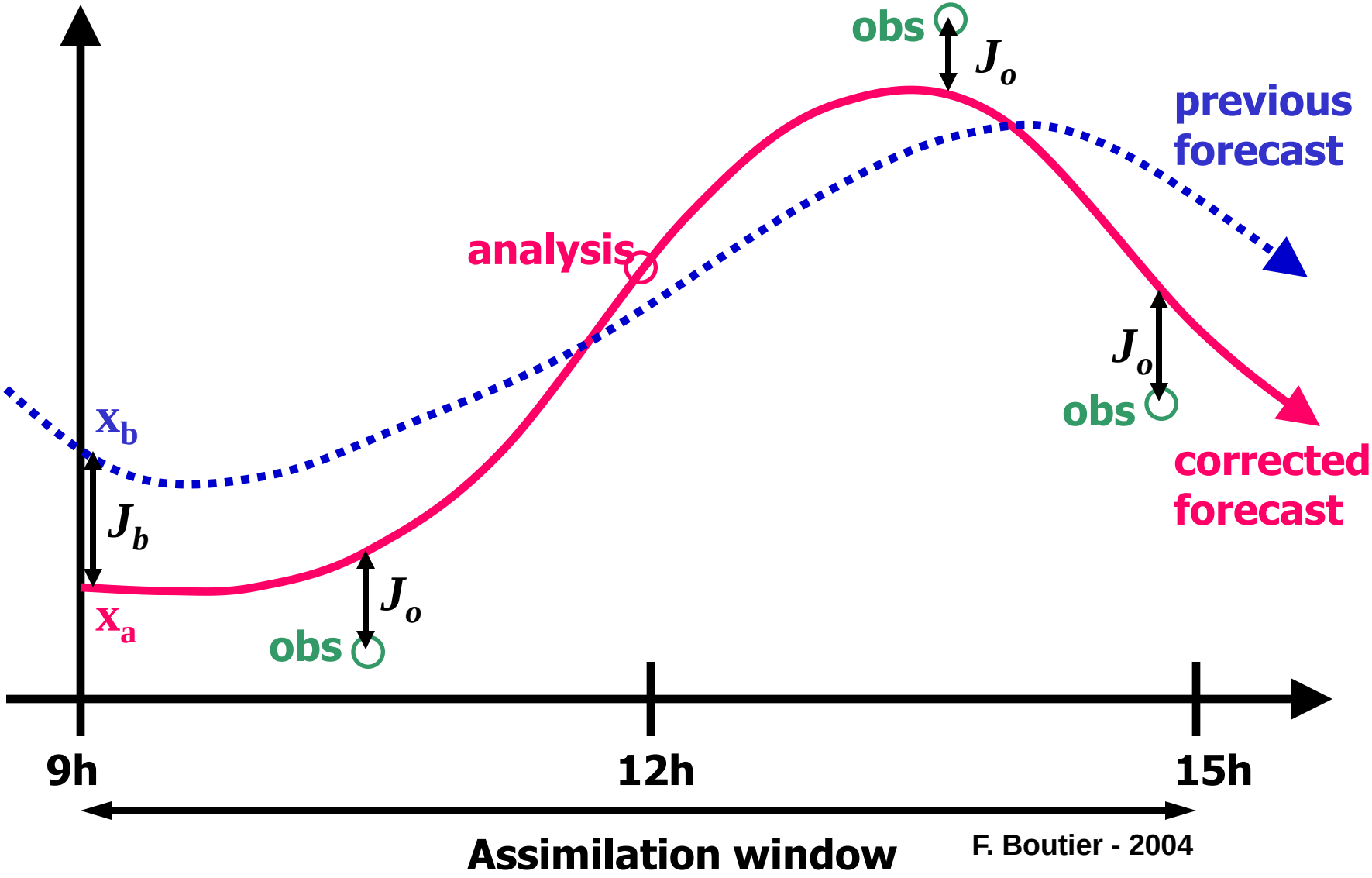
NxM NxN MxM

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right)^{-1}$$

$N \approx 10^8$ $M \approx 10^6$

Can't invert!

Principle of 4D-VAR assimilation



3D-Var used in operational forecasting centers

$$J = \min \frac{1}{2}[(\mathbf{x}_b - \mathbf{x}_a)^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}_b - \mathbf{x}_a) + (\mathbf{y}_o - H\mathbf{x}_a)^{\mathrm{T}} \mathbf{R}^{-1} (y_o - H\mathbf{x}_a)]$$

Distance to forecast **Distance to observations**

- **x is a model state vector, with $10^{8\text{-}11}$d.o.f. x_a minimizes J**
- **y_o is the set of observations, with $10^{6\text{-}9}$ d.o.f.**
- **R is the observational error covariance**
- **B the forecast error covariance.**
- **In 3D-Var B is assumed to be constant: it does not include "errors of the day"**
- **The methods that allow B and A to evolve are very expensive: 4D-Var and Kalman Filtering.**

F. Boutier - 2004

3D-Var

$$J = \min \frac{1}{2}[(\mathbf{x}^a - \mathbf{x}^b)^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}^a - \mathbf{x}^b) + (H\mathbf{x}^a - \mathbf{y})^{\mathrm{T}} \mathbf{R}^{-1} (H\mathbf{x}^a - \mathbf{y})]$$

Distance to forecast — Distance to observations

at the analysis time

4D-Var

$$J = \min \frac{1}{2}[(\mathbf{x}_0 - \mathbf{x}_0^b)^{\mathrm{T}} \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_0^b) + \sum_{i=1}^{s} (H\mathbf{x}_i - \mathbf{y}_i)^{\mathrm{T}} \mathbf{R}_i^{-1} (H\mathbf{x}_i - \mathbf{y}_i)]$$

Distance to background at the initial time

Distance to observations in a **time window interval t_0-t_1**

Control variable $\mathbf{x}(t_0)$

Analysis $\mathbf{x}(t_1) = M[\mathbf{x}(t_0)]$

It seems like a simple change, but it is not! (e.g., adjoint)

What is B? It should be tuned…

Open access peer-reviewed chapter

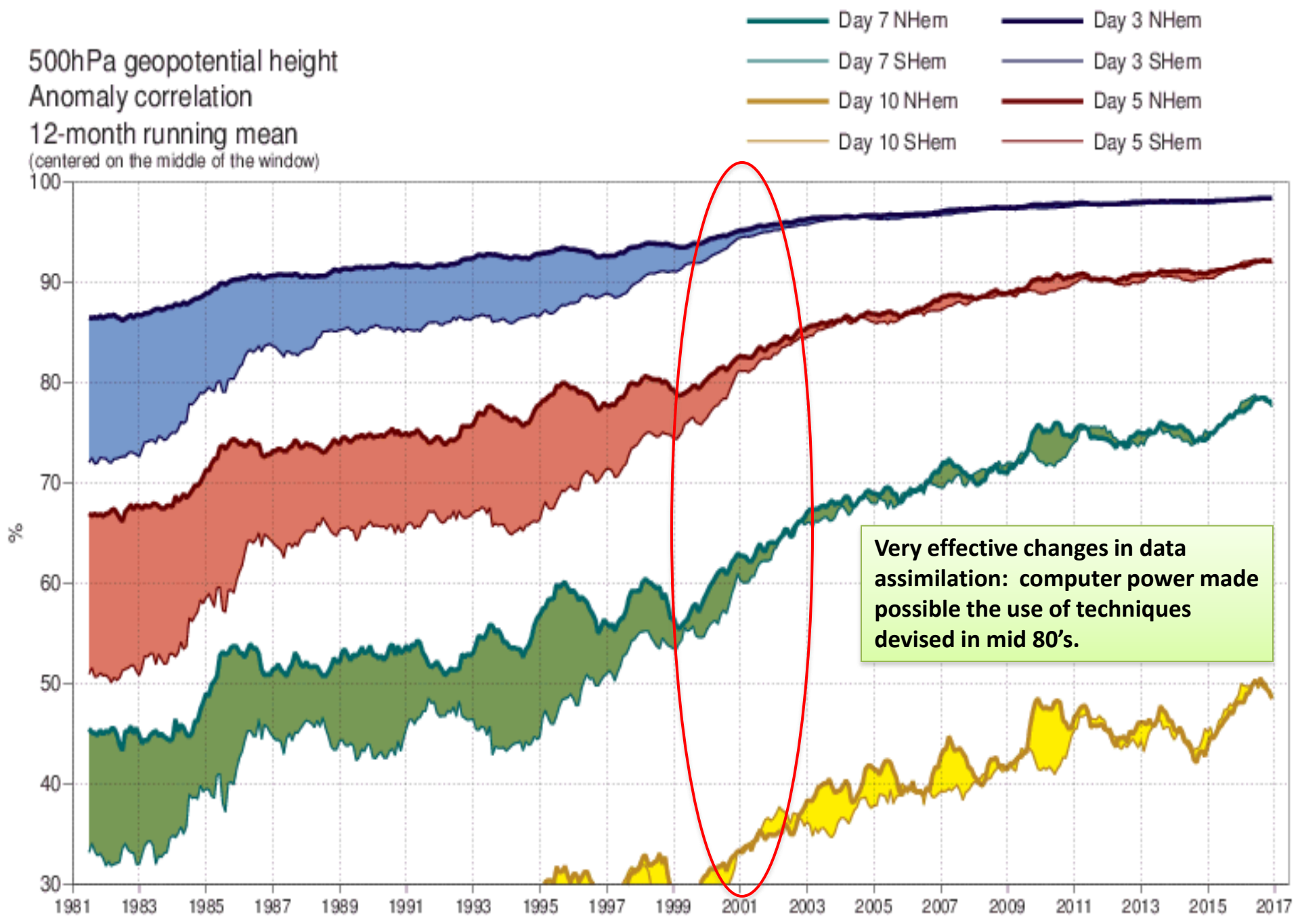
Data Assimilation by Artificial Neural Networks for an Atmospheric General Circulation Model

By Rosangela Saher Cintra and Haroldo F. de Campos Velho

Submitted: June 29th 2017 Reviewed: September 1st 2017 Published: February 28th 2018

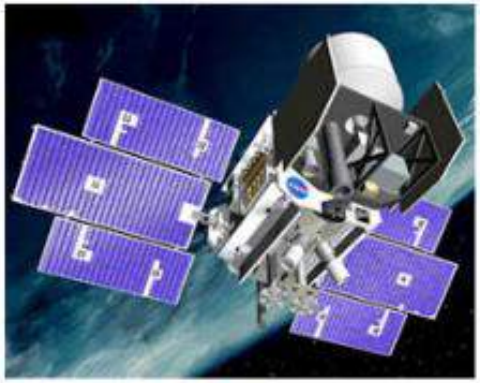
DOI: 10.5772/intechopen.70791

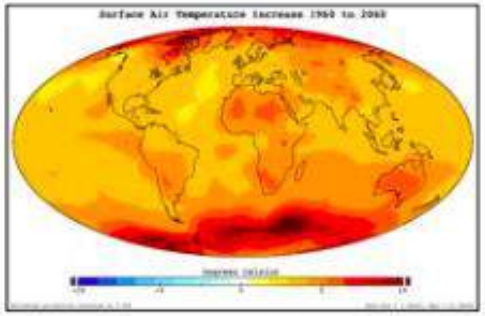
In spite of the initial condition and model uncertainties …..

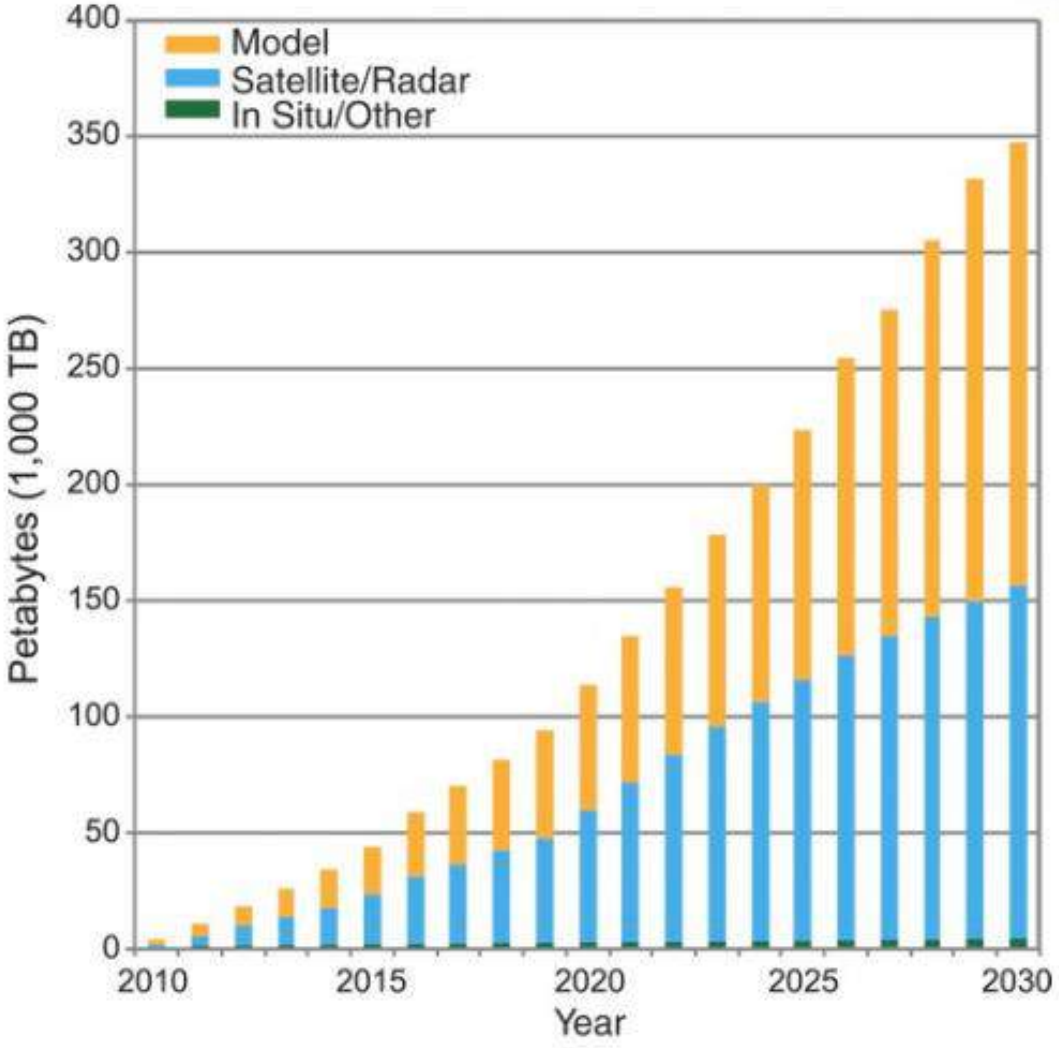


From ECMWF – European Center for Medium Range Weather Forecasting

Climate Change Data Alone is Projected to Grow to Nearly 350 Petabytes by 2030



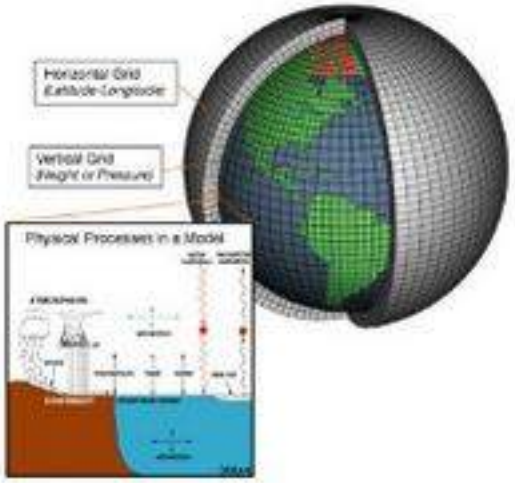




Weather prediction has been a big data problem for a long time

- Weather satellite and radar data
- Global weather prediction models
- Regional weather prediction models
- Surface, upper air, and aircraft data

Petabytes per day



Crowdsourcing of Weather Data: The New Big Data Frontier

- Example: atmospheric pressure from smartphones
- Potentially a billion sensors in a few years
- A number of smartphone apps allow folks to report current weather conditions

IBM, Google, ….several consulting companies are exploring the use of sensors not in the official

Dealing with uncertainties -2

2. Forecast model $f(X,t)$ -> $DX/Dt = f(X,t)$

$f(X,t)$ is not perfectly known:

- *Formulation of physical processes (radiation, turbulence, phase change of water – vapour, liquid, ice);*
- *Spatial resolution of the models (100km, 10km, 1km ….): how do we deal with upscaling?*
- *Numerical Methods: discretization*
- *F(X,t) is a multivariate function: physical interactions among components not completely understood – uncertainties in parameters.*

Uncertainty with respect to initial conditions:

- ***Ensemble forecasting: perturbation of initial conditions:***

Initial condition uncertainties: Ensemble forecasting/prediction

- Initial conditions not completely determined (X is a multivariate variable – large dimension - observations (Y) a few order of magnitudes smaller in dimension that X.

- One integration from a particular IC in general does not capture the complexity of the situation:

- Several integrations are needed, spanning the uncertainty in the model dynamics and initial conditions; perturbations/different models – merge with data assimilation uncertainties => **Ensemble Forecasting** – *huge amount of data!*

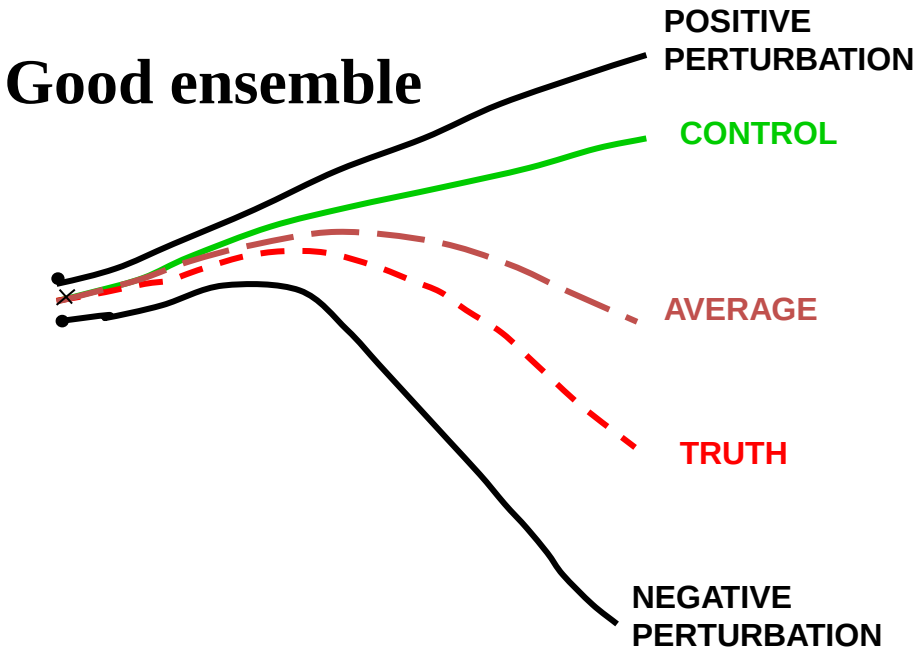
- Spread and differences between control and ensemble mean tell a lot about the system;
- **Need to be applied more generally in modelling the complex earth system (X_a, X_o, X_v, X_s, X_c, etc).**

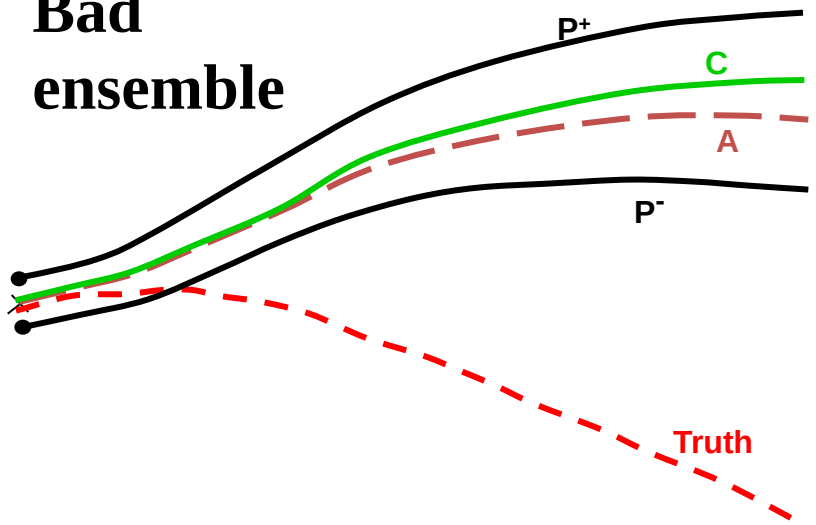
Components of ensemble forecasts

An ensemble forecast starts from initial perturbations to the analysis…

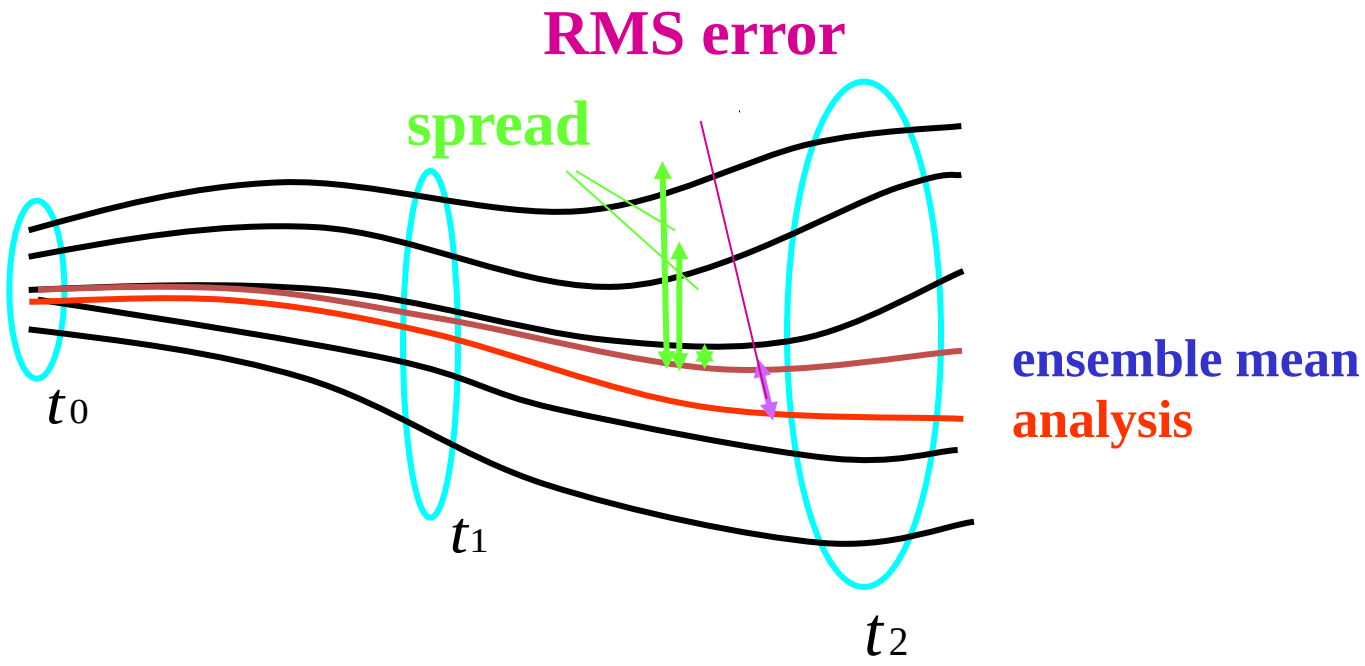
In a good ensemble "truth" looks like a member of the ensemble

The initial perturbations should reflect the analysis "errors of the day"





Representing initial state uncertainty by an ensemble of states



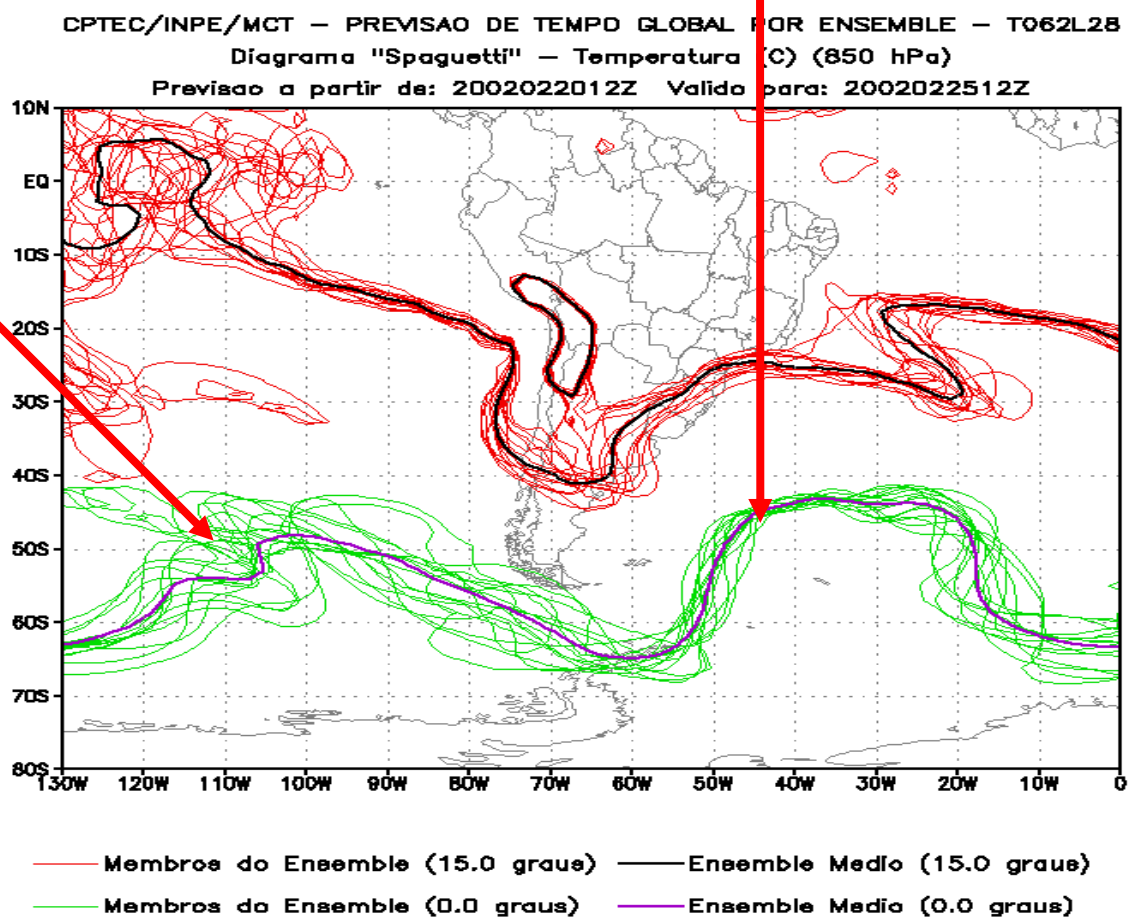
- Represent initial uncertainty by ensemble of atmospheric flow states
- Flow-dependence:
  - Predictable states should have small ensemble spread
  - Unpredictable states should have large ensemble spread
- ***Ensemble spread should grow like RMS error***

Example of ensemble prediction

Low predictability

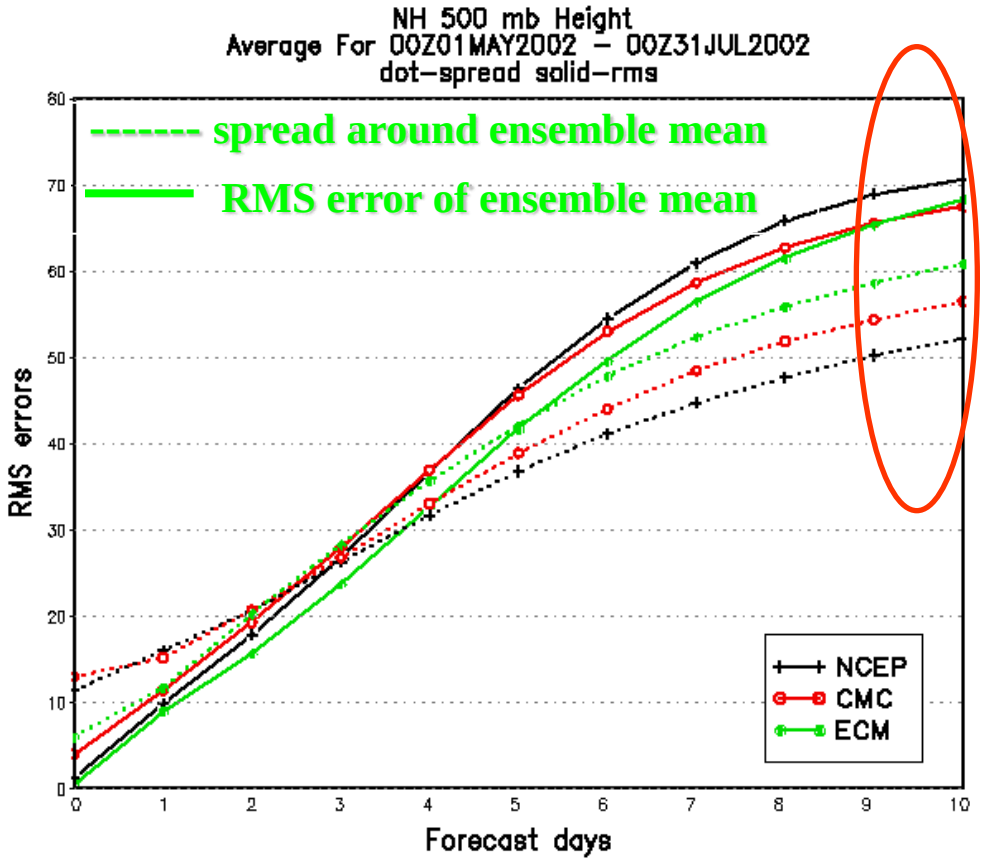
Operational product from Center for weather Prediction and Climate Studies - BRAZIL

Initial condition perturbation



Errors of the day tend to be localized (Patil et al, 2001)

Under-dispersion of the ensemble system



Buizza et al., 2004

The RMS error grows faster than the spread

- **Ensemble is under-dispersive**
- **Ensemble forecast is over-confident**

- **Under-dispersion is a form of model error**
- ***Forecast error = initial error + model error***

- A major restriction for accurate prediction is dynamical instability.

- Successful predictions in climate change science are hampered by the fact that the actual dynamics is a turbulent large-dimensional system with positive Lyapunov exponents on essentially all spatiotemporal scales;

- With unstable systems, the Monte Carlo (MC) sampling scheme may suffer from lack of robustness in the estimates it produces if employed with small samples and very costly if used with very large samples required to estimate probabilistic estimates.

- Several techniques for reducing sample sizes have been defined:
  - Bred vectors, singular vectors, influence function MC …fertile area for data analysis!!!!!

Roadmap for improving Ensemble Forecasting

Possible contributions:

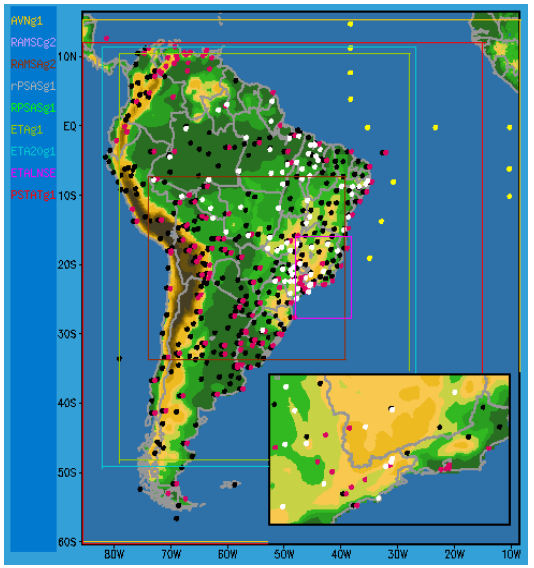
- **Add more members – improving estimate of uncertainties associated with model: weather forecasting (association of major weather forecasting centers – TIGGE/WMO), climate forecasting (IRI, Eurobrisa…) and IPCC (climate scenarios for global warming); ➔ (very high computational cost!!!)**

- **Application of Estimation Theory techniques for identifying optimal estimates (Bayesian techniques, ensemble Kalman filters and**
- **Genetic Programming, Neural Networks – deep learning techniques….);**

An example on how to sample model uncertainties:

Model Intercomparison – Super Model Ensemble

Participants:



- **Center for Weather Prediction and Climate Studies (CPTEC/INPE)**
- **Brazilian National Meteorological Institute - INMET**
- **Laboratory of Meteorology Applied to Regional Weather Systems (MASTER) – Univ. of São Paulo**
- **Laboratory of Mesoscale Forecasting (LPM/Fed Univ. of Rio de Janeiro)**
- **Center of Land-Ocean-Atmosphere (CATO/LNCC)**
- **Department of Meteorology of University of Maryland**
- **Brazilian Marine Meteorological Service (SMM/CHM)**
- **Center of Environmental Resources Information and Hydrometeorology (CIRAN/EPAGRI)**
- **Public available information from NCEP and other institutions**
- **New participants: UK Met.Office**

Colaborative work!!!

www.master.iag.usp.br/intercomp

Silva Dias and Moreira 2006)

Intercomparação de Modelos

O laboratório MASTER do IAG/USP, o Centro de Previsão de Tempo e Estudos Climáticos do INPE, o Laboratório de Prognósticos em Mesoescala do IG/UFRJ, o Centro de Modelagem do Sistema Atmosfera Terra Oceano do LNCC, o Department of Meteorology of University of Maryland, o Serviço Meteorológico Marinho (CHM) do Centro de Hidrografia da Marinha (CHM), o Centro de Informações de Recursos Ambientais e de Hidrometeorologia do EPAGRI, o Grupo de Estudos em Previsão Regional Atmosférica da FURG, o Centro de Investigaciones del Mar y la Atmosfera da UBA e o Grupo de Modelado en Mesoescala da UBA e o Grupo de Modelagem Atmosférica de Santa Maria da UFSM, participam de um esforço coordenado para avaliar as previsões numéricas de tempo disponíveis para o público em geral e desenvolver esquemas de previsão baseados na multiplicidade dos produtos de previsão, disponíveis em tempo real. Produtos públicos disponibilizados por instituições estrangeiras (como o NCEP/EUA) são também incluídos no processo de avaliação. Este trabalho está em sintonia com os objetivos do programa THORPEX/TIGGE da OMM. Este esforço é parcialmente financiado pela FINEP no projeto BRAMSNET

MODELOS REGIONAIS

☐	**RAMSC/MASTER** (Modelo BRAMS iniciado com o Global do CPTEC - Resolução de 25 km)
☐	**RAMSV/MASTER** (Modelo BRAMS iniciado com o ETA20 - Resolução de 2 km)
☐	**TEB/MASTER** (Modelo BRAMS acoplado ao TEB iniciado com o Global do AVN - Resolução de 4 km)
☐	**DR_ET_g1/MASTER** (Modelo BRAMS iniciado com o Global do CPTEC - Resolução de 14 km)
☐	**RAMSB_g2/MASTER** (Modelo BRAMS iniciado com o Global do AVN - Resolução de 10 km)
☐	**RAMSQ/MASTER** (Modelo BRAMS iniciado com o Global do AVN - Operação para o Rio Grande - Resolução de 32 km)
☐	**RAMSS/MASTER** (Modelo BRAMS iniciado com o Global do CPTEC - Resolução de 20 km) - Experimental
☐	**RAMSM/MASTER** (Modelo BRAMS iniciado com o Global do AVN (acoplado com o Dalt) - Resolução de 12 km)
☐	**GEPFA/GEPFA-FURG** (Modelo BRAMS - Resolução de 10 km)
☐	**GRUMA/GRUMA-UFSM** (Modelo BRAMS - Resolução de 20 km)
☐	**B_UBA_g1/BRAMS-UBA** (Modelo BRAMS - Resolução de 80 km)
☐	**B_UBA_g2/BRAMS-UBA** (Modelo BRAMS - Resolução de 20 km)
☐	**WRFAR_g1/CIMA-UBA** (Modelo WRF - Resolução de 60 km)
☐	**WRFAR_g2/CIMA-UBA** (Modelo WRF - Resolução de 20 km)
☐	**CATT_g2/CPTEC** (Modelo CATT-BRAMS iniciado com o Global do CPTEC - Resolução de 30 km)
☐	**CATT_g3/CPTEC** (Modelo CATT-BRAMS iniciado com o Global do CPTEC - Resolução de 15 km)
☐	**RPSAS/CPTEC** (Modelo ETA com assimilação de dados observados - Resolução de 40 km)
☐	**rPSAS/CPTEC** (Modelo ETA com assimilação de dados observados - Resolução de 40 km) - Experimental
☐	**ETA/CPTEC** (Modelo ETA - Resolução de 40 km)
☐	**ETA20/CPTEC** (Modelo ETA - Resolução de 20 km)
☐	**ETA15_SE/CATO-LNCC** (Modelo ETA - Grade para a região Sudeste - Resolução de 15 km)
☐	**ETA15_RJ/CATO-LNCC** (Modelo ETA - Grade para o Rio de Janeiro - Resolução de 10 km)
☐	**ETA80km/UMD** (Modelo ETA - Resolução de 80 km - provou somente a precipitação)
☐	**ETA22km/UMD** (Modelo ETA - Resolução de 22 km - provou somente a precipitação)
☐	**HRM/CHM** (Modelo HRM - iniciado com o Modelo Alemão (GME)) - Resolução de 30 km)
☐	**HRM_zoo/CHM** (Modelo HRM - iniciado com o Modelo Alemão (GME)) - Resolução de 13 km)

MODELOS GLOBAIS

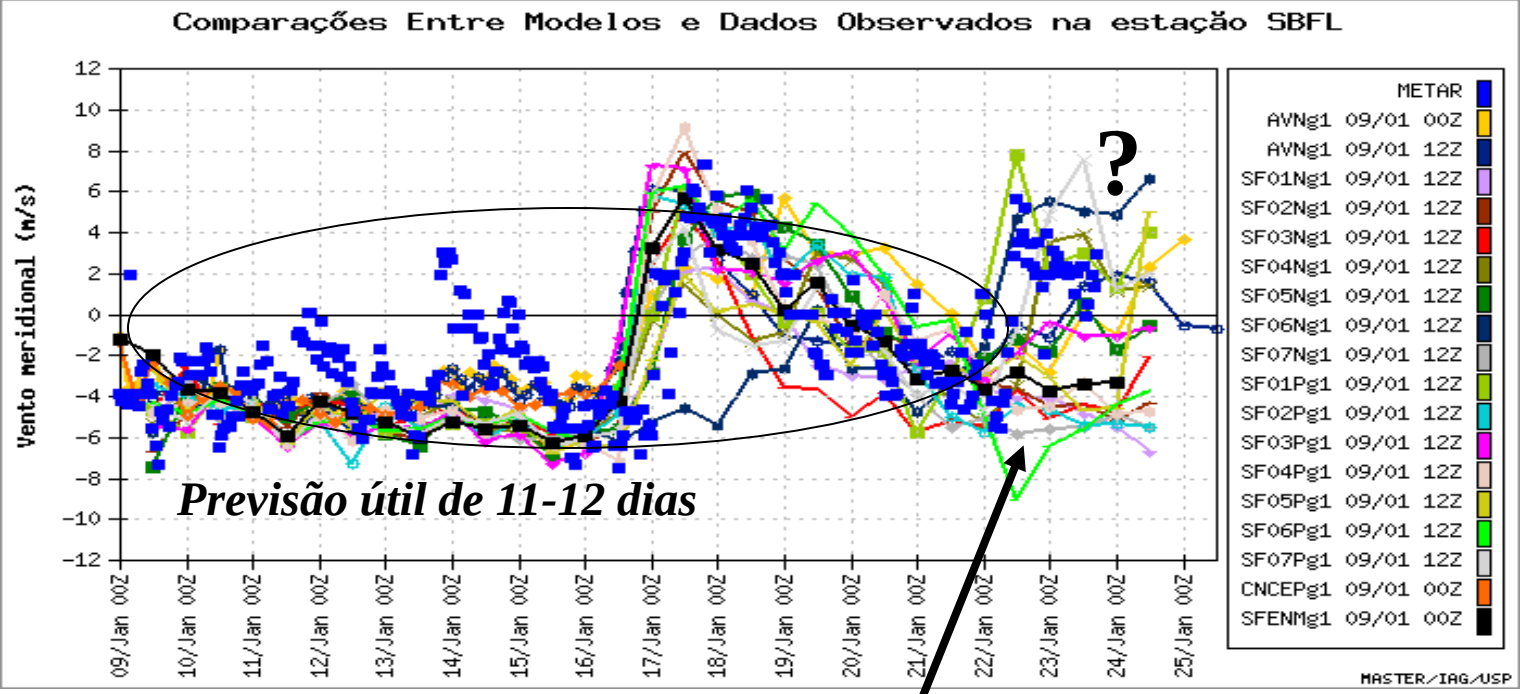
☐	**MRF/NCEP** (Modelo MRF - Resolução de 2.5°)
☐	**AVN/NCEP** (Modelo GLOBAL do NCEP - Resolução de 1°)
☐	**Conjunto Médio/NCEP** (Média dos membros do conjunto - Resolução de 1°)
☐	**T126/CPTEC** (Modelo GLOBAL do CPTEC - Resolução de 100 km)
☐	**Acoplado/CPTEC** (Modelo T126 acoplado com modelo oceânico - Resolução de 100 km) - Experimental
☐	**"Conjunto" Acoplado/CPTEC** (Conjunto obtido via 10 previsões consecutivas do modelo Acoplado/CPTEC) - Experimental
☐	**T213/CPTEC** (Modelo GLOBAL do CPTEC - Resolução de 63 km)
☐	**GPSAS/CPTEC** (Modelo com análise do CPTEC - Resolução de 100 km)
☐	**Conjunto Médio/CPTEC** (Média dos membros do conjunto - Resolução de 100 km)

PREVISÃO ESTATÍSTICA

☐	**MSMES** - *MASTER Super Model Ensemble System* (Previsão média de todos os modelos disponíveis, ponderada pelo erro médio quadrático com remoção prévia do viés) - clique **aqui** para mais informações

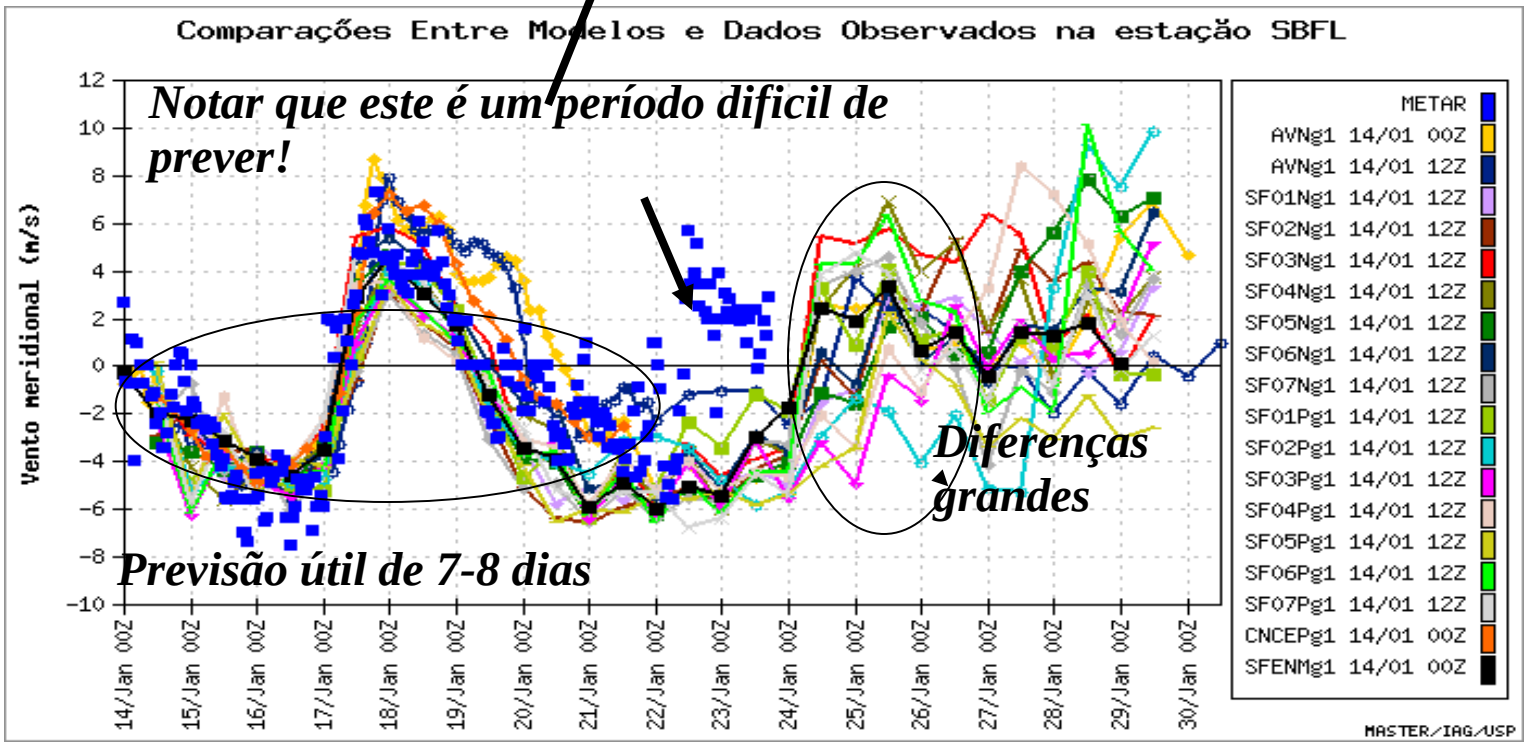
Submeter

Configurações pré-definidas:



Exemplo da previsão da componente meridional do vento - membros do CPTEC e NCEP

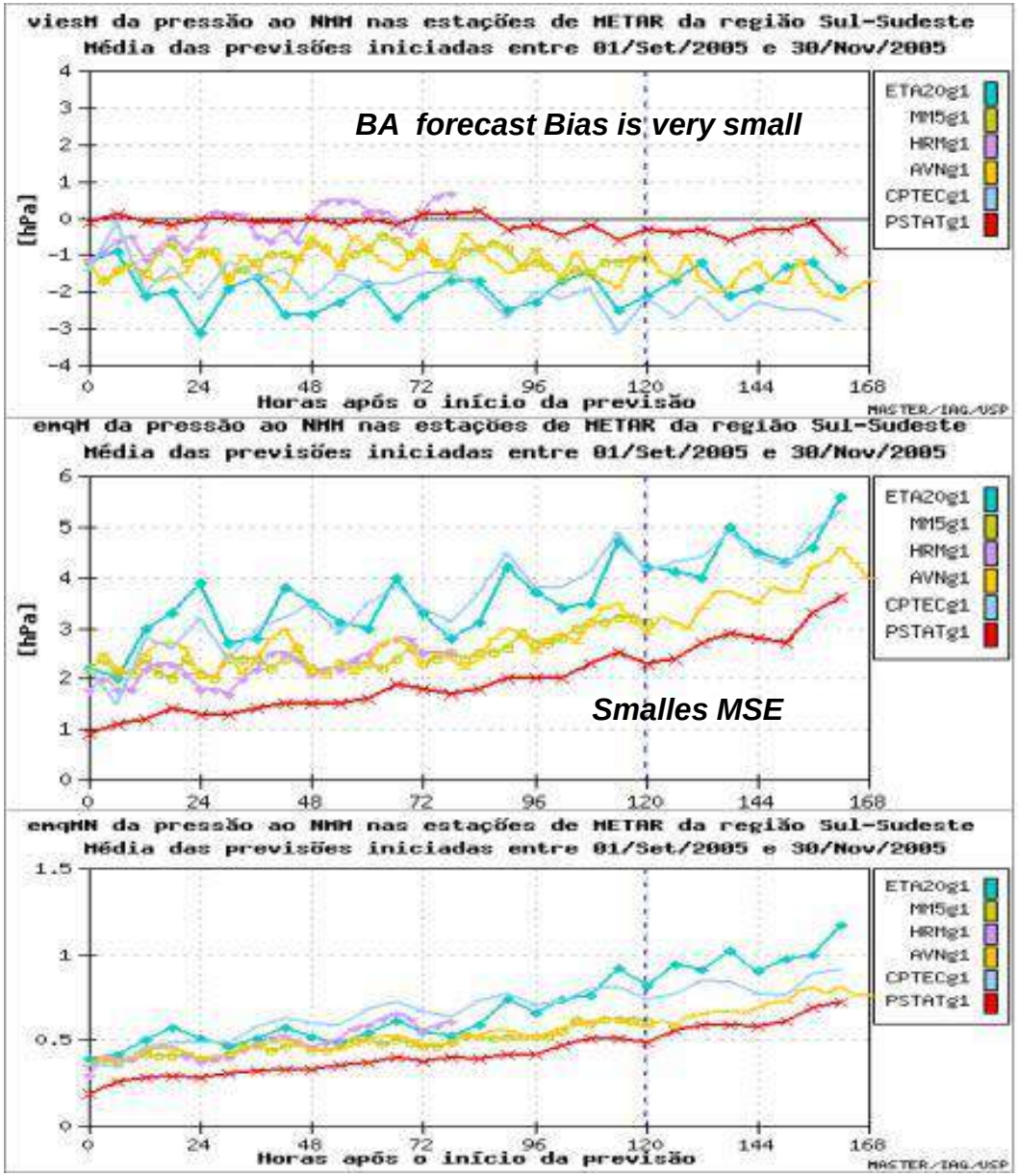
Quadrados azuis: observações no aeroporto SBFL

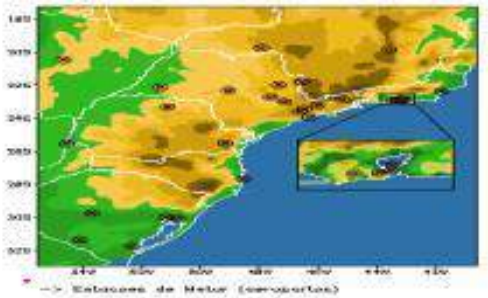


Silva Dias et al. 2006)

MSL Pressure Forecasts

Red - Bayesian Average Approach





Bias – up to 168h forecast

Mean square error - EMQM

Normalized mean square error - EMQMN

Técnicas Estatísticas Tradicionais

Seja a previsão por conjunto E:

$$\boxed{E = \sum_{i=1}^{m} w_i (M_i - B_i) = \sum_{i=1}^{m} w_i M_i'}$$

- Média simples:

$$w_i = \frac{1}{m} \quad \forall i$$

- *MASTER Super Model Ensemble System* – MSMES:

$$w_i = \frac{\frac{1}{\mathrm{MSE}_i}}{\sum_{i=1}^{m} \frac{1}{\mathrm{MSE}_i}} \quad \text{com} \sum_{i=1}^{m} w_i = 1,$$

onde o MSE (*Mean Squared Error*) do membro i é dado por

$$\mathrm{MSE}_i = \frac{1}{t} \sum_{j=1}^{t} \left(M_{ij}' - O_j \right)^2$$

Silva Dias, P. L., Moreira, D. S., e Dolif, G. (2006). The master super model ensemble system (MSMES). pgs. 1751–1757. 8 ICSHMO

Técnicas Estatísticas Tradicionais

- *Bayesian Model Averaging* – BMA:

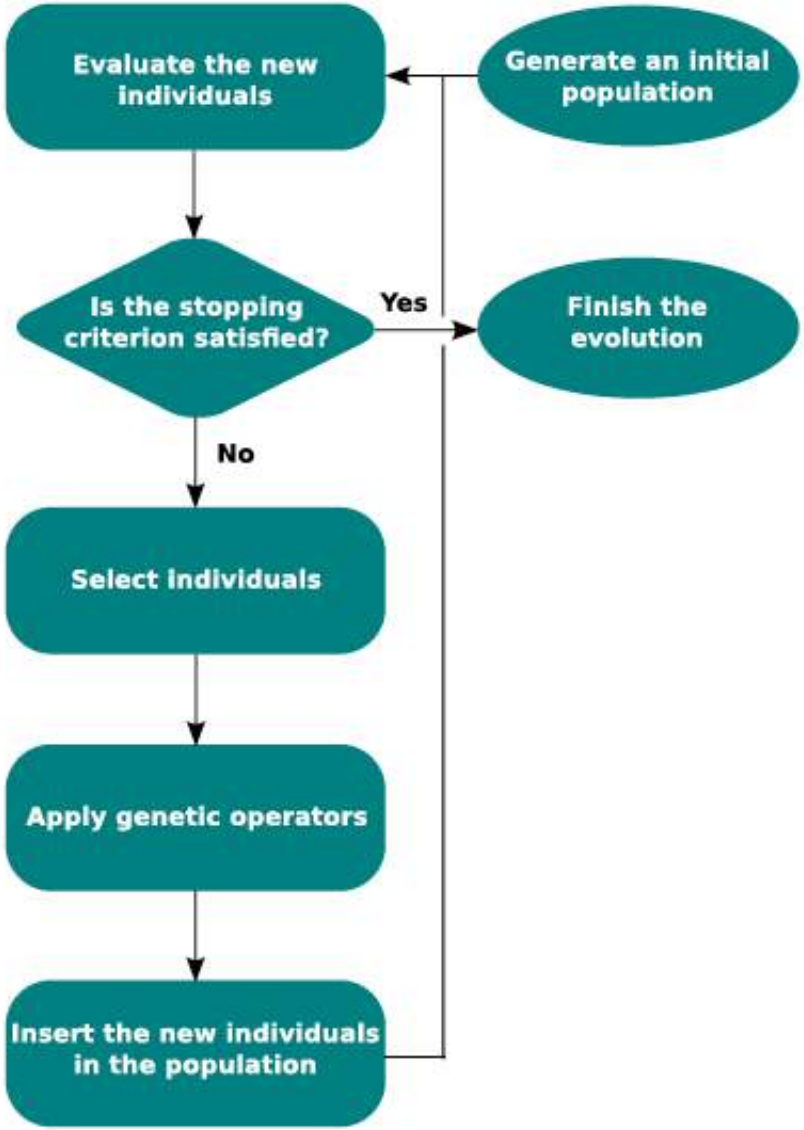
$$L(w_1, ..., w_m, ...) = \sum_{j=1}^{t} \sum_{i=1}^{m} w_i f_i(O_j | M'_{ij}) \quad \text{com} \sum_{i=1}^{m} w_i = 1,$$

onde w_i e alguns parâmetros de f_i são obtidos através da maximização da função de verossimilhança L através do algoritmo *Expectation-Maximization*.

$$w_i = \frac{1}{t} \sum_{j=1}^{t} z_{ij}, \quad z_{ij} = \frac{w_i f_i(O_j | M'_{ij})}{\sum_{k=1}^{m} w_k f_k(O_j | M'_{kj})}$$

Genetic Programming

It **starts** with the **generation** of an **initial population** of **candidate solutions**. **Each solution** is **evaluated** according to a **fitness function**. Based on this fitness, some **solutions** are **stochastically selected** from the **population**. The algorithm follows with the **application** of the **genetic operators over** the **selected solutions**. Two of the most important genetic operators are **crossover** and **mutation**. The **new solutions** are **introduced** into the **population**. The **evolutionary process** of **evaluation**, **selection**, **genetic operators**, and **replacement** is **iterated** until a **stopping criterion** is **satisfied**. The **final solution** is the **best solution** of the **last iteration**.



Computers & Geosciences 106 (2017) 139–149

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Application of evolutionary computation on ensemble forecast of quantitative precipitation

Amanda S. Dufek[a,*], Douglas A. Augusto[b], Pedro L.S. Dias[c], Helio J.C. Barbosa[a,d]

[a] National Laboratory for Scientific Computing, 333 Getúlio Vargas Avenue, Quitandinha, Petrópolis, RJ 25651-075, Brazil
[b] Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil
[c] Institute of Astronomy, Geophysics and Atmospheric Sciences, University of São Paulo, SP, Brazil
[d] Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil

ARTICLE INFO

Keywords:
Ensemble weather forecast
Quantitative precipitation
Evolutionary computation
Genetic programming

ABSTRACT

An evolutionary computation algorithm known as genetic programming (GP) has been explored as an alternative tool for improving the ensemble forecast of 24-h accumulated precipitation. Three GP versions and six ensembles' languages were applied to several real-world datasets over southern, southeastern and central Brazil during the rainy period from October to February of 2008–2013. According to the results, the GP algorithms performed better than two traditional statistical techniques, with errors 27–57% lower than simple ensemble mean and the MASTER super model ensemble system. In addition, the results revealed that GP algorithms outperformed the best individual forecasts, reaching an improvement of 34–42%. On the other hand, the GP algorithms had a similar performance with respect to each other and to the Bayesian model averaging, but the former are far more versatile techniques. Although the results for the six ensembles' languages are almost indistinguishable, our most complex linear language turned out to be the best overall proposal. Moreover, some meteorological attributes, including the weather patterns over Brazil, seem to play an important role in the prediction of daily rainfall amount.

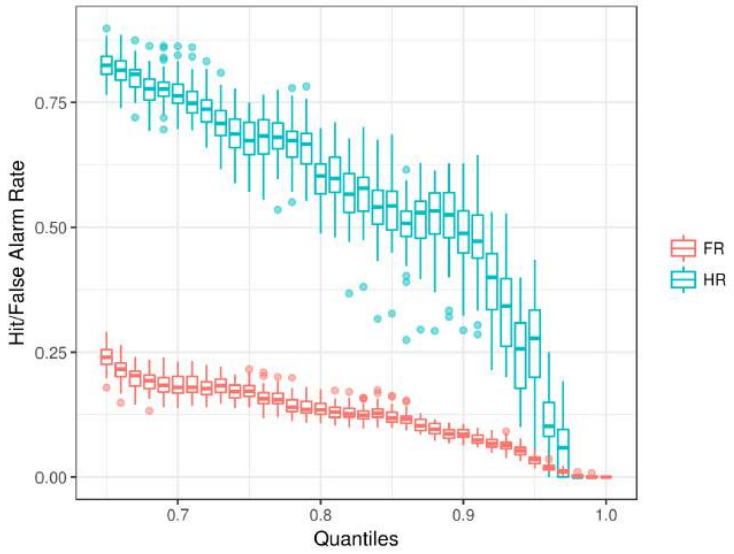


Figure 6.7: Hit Rate (HR) and False Alarm Rate (FR) of the DNN 24-hour precipitation forecast in the city of São Paulo. In this case, the DNN was trained using GFS explanatory data.

DNN – Deep Neural Network

Perez, G.M.P. – Improving the Quantitative Precipitation Forecast: a Deep Learning Approach – PhD Thesis – IAG/USP 2018.

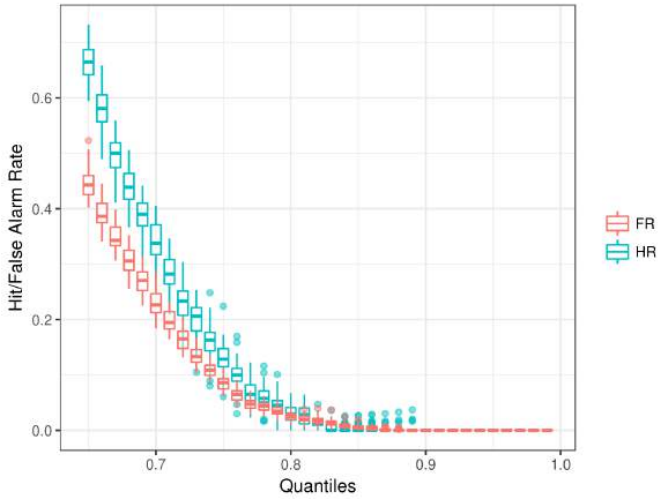
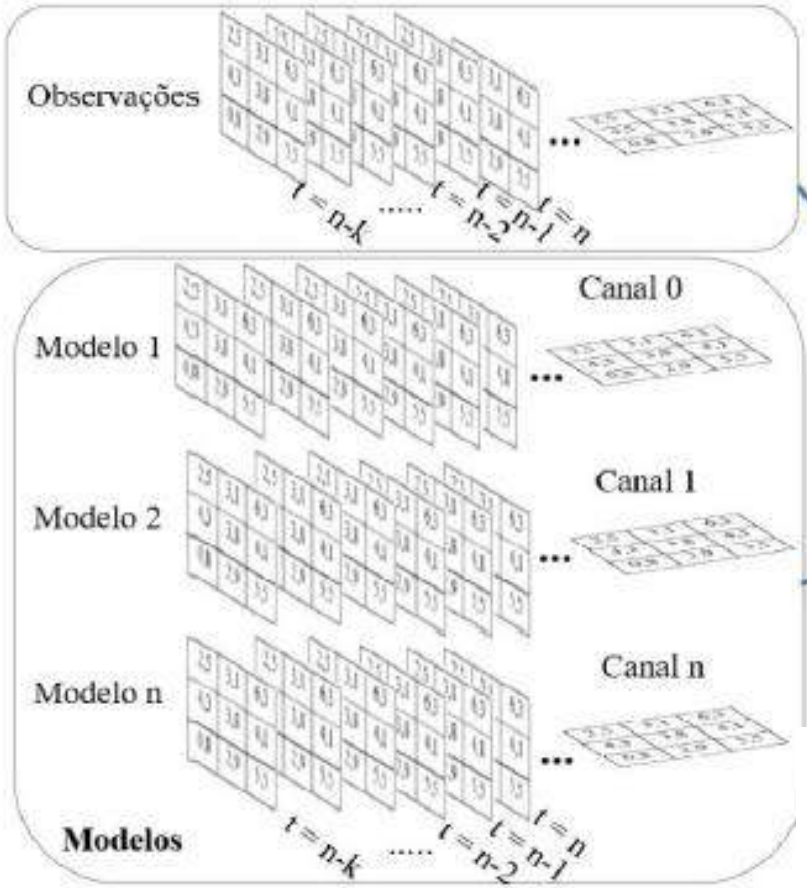
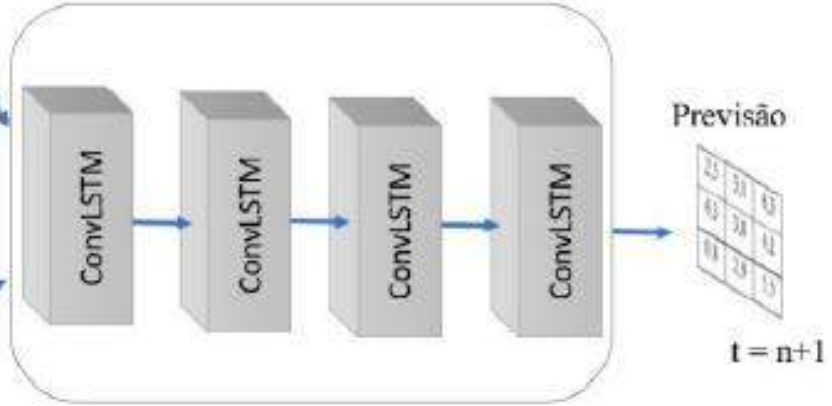
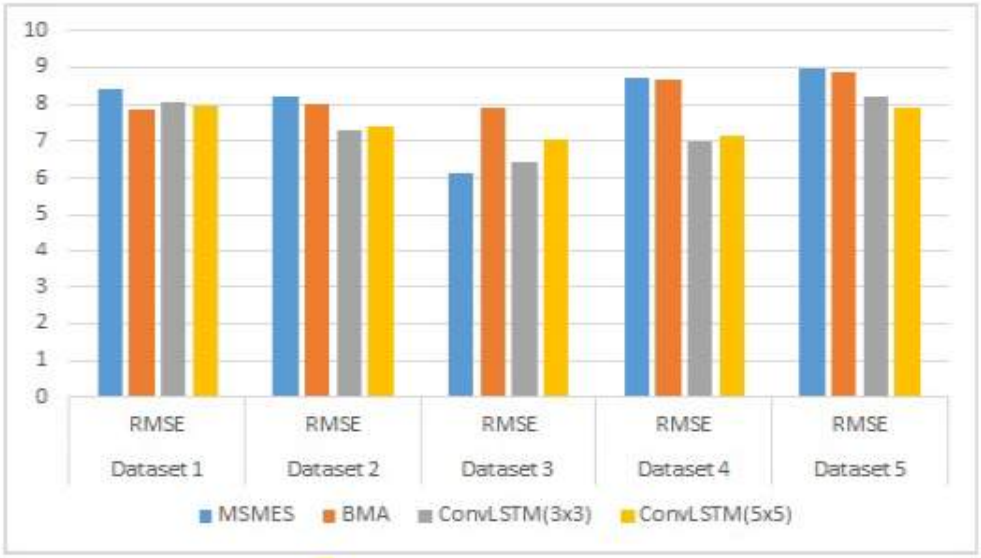


Figure 6.8: Hit Rate (HR) and False Alarm Rate (FR) of the GFS 24-hour precipitation forecast in the city of São Paulo.

ConvLSTM. Application to ensemble prediction – Precipitation

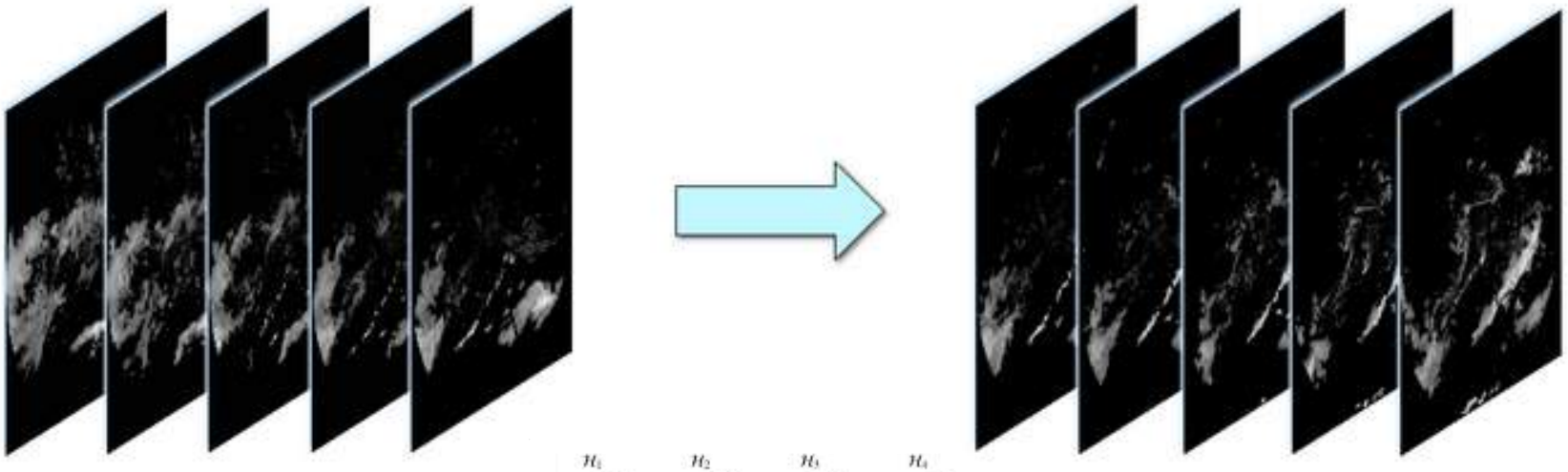


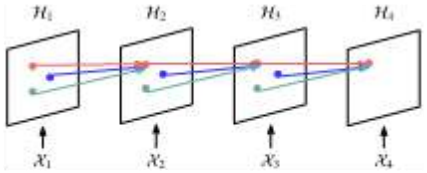




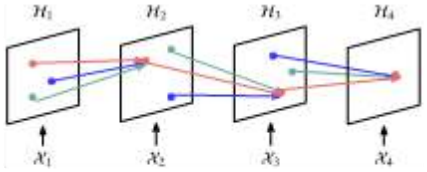
Souto, Y. Souto – LNCC 2018

Spatiotemporal Machine Learning Models for Precipitation Nowcasting





(a) For convolutional RNN, the recurrent connections are fixed over time.



(b) For trajectory RNN, the recurrent connections are dynamically determined.

Summary -1

- **Data Assimilation techniques in the early 2000's had a remarkable impact!!! More than model improvements.**

- **3DVar < 4DVar - Ensemble Kalman Filter competitive and numerically very efficient**

- **Insufficient ensemble model spread indicates the need to account for the statistical aspects of model error.**

- **Need to increase model independence (model diversity…)**

- **Super model ensemble based on simple Bayesian procedure : quite successful. Can be improved with AI – e.g. Genetic Programming Approach or Deep Neural Networks**

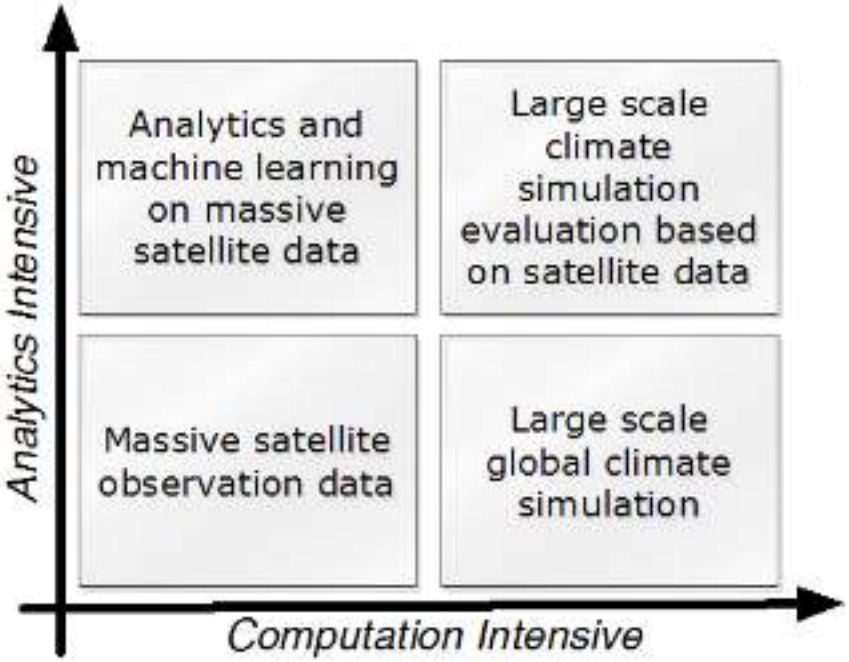


Fig. 1. Categorization of HPC and Big Data related challenges in cloud radiation and energy.

Training Challenges

- **Lower cyberinfrastructure adoption on advanced data and computing techniques in the current atmospheric sciences curriculum;**
- **Lack of training research challenges in applicable domains for graduate students in Computing and Applied Mathematics in Atmospheric Science topics;**
- **Lack of customized training on the use of data science tools for students in Atmospheric Sciences.**
- **Lack of team-based multidisciplinary training and frontier research projects.**

Graduate program	Existing courses can be leveraged	Other main courses offered	Additionally required knowledge
Information systems	• Programming • Data Mining and Machine Learning • Distributed Systems • Introduction to Data Science	• Databases • Artificial Intelligence • Decision Making • System Analysis and Design	• Computational Physics • Parallel Computing • Partial Differential Equations • Big Data Techniques and Systems
Applied Mathematics	• Partial Differential Equations • Computational Mathematics and Programming • Introduction to Parallel Computing	• Ordinary Differential Equations • Optimization Techniques • Combinatorics and Graph Theory • Linear Algebra	• Computational Physics • Data Mining and Machine Learning • Big Data Techniques and Systems
Atmospheric Physics	• Computational Physics	• Atmospheric Physics • Atmospheric Dynamics • Atmospheric Radiative Transfer • Atmospheric Remote Sensing • Atmospheric Modelling	• Parallel Computing • Partial Differential Equations • Data Mining and Machine Learning • Big Data Techniques and Systems

Summary -2

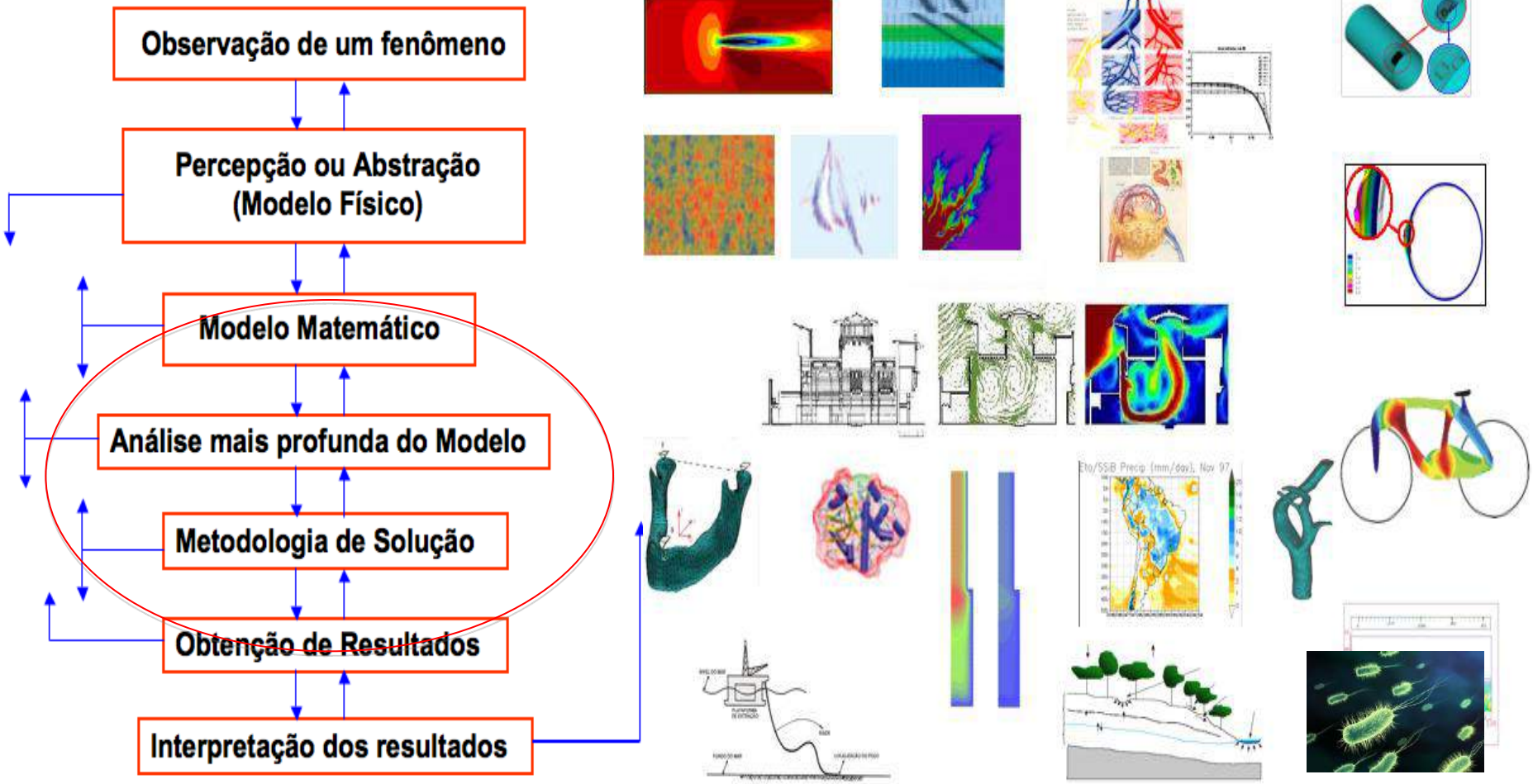
- **Training in Data Science for Atmospheric Sciences students:**
  - **How to prepare the next generation scientists?**
    - **Data Science**
    - **High Performance Computing as an indispensable tool**
    - **Complex network analysis and climate science**
    - **Team work in challenging problems**

Based on Wang et al. https://pdfs.semanticscholar.org/099b/6254a2da68fddd2af2fad97e5d9d0b022d93.pdf

“The whole problem with the world is that fools and fanatics are always so certain of themselves, and wiser people so full of doubts”
– Bertrand Russell
Thanks
Pedro Leite da Silva Dias –
pedro.dias@iag.usp.br

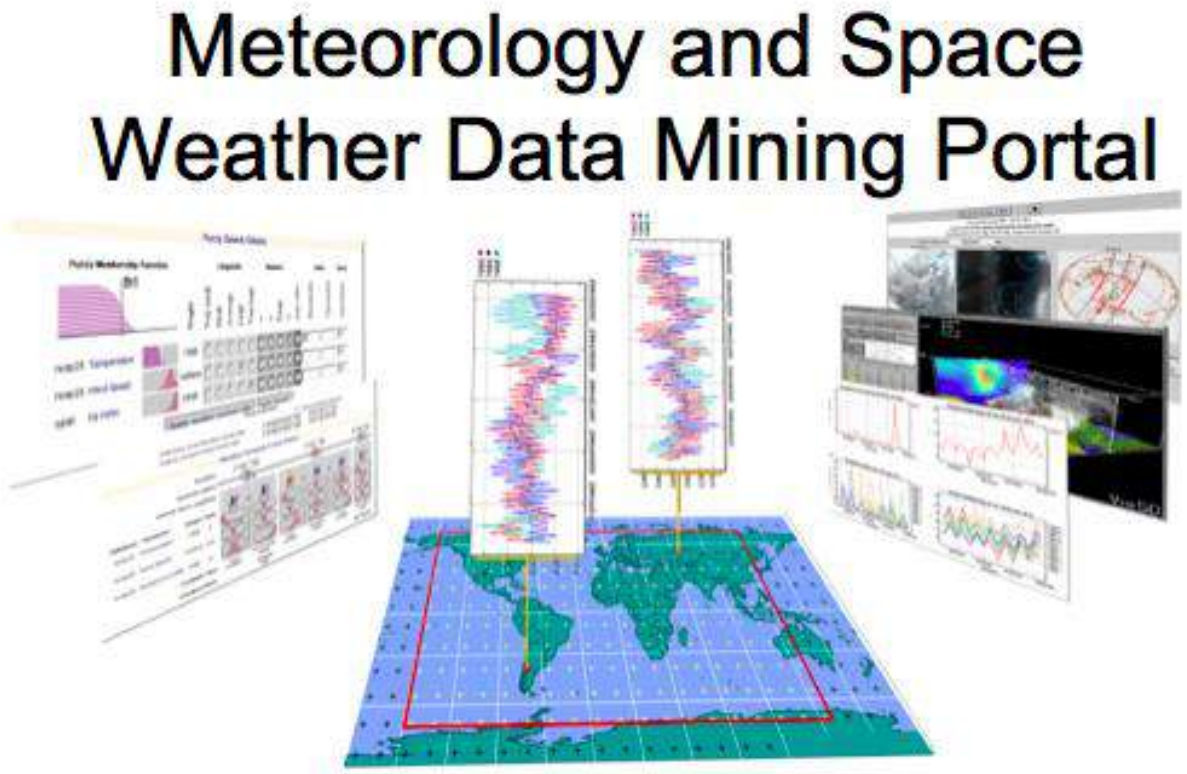
Modelagem Matemática

Construção de modelos matemáticos e técnicas de soluções numéricas utilizando computadores para analisar e resolver problemas científicos e de engenharia.



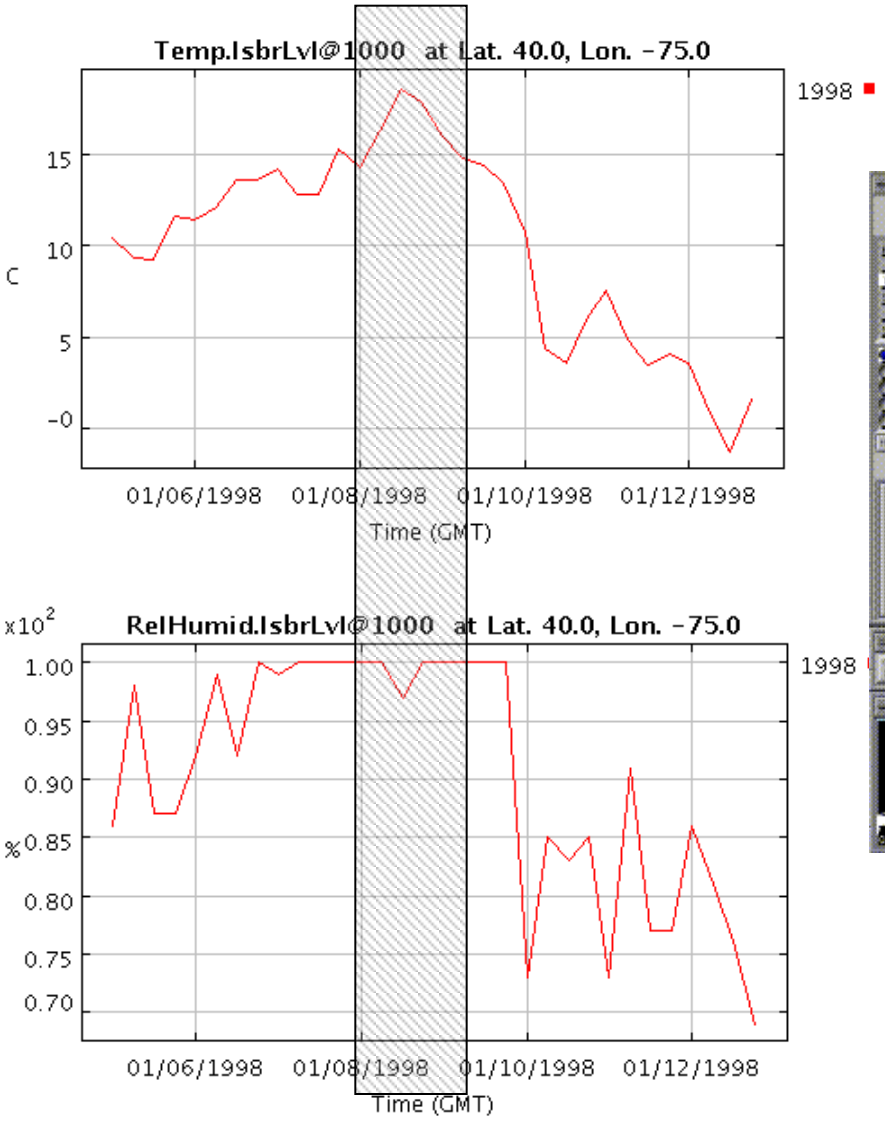
Challenge :

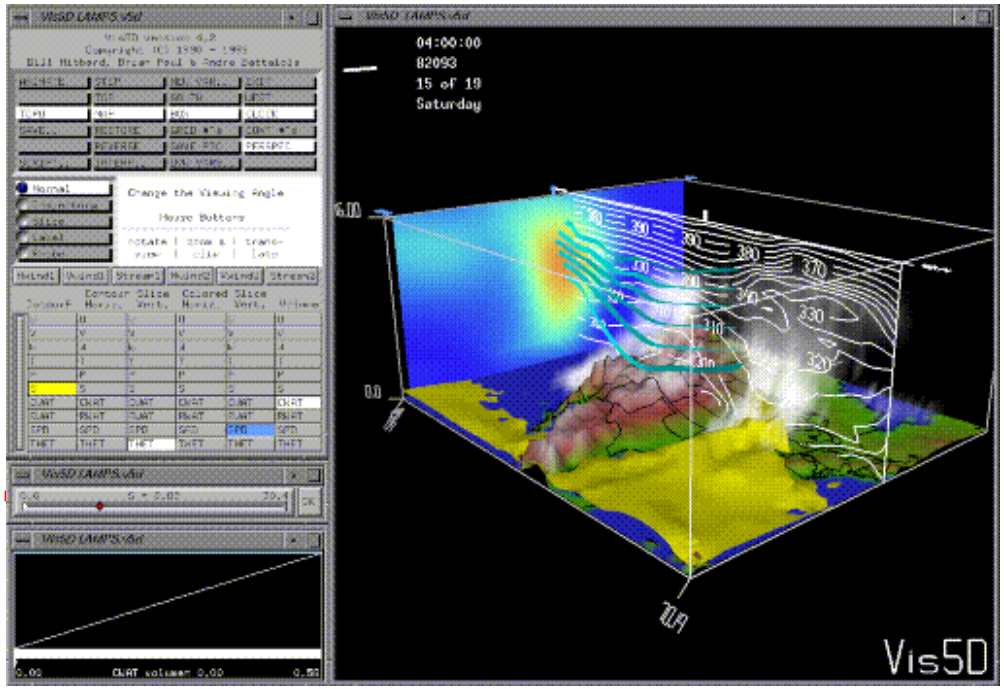
- ***Data Mining of massive data sets (3D animation, multiple fields)***



- ***Challenge : Visualization of massive data sets (3D animation, multiple fields)***

Viewing the event in time and space





Vis5D time-space-parameter animation