Data Mining and Knowledge Discovery of Land Cover and Terrestrial Ecosystem Processes from Global Remote Sensing Data

Mark Friedl
Department of Geography & Environment, Boston University
friedl@bu.edu

Carla Brodley
Department of Computer Science, Tufts University
brodley@cs.tufts.edu

Surajit Ray
Department of Mathematics and Statistics, Boston University
sray@math.bu.edu

Support from NASA (NASA TEP, LCLUC, MODIS, and IDU)

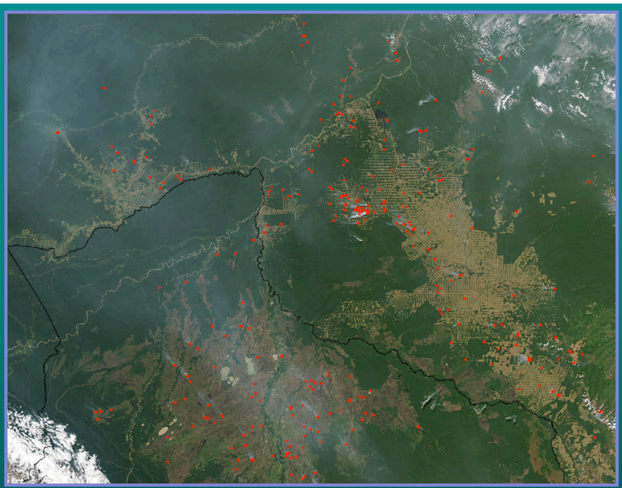
Context: Global Change Studies

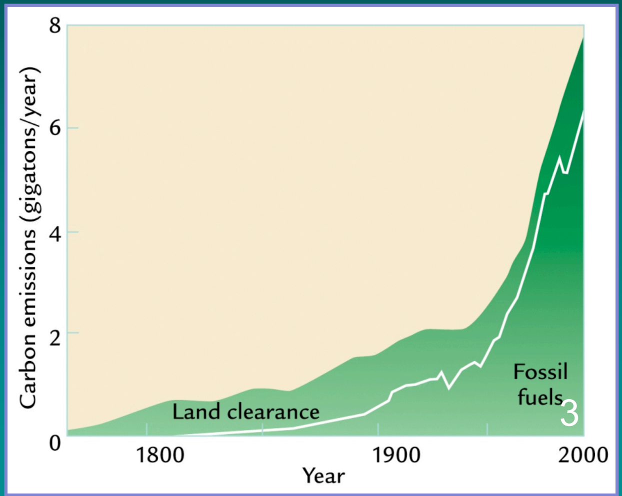
- *Global Observation Systems*
  - EOS/GEOSS, remote sensing, in-situ, sensor webs….
  - Large, heterogeneous, complex data sets
    - High dimensional: multi-spectral, multi-temporal, multi-resolution…
    - Significant analysis problems: noise, missing data…
- *Dynamics in Earth System*
  - Characterized by high complexity, variability at multiple scales
    - Climate change vs variability
    - Ecosystem response (species composition, phenology, & population dynamics)
    - Human activities

Why Does this Matter?

Monitoring and quantification of human impacts

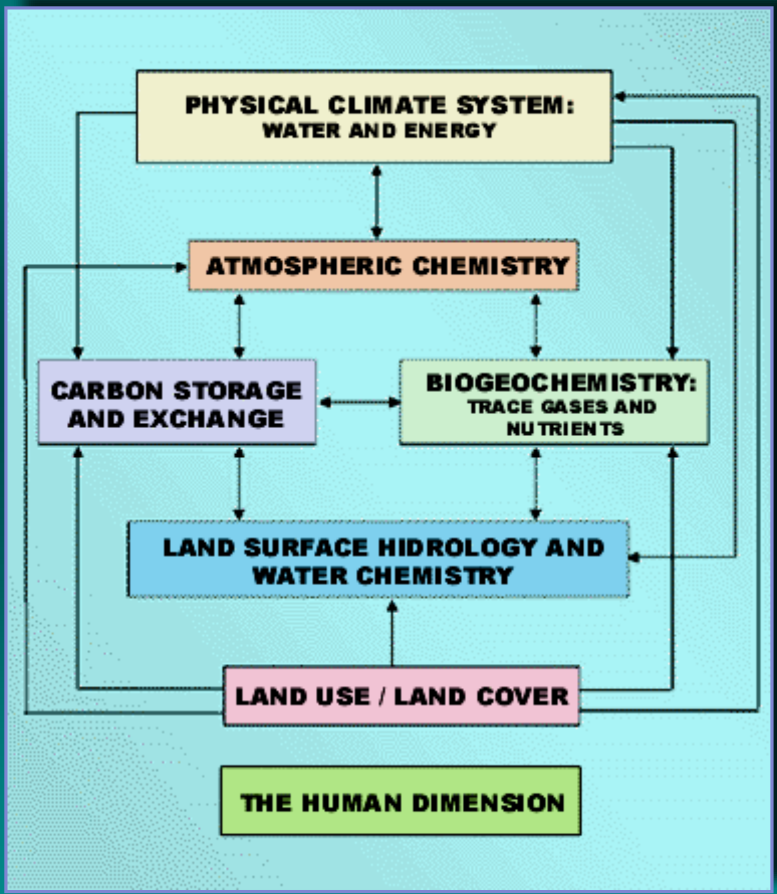
- Land conversion and land use by humans represent the largest single mechanism of environmental change
- Carbon storage/release
- Biodiversity
- Ecosystem Services
  - Land resources & food security
  - Hydrology and water resources
- Etc……





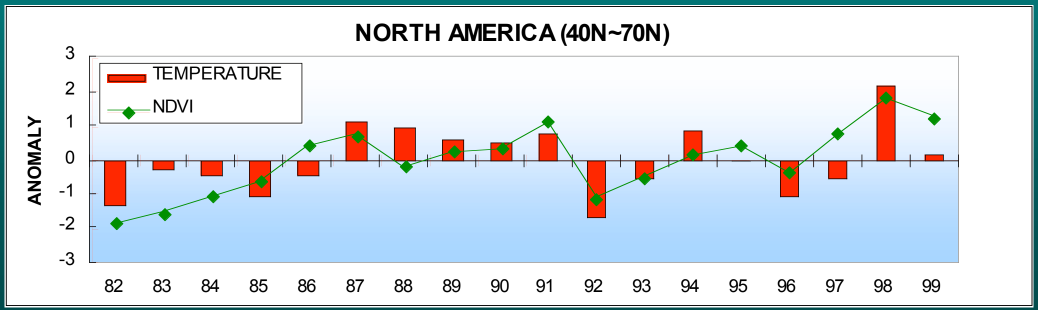
Global Land System

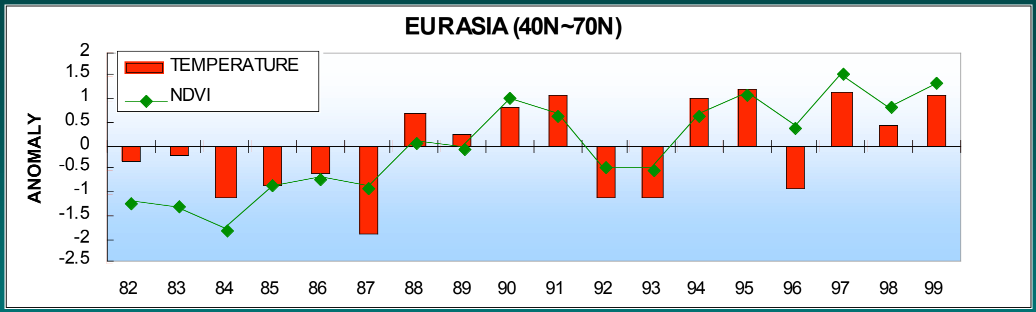
- Modeling Perspective
  - Global ecosystems and land surface provides key boundary condition to global weather and climate system



(credit: NASA LBA)

Ecosystem Response to Climate Change





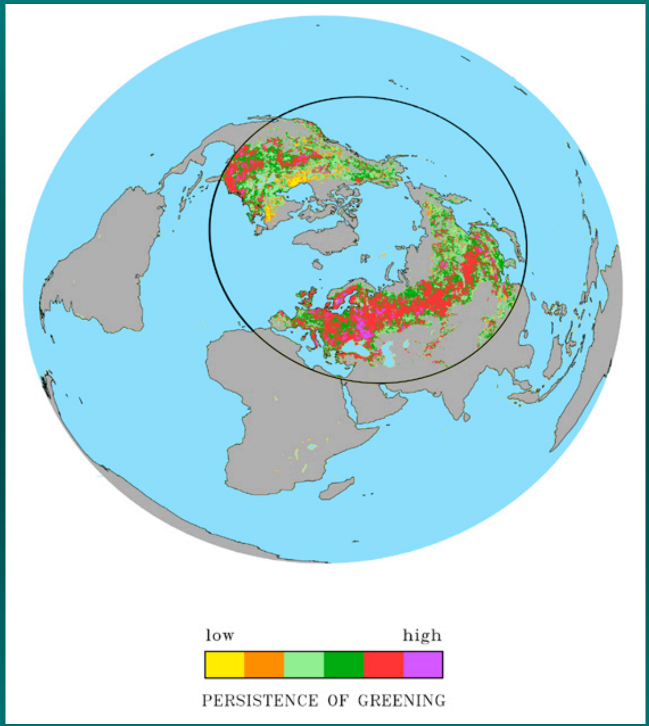


Image Credits: Ranga Myneni

Growing evidence that ecosystems are responding to changing climate at a variety of space-time scales
(Myneni et al, *Nature*, 1997; Nemani et al, *Science*, 2003; others…)

IDU Challenge

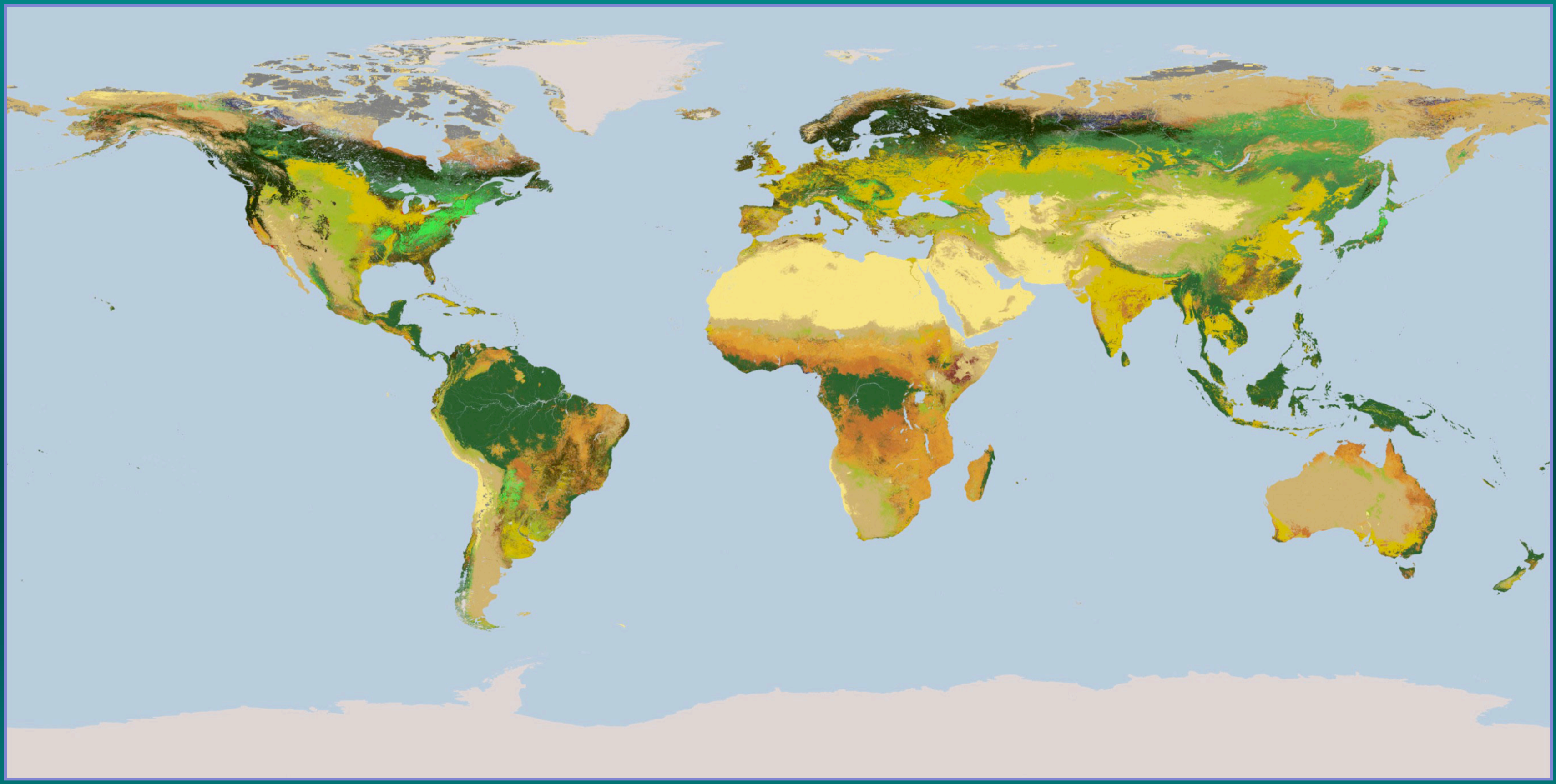
...Better tools for processing, analyzing, detecting change, and understanding patterns and process from large scale Earth science data sets……

- Machine learning, data mining,statistical tools are not the answer, but they can be part of the solution.
- *Danger*: Fishing expeditions
- *Require*: Earth scientists to better understand tools, data modelers to better understand problem domains.

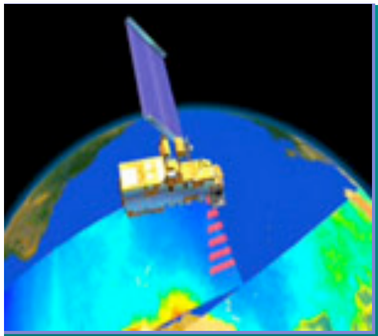
Overview of Talk

- Three Problem Domains/Applications
  - Supervised classification of global land cover
    - *Map global land cover from remote sensing*
  - Unsupervised deomposition of space-time variance in remote sensing time series
    - *Search, mine, discover patterns* related to data artifacts and patterns in coupled climate-ecosystem dynamics
  - Use of functional models, clustering & mixture models
    - *Reduce dimensionality & understand class structure in data*.

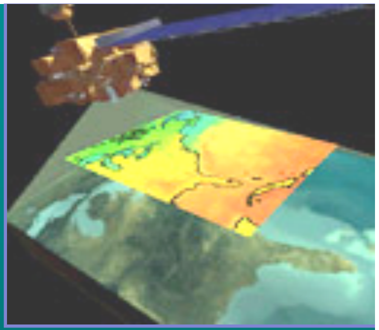
Supervised MODIS Land Cover Classification



0 Water	6 Closed Shrublands	12 Croplands
1 Evergreen Needleleaf Forest	7 Open Shrublands	13 Urban and Built-Up
2 Evergreen Broadleaf Forest	8 Woody Savannas	14 Cropland/Natural Veg. Mosaic
3 Deciduous Needleleaf Forest	9 Savannas	15 Snow and Ice
4 Deciduous Broadleaf Forest	10 Grasslands	16 Barren or Sparsely Vegetated
5 Mixed Forests	11 Permanent Wetlands	17 Tundra

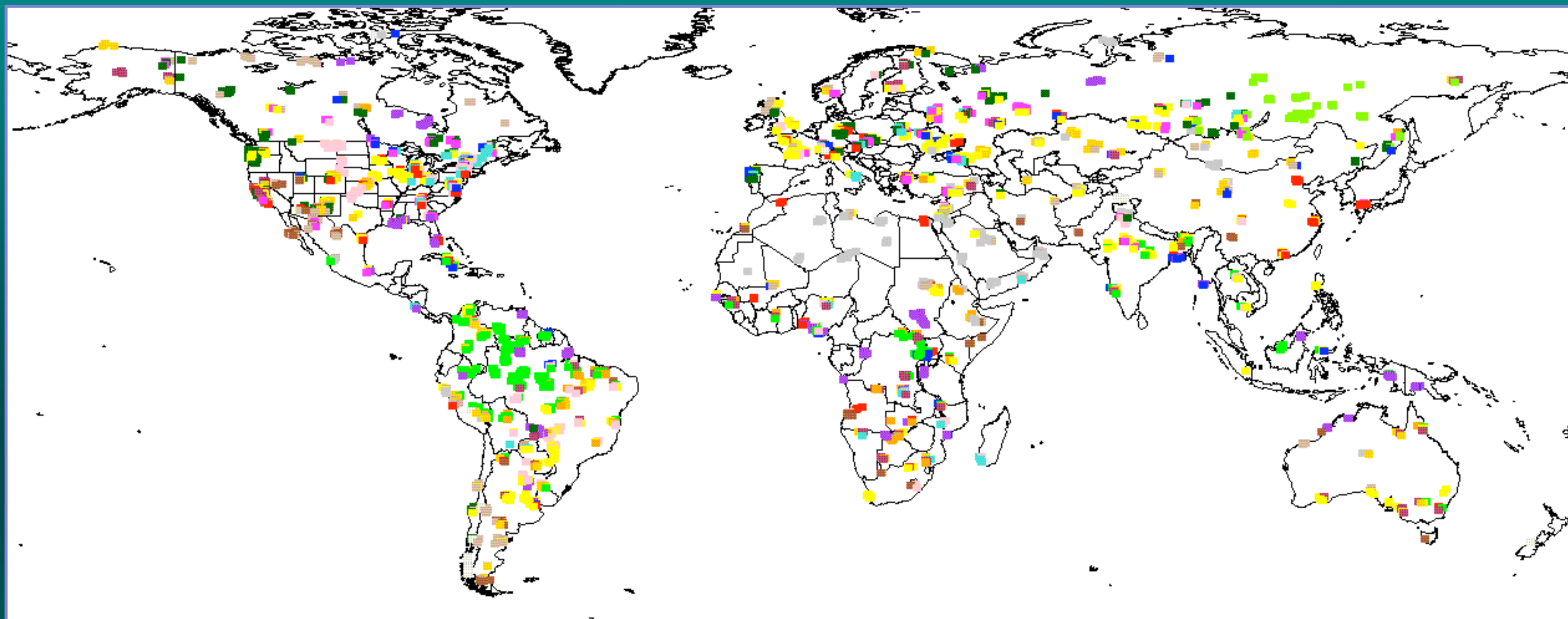


MODIS

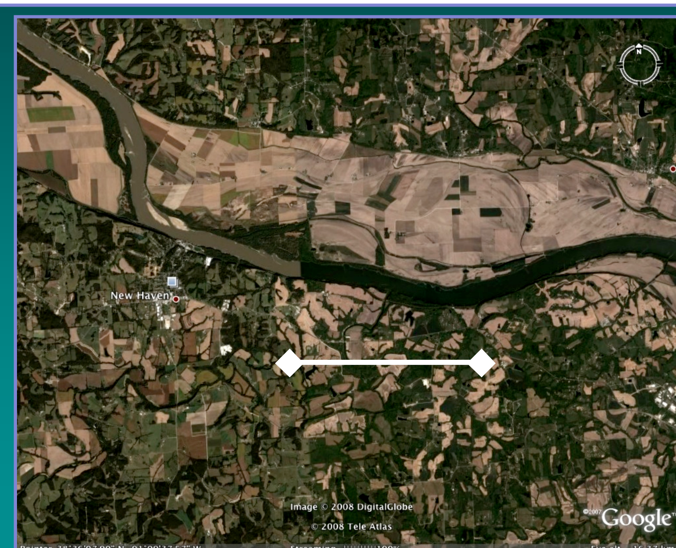


- Moderate Resolution Imaging Spectroradiometer
- Onboard EOS Terra (10:30 AM descending); and EOS-Aqua (1:30 PM ascending) local solar equatorial crossing
- Sun synchronous, near polar orbit; 705.3 km
  - 36 spectral bands, VNIR, SWIR, TIR (0.4–14 µm)
  - Spatial resolution 500-m; scan angle: +/-55^{o}; 2330 km swath
  - 2-day global repeat, 1-day or less poleward of 30^{o}
  - Onboard calibration; Band-to-band registration, etc.
- *Ingest: global, 500-m, 9-bands, 8-day intervals for one year*
  - *~2.8 x10^{11} input elements to produce a map with ~175x10^{6} cells*

Supervised: Training Data



~2000 Sites derived from hi-res imagery & spanning all major regions & ecosystems, but sampling based on “opportunistic” criteria



Technical Challenges

- Algorithms cannot compensate for inadequate features
  - *Use of spatially varying priors*
- Unbalanced, misrepresentative representative training data
  - *Bias correction via global priors*
- *Each of these "corrections" reduce accuracy of predictions relative to training data, but improve quality of final maps!*

- (Year-to-year classification variability vs real change?
  - *Heuristic for updating labels based on estimated posterior probs)*

MODIS Land Cover Processing Chain

Extract Exemplars From Site Database

Estimate DTs from Training Data (AdaBoost, c4.5)

Apply Boosted DT's to Global Data

Class-conditional probability for each class at each pixel
(Friedman et al., 2000)

Adjust Sample Bias, Apply Priors via Bayes' Rule

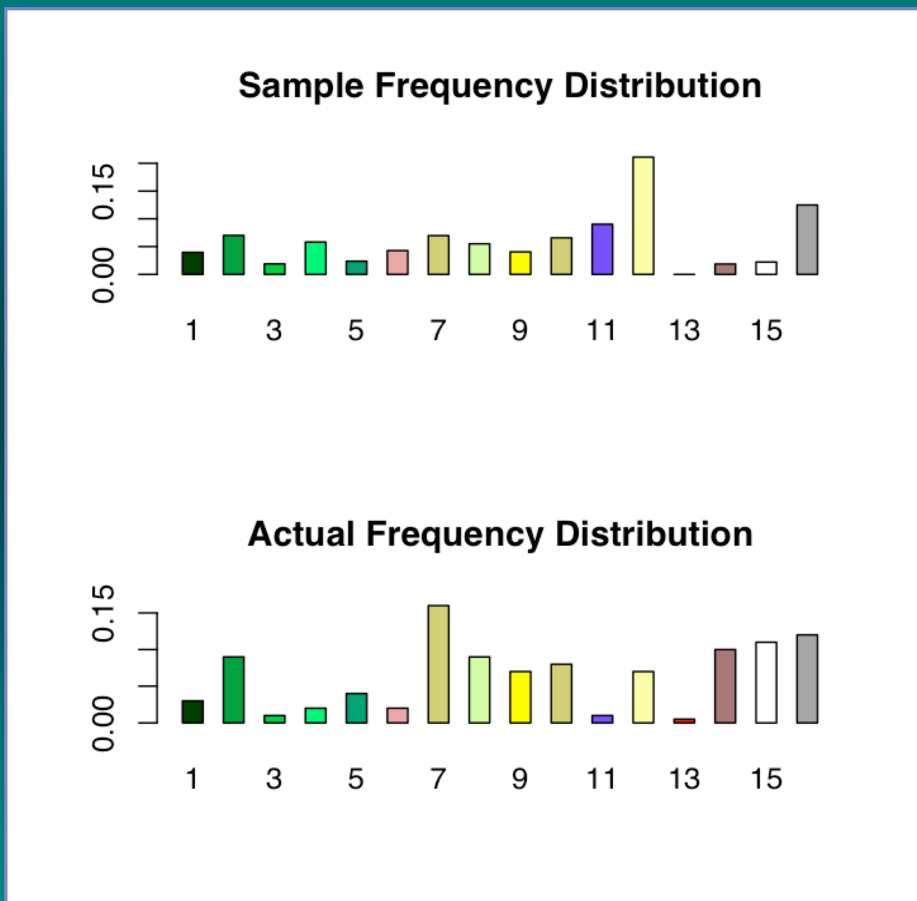
Posterior probability for each class at each pixel

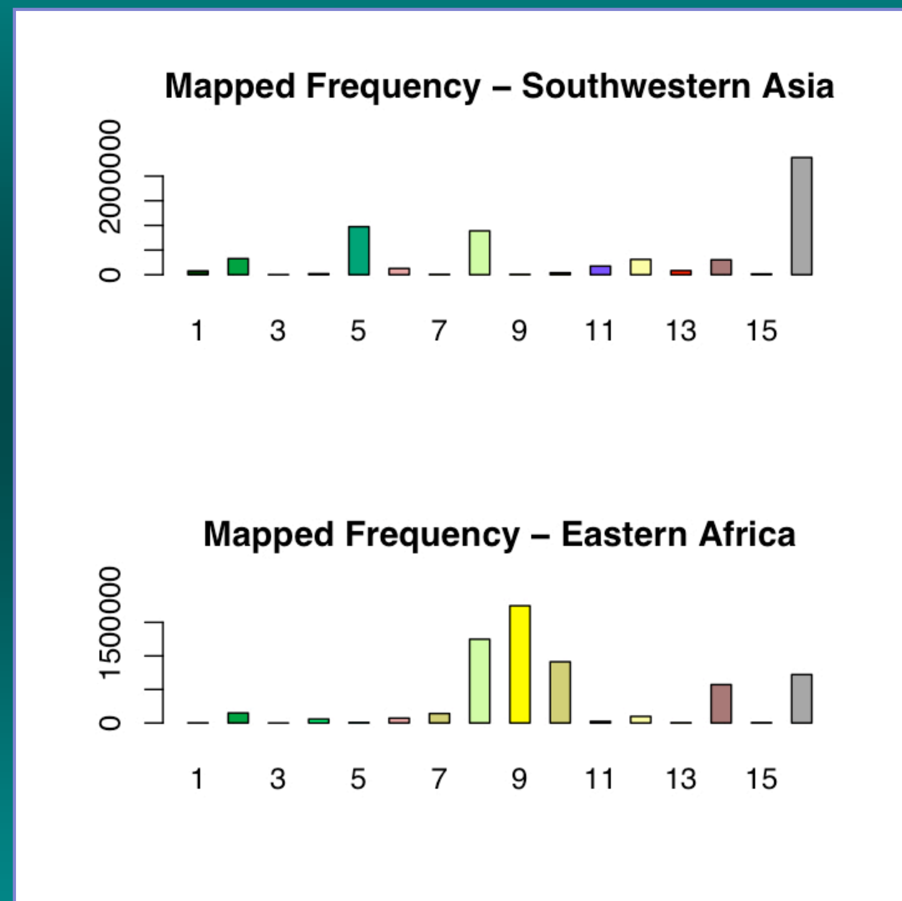
Maps

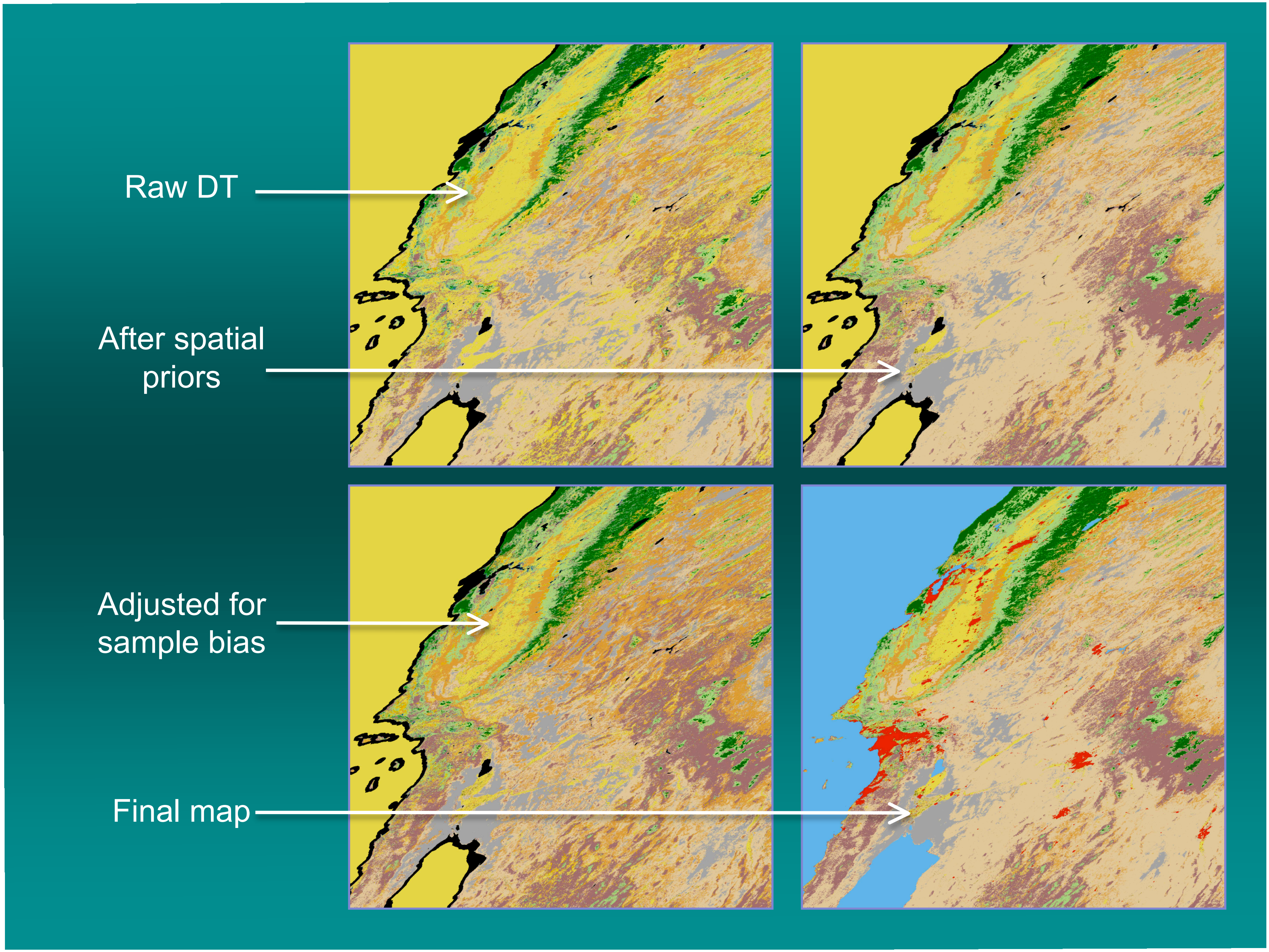
1st & 2nd most likely class at each pixel

Friedl and Brodley, 1997; Friedl et al, 1999, 2002, 2008; McIver and Friedl, 2001, 2002

Sample Bias and Spatial Priors






Raw DT
After spatial priors
Adjusted for sample bias
Final map

Unsupervised Analysis of Gridded Time Series

A. Independent Component Analysis (ICA)

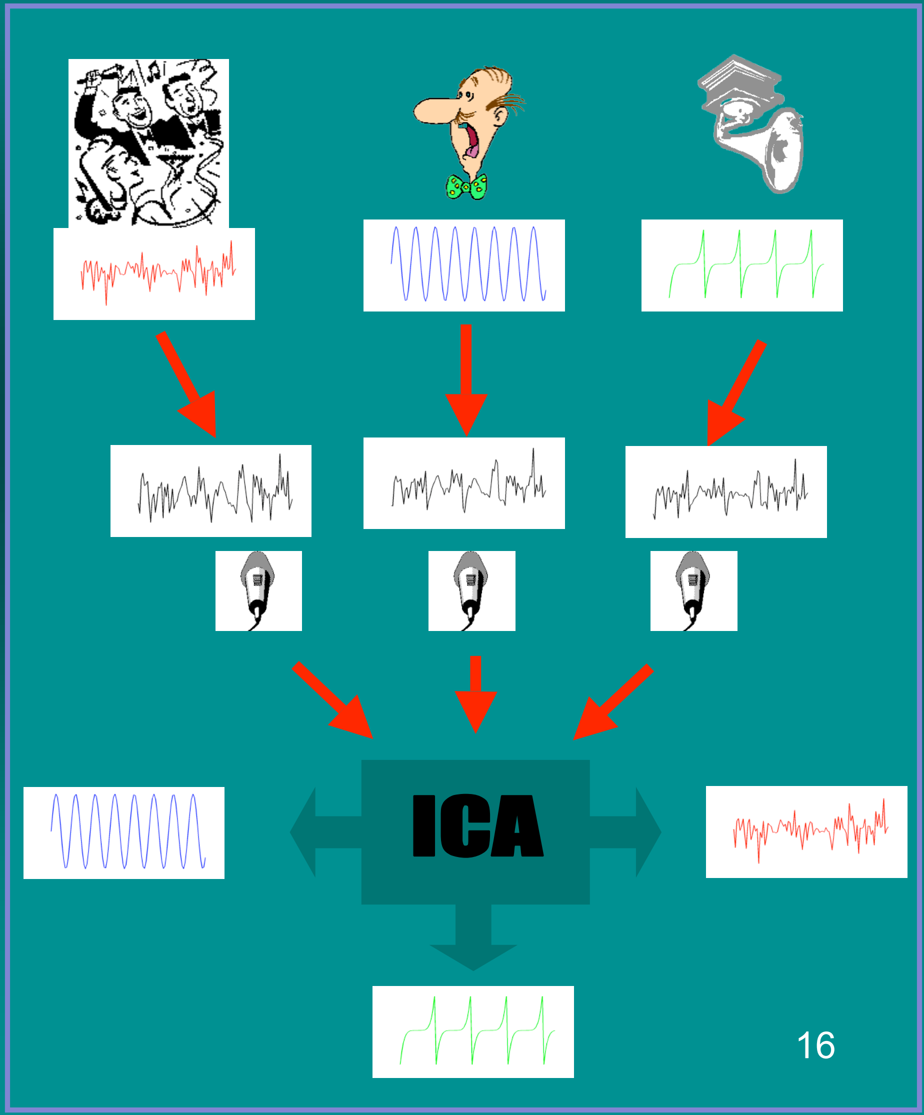
- Non-linear decomposition of temporal variance
- Feature extraction from NDVI time series

B. Principal/Canonical Correlation Analysis (PCA/CCA)

- Joint (linear) variability of global vegetation and precipitation
- Analysis of NH drought and SST patterns

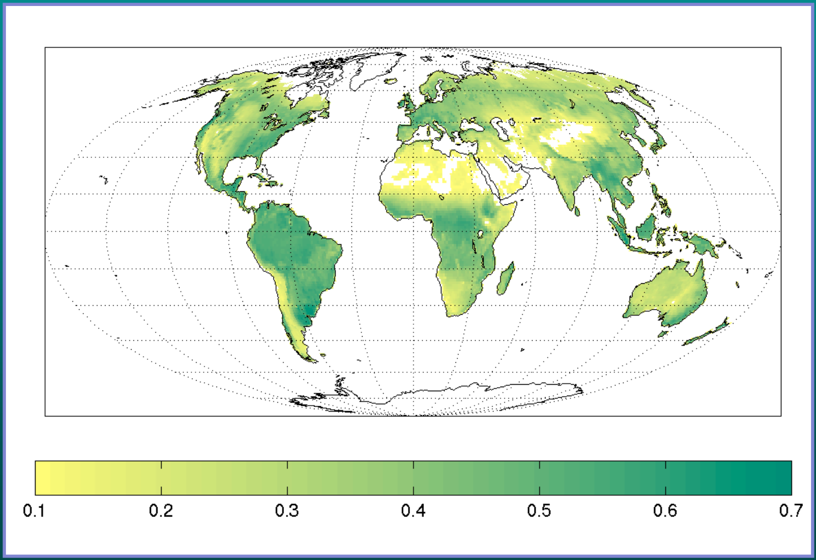
Independent Component Analysis of Time Series NDVI

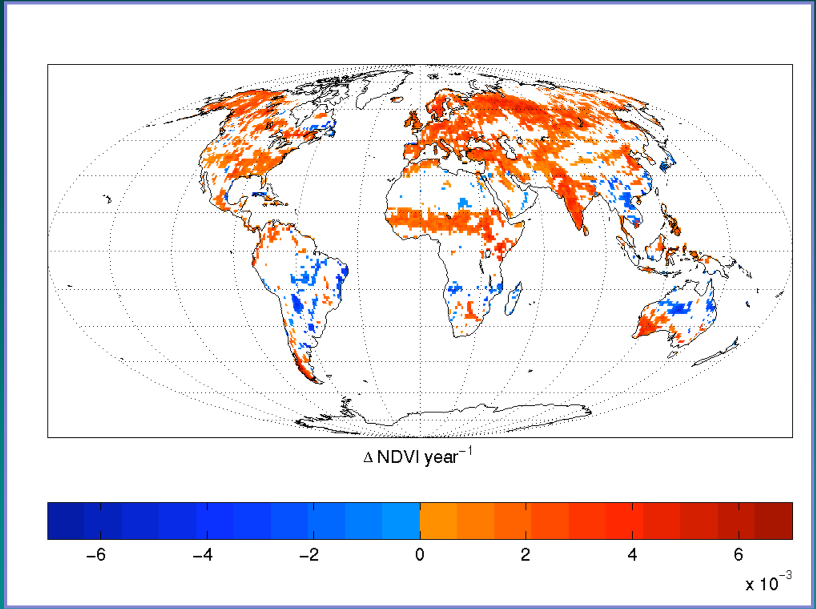
- Independent signals are convoluted and recorded by a sensor (e.g. microphones, satellite instrument)
- ICA separates the signal mixtures into the original source signals
- Independent, not just uncorrelated
- Blind Source Separation – no *a priori* knowledge about the sources
- Looking for hidden sources of variance in time series



FASIR-NDVI

- ***F**ourier **A**djusted*
  ***S**olar zenith angle corrected*
  ***I**nterpolated*
  ***R**econstructed*
  ***N**ormalized **D**ifference **V**egetation **I**ndex*
- NOAA (7,9,11,14)-AVHRR
- Monthly 1982-1998
- 1x1 degree spatial resolution

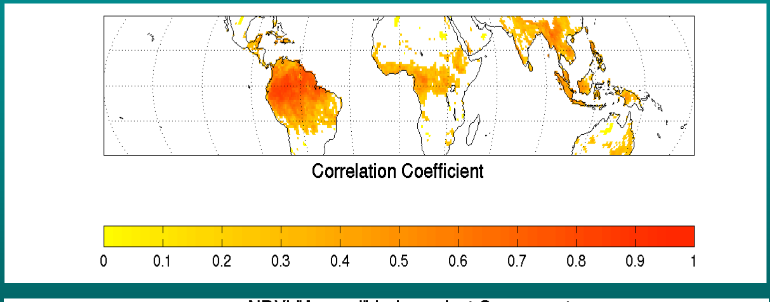


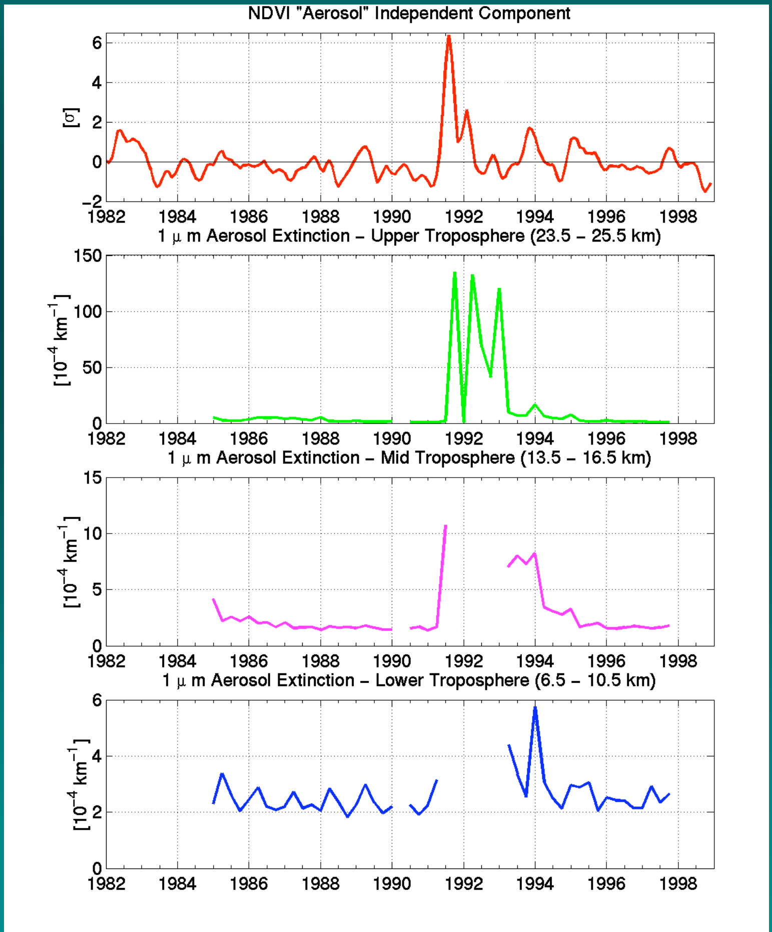


Los et al. (1994), Tucker et al. (2001)

“Aerosols” IC

- Residual aerosol signal in tropics
- Co-variation with Stratospheric Aerosol and Gas Experiment (SAGE) II data 1985-1998
- Not revealed via linear methods like PCA

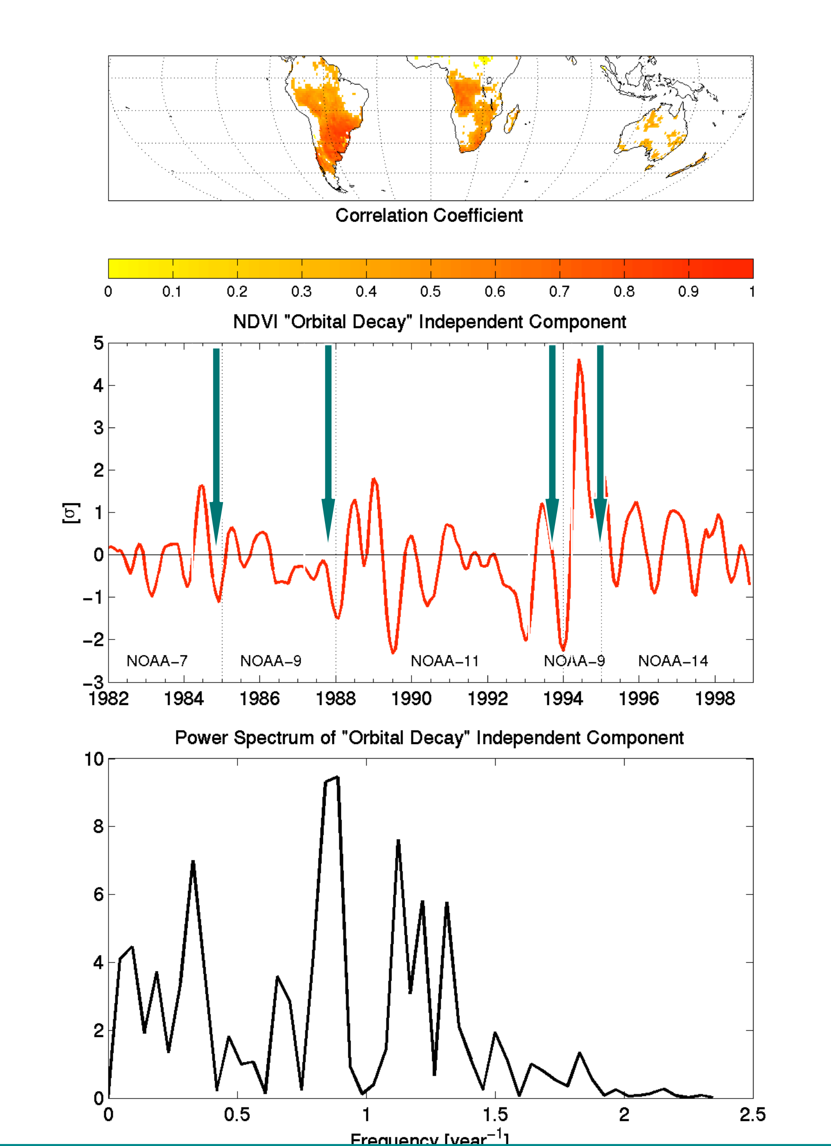




Lotsch et al., IEEE TGARS, 2003

“Orbital Drift” IC

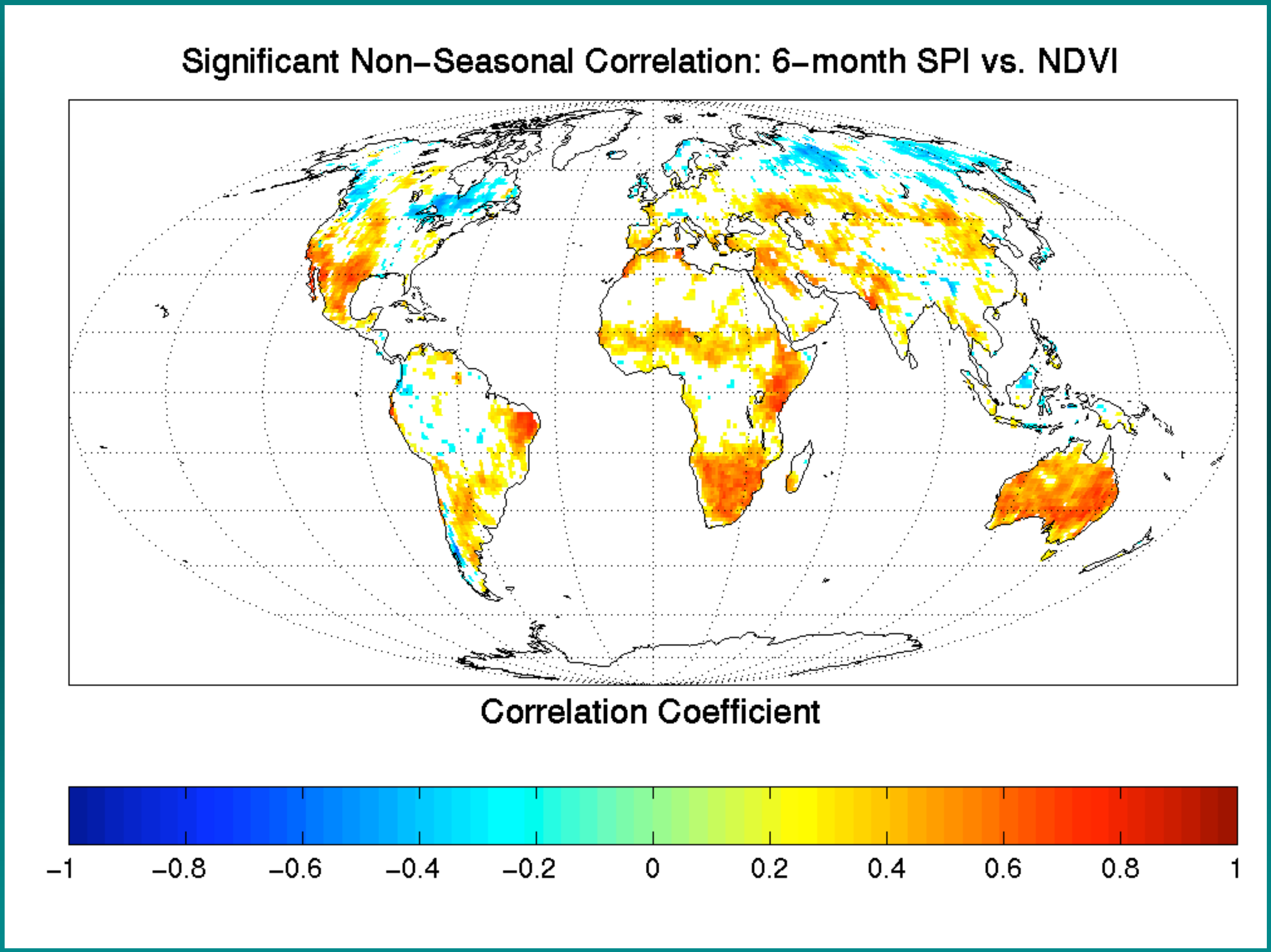
- Discontinuities coincide with AVHRR sensor changes
- Reflect changes in sensor view geometry & orbital drift
- Limited to southern latitudes



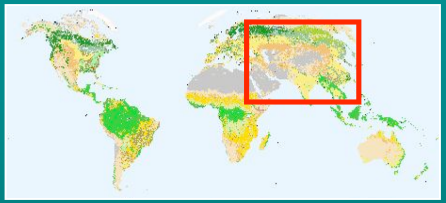
Lotsch et al., IEEE TGARS, 2003

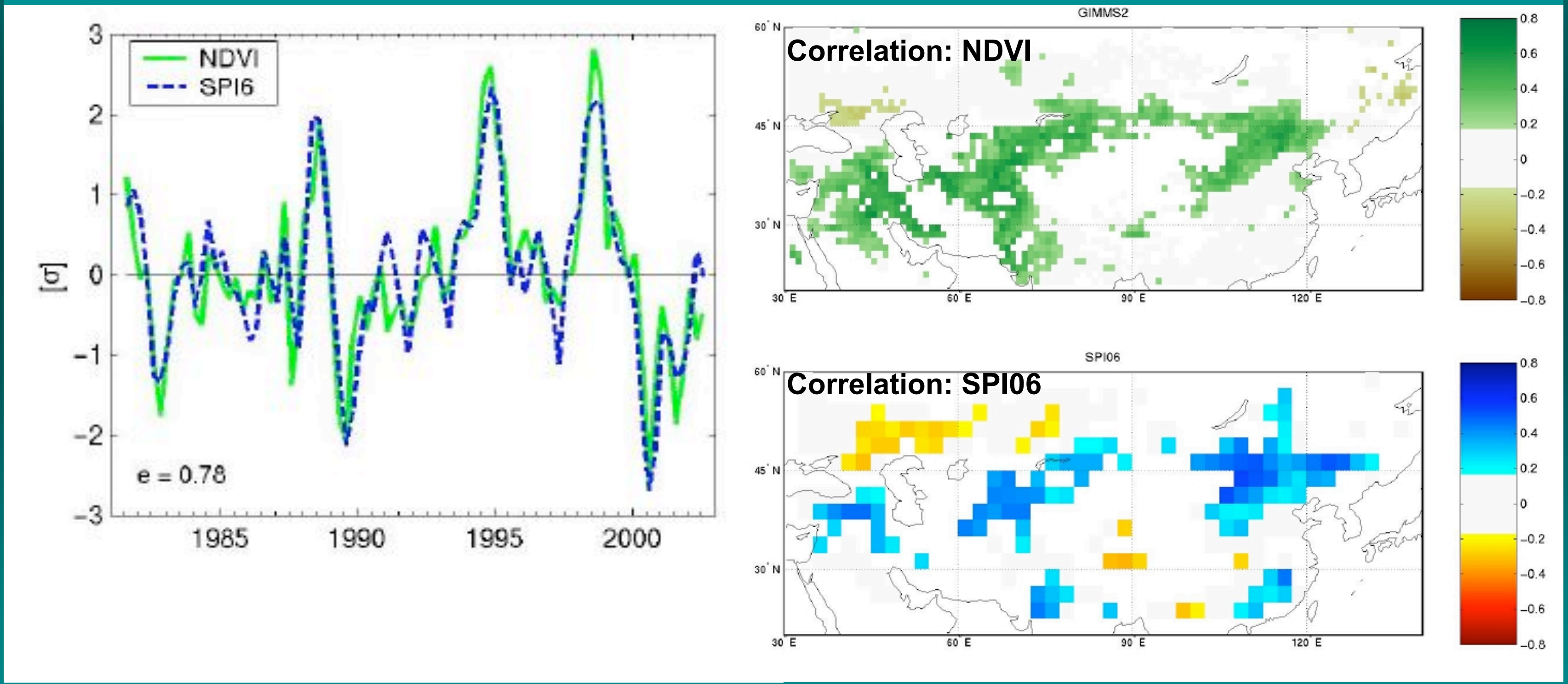
Joint Variability in Climate & Vegetation

(GIMMS-NDVI vs Standardized Precipitation Index 7/1981-3/2003)



Canonical Correlation
Example: Eurasia (CF1)



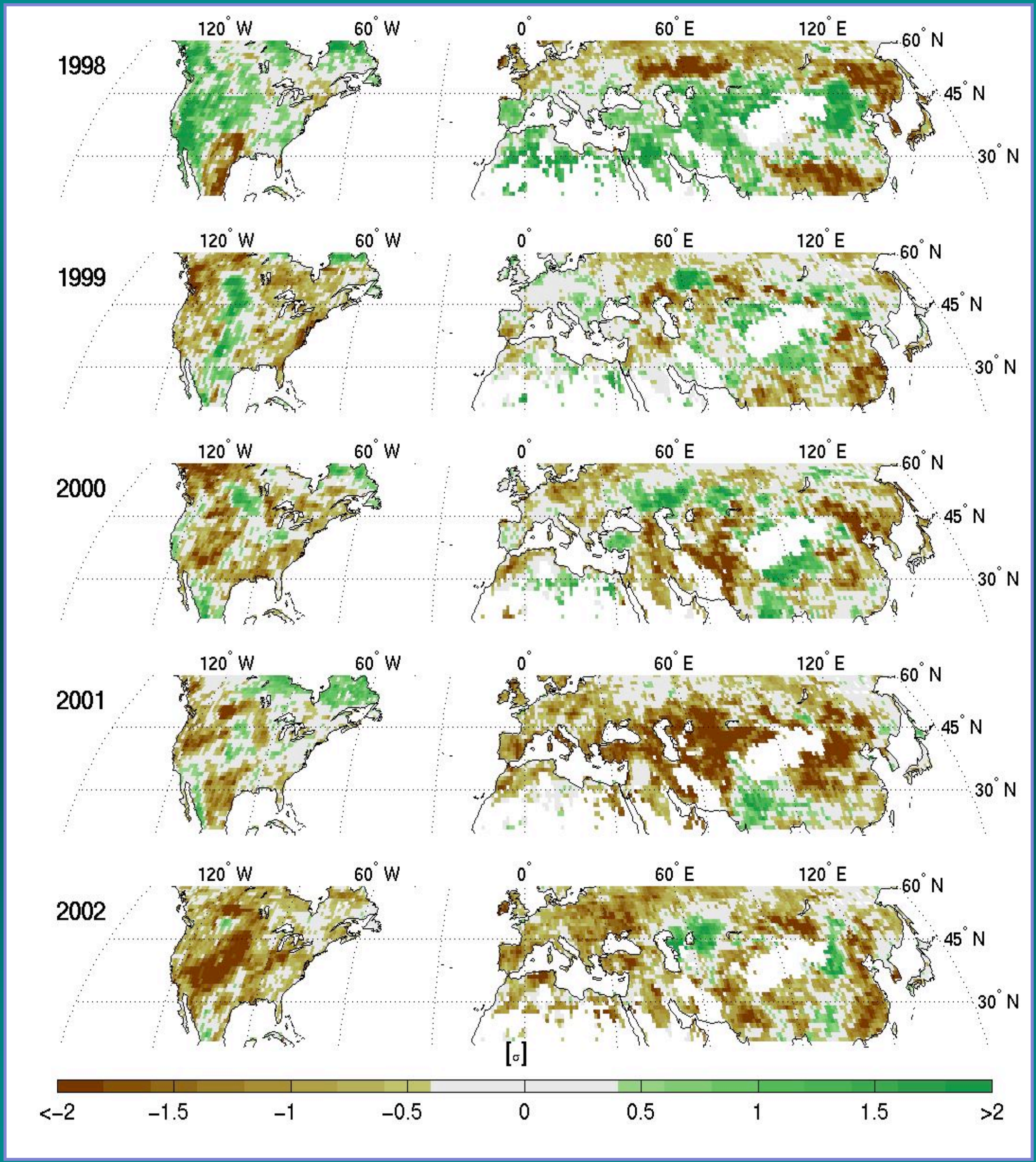


Lotsch et al, *Geophysical Research Letters*, 2003

1998 –2002 Northern Hemisphere Mid-Latitude Browning

June-August standardized anomalies in NDVI relative to 1981-2002 mean

Motivated by Hoerling and Kumar 2003, Perfect Ocean for Drought, *Science*



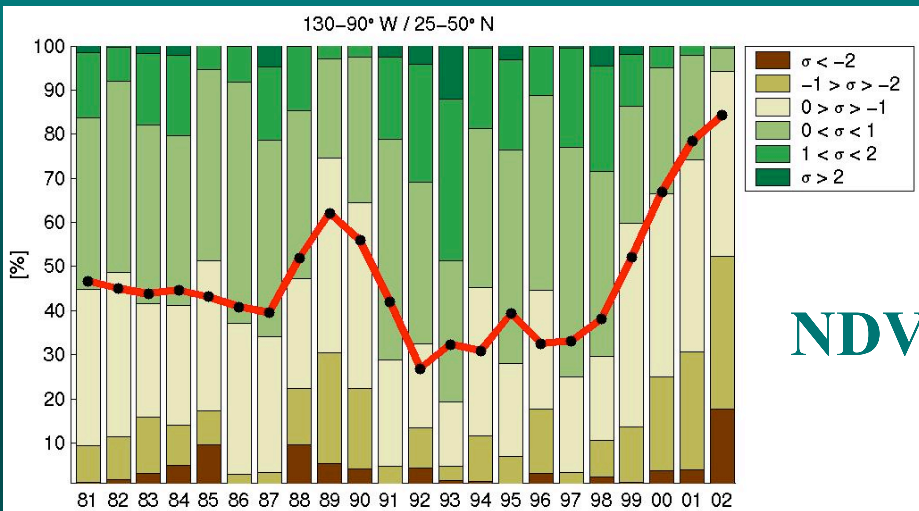
Lotsch et al. (2005) *Geoph. Res. Letters*

NDVI and SPI Anomalies

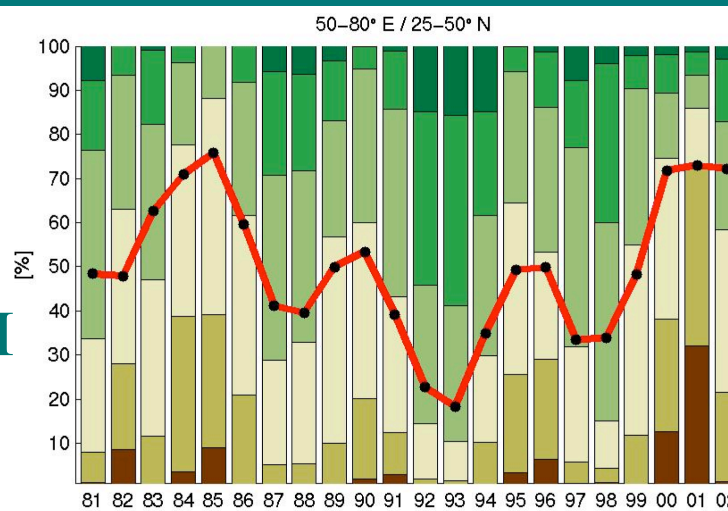
May-September 1981-2002

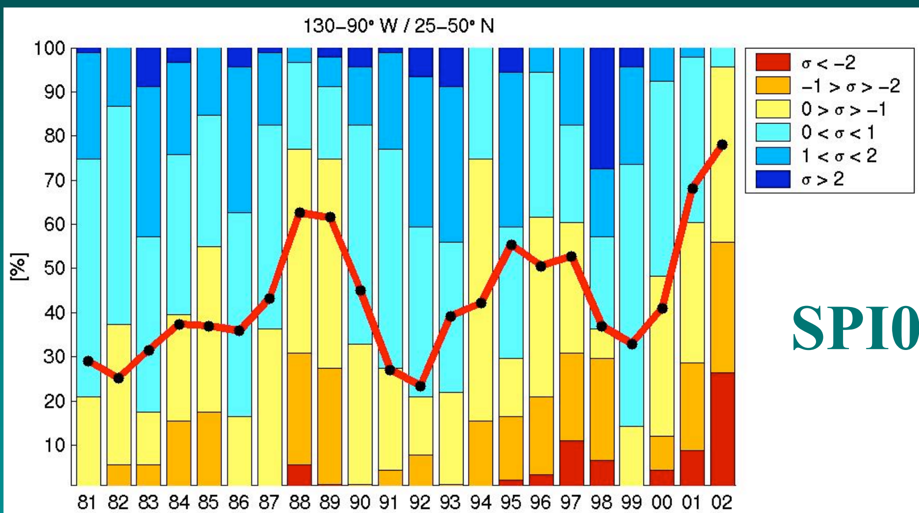
North America *130°-90°W*

CSW Asia *50°-80°W*

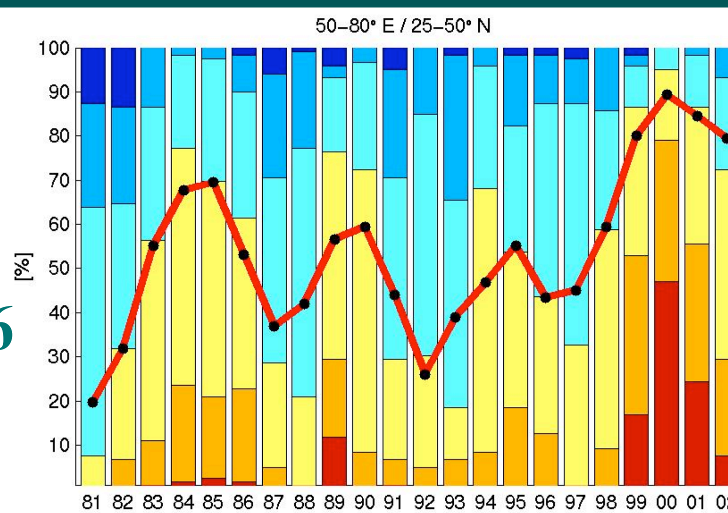


NDVI





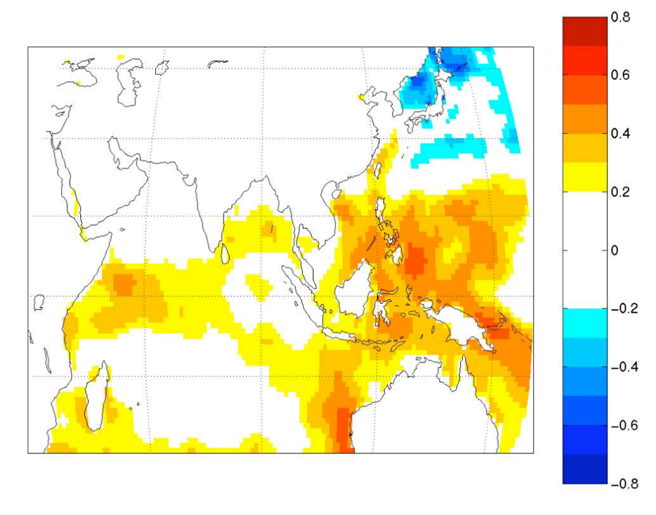
SPI06

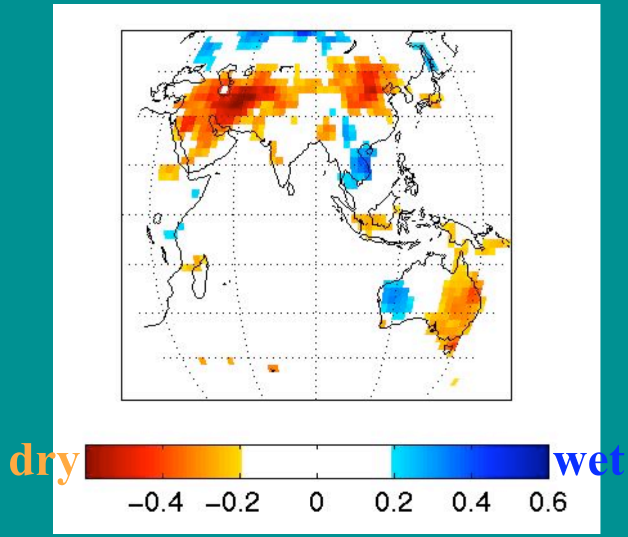


Ocean-Drought Teleconnections

e.g., Eurasia & Australasia 1948-2002

Correlation with SST (MAM)

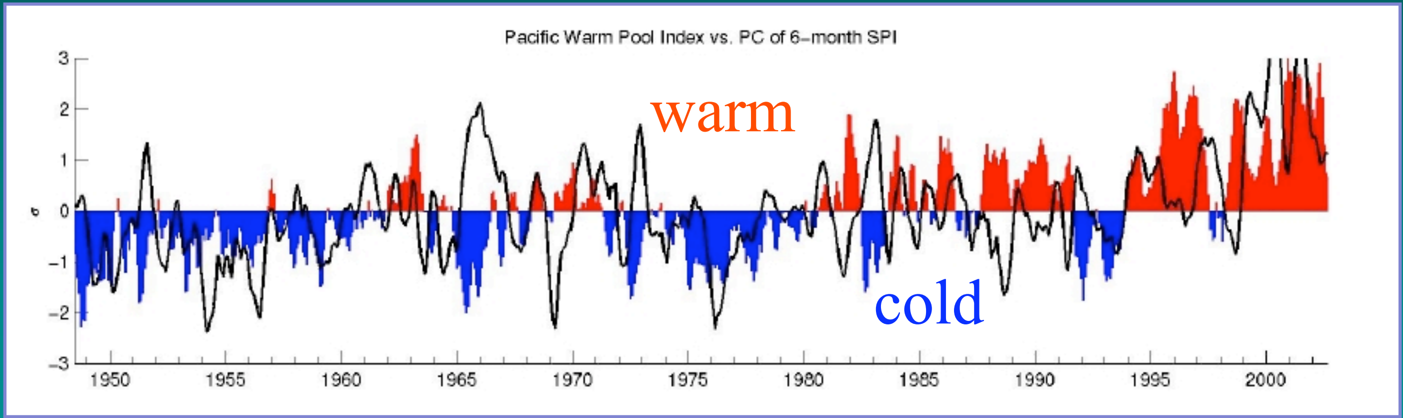




SPI pattern

Pacific Warm Pool

PC1 of SPI06



Conclusion

Unprecedented reduction of plant photosynthetic activity linked to synchronous patterns of sea surface temperature fluctuations and extensive patterns of drought in the Northern Hemisphere mid-latitudes during 1998-2002

ΔSST

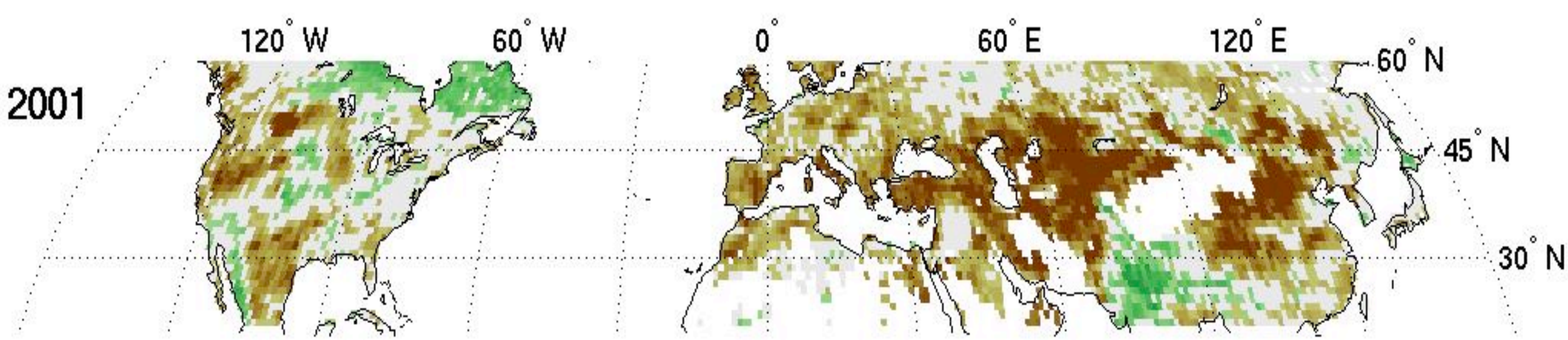
Pacific + Atlantic + Indo-Pacific

⬇

NH Precip Regimes

⬇

Plant Photosynthesis



Ongoing Work

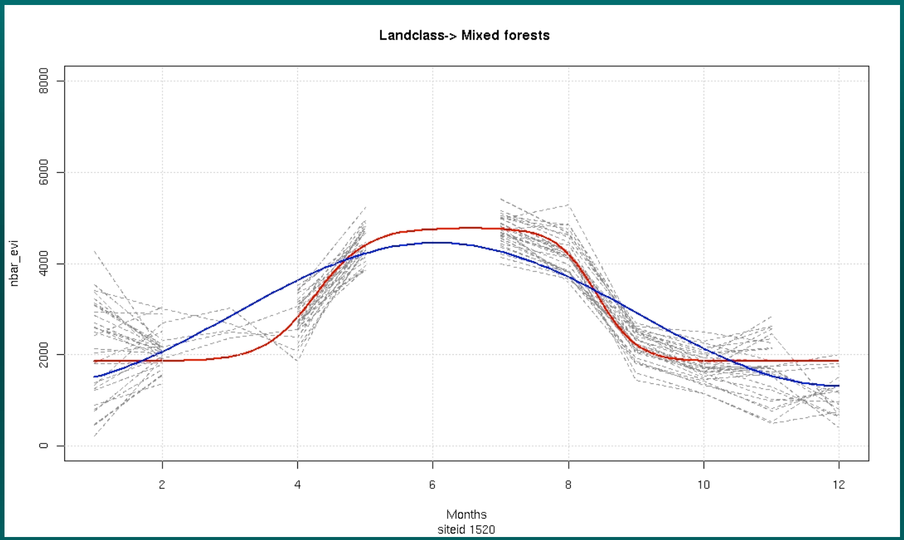
- Functional Data Analysis & Modal Clustering
  - Basic Question: *How to best characterize temporal patterns and reduce feature dimensionality?*
  - Functional Model - Double Logistic:

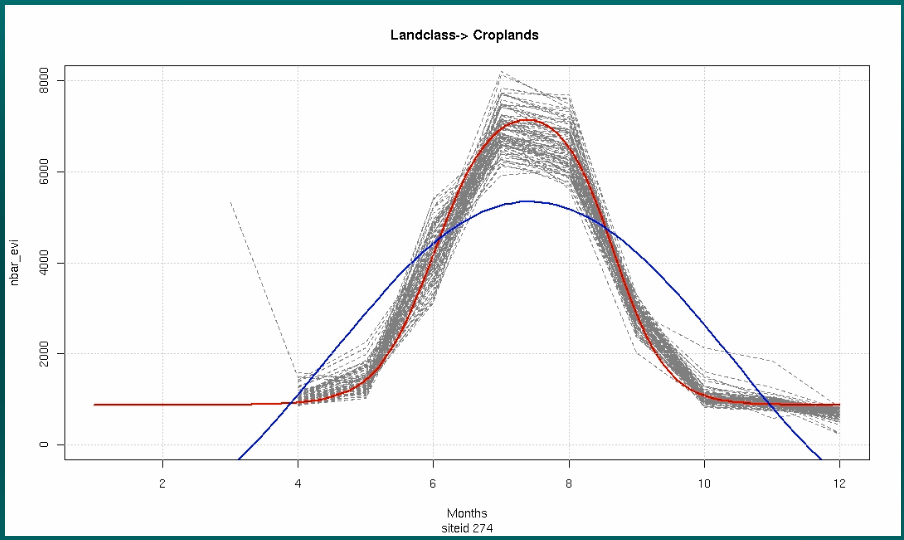
$$Y(x) = a_1 + (a_2 - a_1)\left(\frac{1}{1+\exp(-a_3(x-a_4))} + \frac{1}{1+\exp(a_5(x-a_6))} - 1\right)$$

  - captures timing, magnitude & form of temporal variation

(a_1= min; a_2= max, a_3=angle of inflection 1; a_4= time of inflection 1
a_5=angle of inflection 2; a_6= time of inflection 2)

Sample Double Logisitic Fits



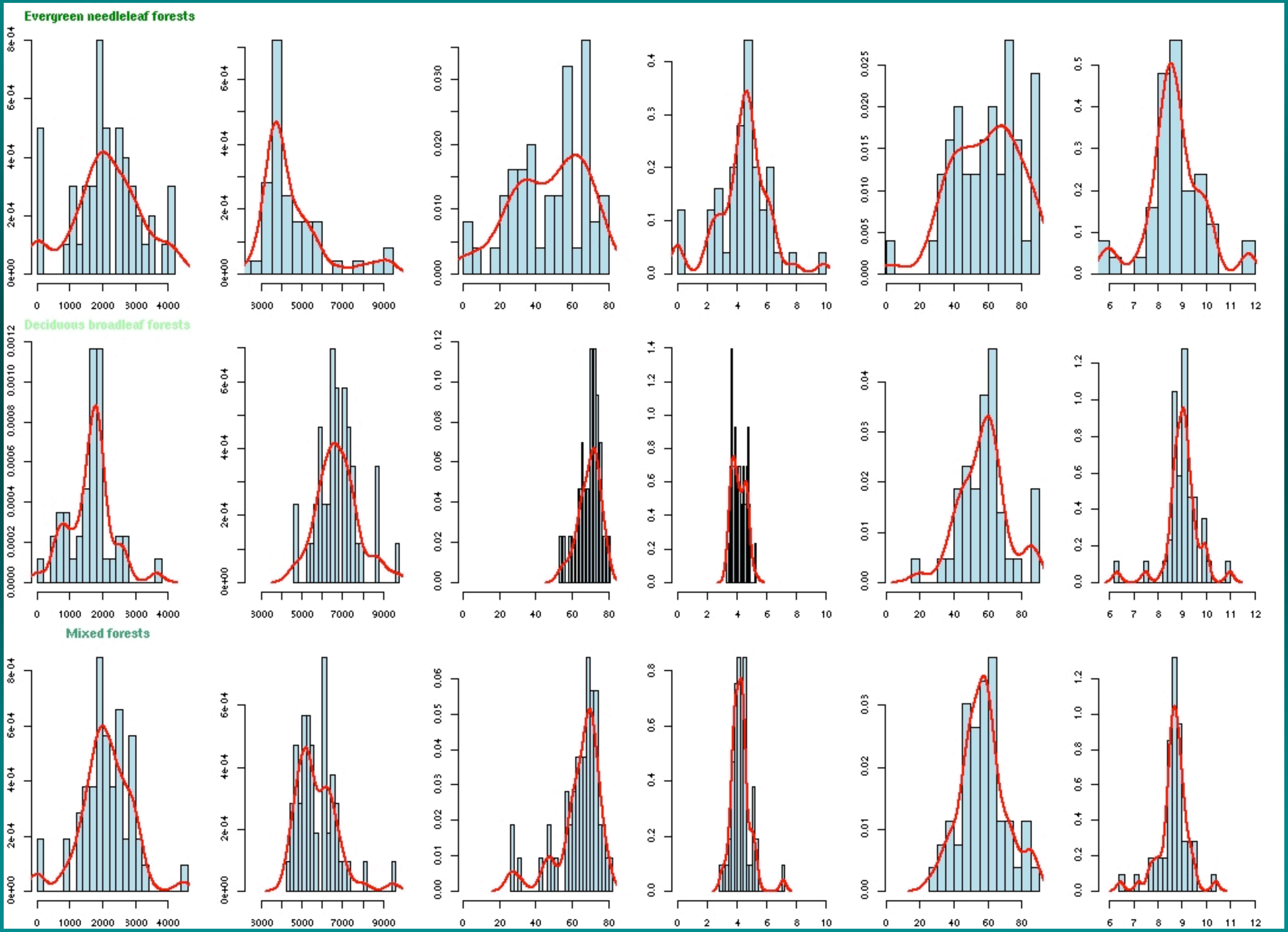


Blue: symmetric (Fourier-based) model;

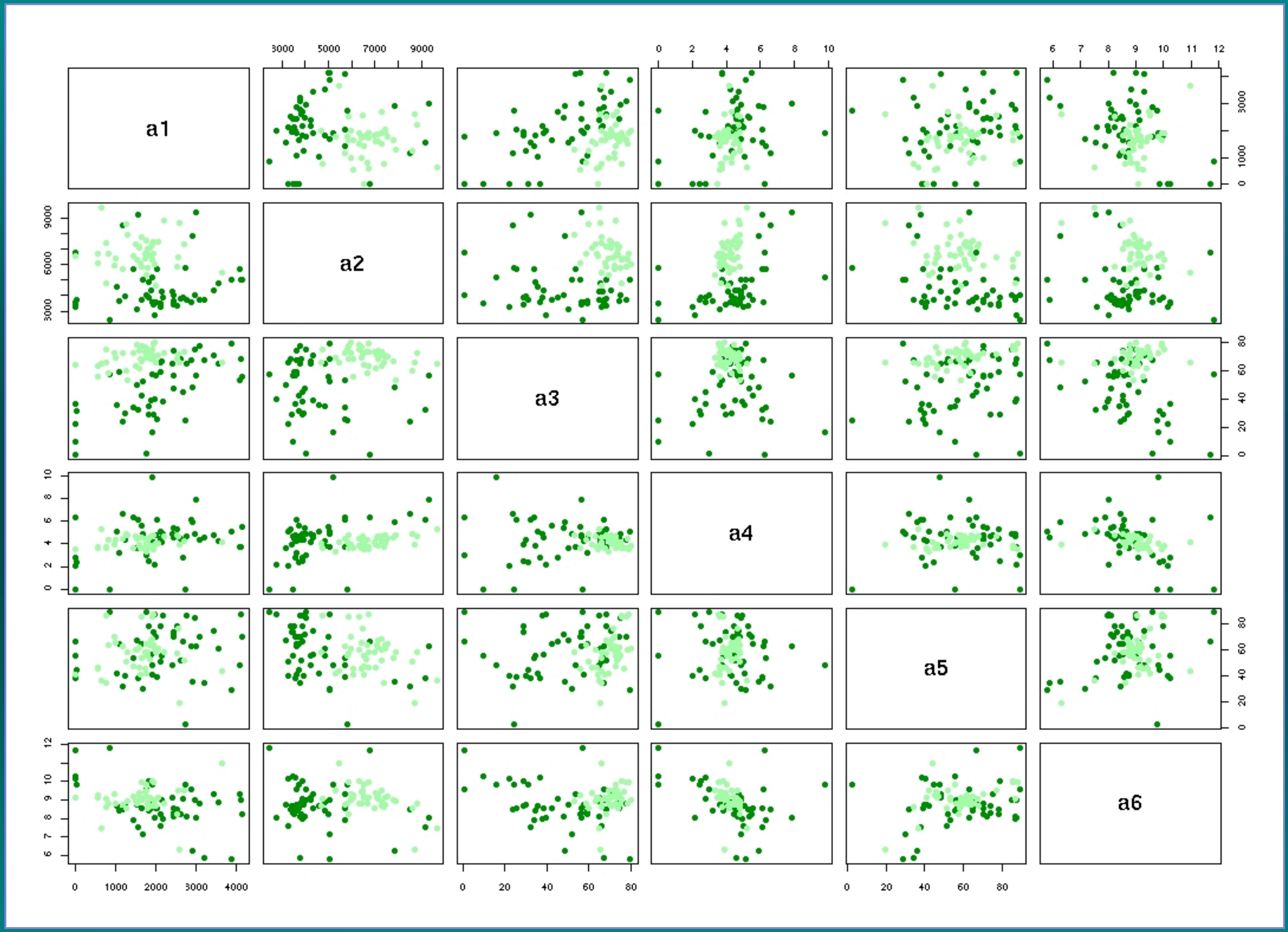
Red: double logistic

Note fit, missing values

Distribution of Coefficients Across Classes



Clustering



Next step: compare clustering of original data w/coefs from functional model using modal clustering to deal with non-normal distributions

Ray, S. and Lindsay, B. G. (2008). Model selection in high dimensions: a quadratic-risk-based approach. *J. Roy. Statist. Soc. Ser. B*, 70(1):95–118.
Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods., *J. Roy. Statist. Soc.* Ser. B, 68(2):305–332.

Conclusions: Technical

- *Supervised Learning*
  – It's not just the learning algorithm…..
  – Data and biases associated with training data are what count
    - Unbalanced training data
    - Feature selection
    - Active Sampling or identifying redundant training data
    - How to stabilize classification results across years
- *Unsupervised*
  – Linear vs non-linear methods; Gaussian vs non-Gaussian
  – Danger of fishing expeditions
    - Analyses need to be hypothesis driven
  – Toolkit feels less mature, esp for very large data sets.
    - Clustering, PCA, CCA, etc. (may reflect my ignorance)
    - Dimensionality, feature selection key challenges.

Conclusions: General

- *Data mining in Earth Sciences is hard*
  - Looking for causal relations, not just patterns
  - Need teams to prevent natural scientists from doing naïve analysis and computational scientists from doing naïve science
  - NASA should be supporting this – interests in missions and measurements in support of science
- *Need to foster community*
  - Funding?
  - Publishing:
    - Where to publish this work?
    - Is it technical or is it science?
  - Where to present? What meetings?