IDS.012 / IDS.131 / 6.419 / 6.439
Statistics, Computation & Applications

Lecture 10:

Environmental Data & Gaussian Processes

October 10, 2018

Announcements

- **Projects**: If you are assigned to a group but are dropping the class, let your team members and us know now.

Environmental Data – Examples

- Air Quality, Water Quality (pollutants)
- Weather & climate (temperature, winds, moisture, precipitation, extreme conditions, . . .)
- Storms
- Ocean dynamics
- Vegetation (forests, algae, . . .)
- Wildlife monitoring
- Earthquake magnitudes

Why do we care? What are questions of interest?

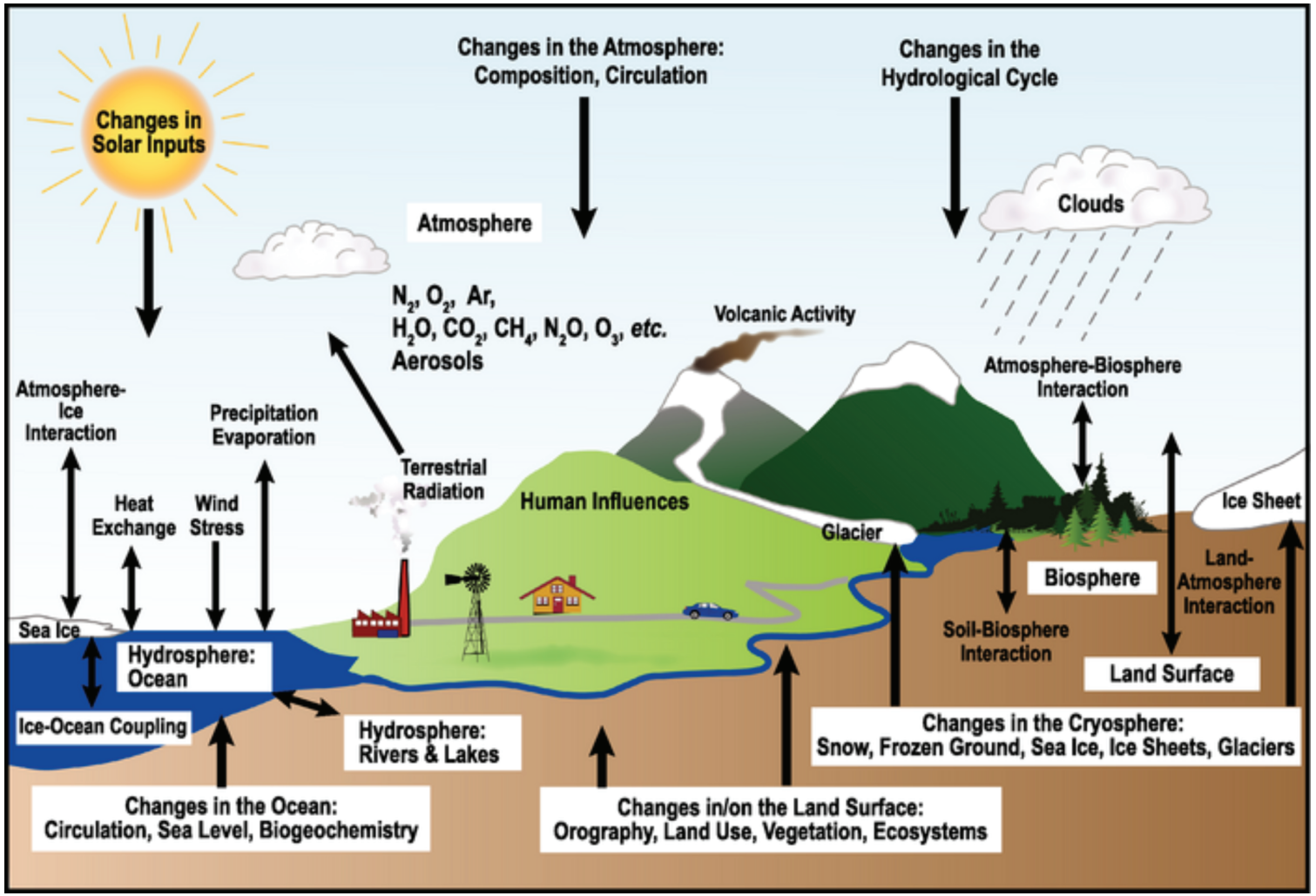
Environmental Data – Why do we care?

- understand underlying processes, changes
  (e.g. climate change: "statistics of weather over time")
- impacts on environment, health, economics, society
- policies
- forecast events, warnings (e.g. community seismic network, storms, ...)
- energy management with renewable energies
- use in planning, routing, backtracking, control (ships, airplanes, ...)

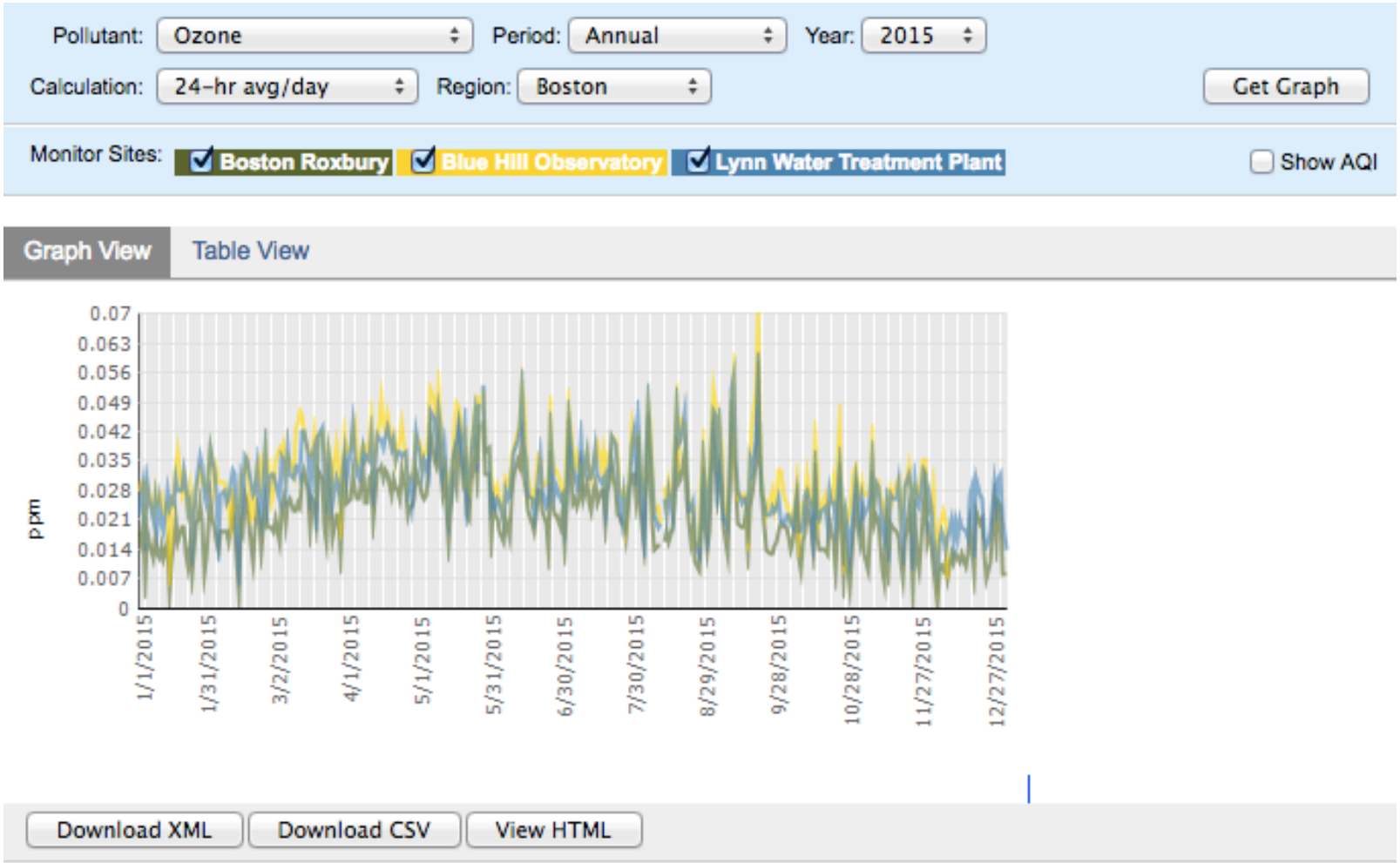
Questions

- relationships (correlations, association)
- trends; forecasting
- planning
- quantifying uncertainty, adaptive sensing

Environmental Data – What is special?



Environmental Data – What is special?



spatial correlation

from `https://www.epa.gov/outdoor-air-quality-data`,
`http://public.dep.state.ma.us/MassAir/Pages/MapCurrent.aspx?&ht=1&hi=101`

Environmental Data – What is special?

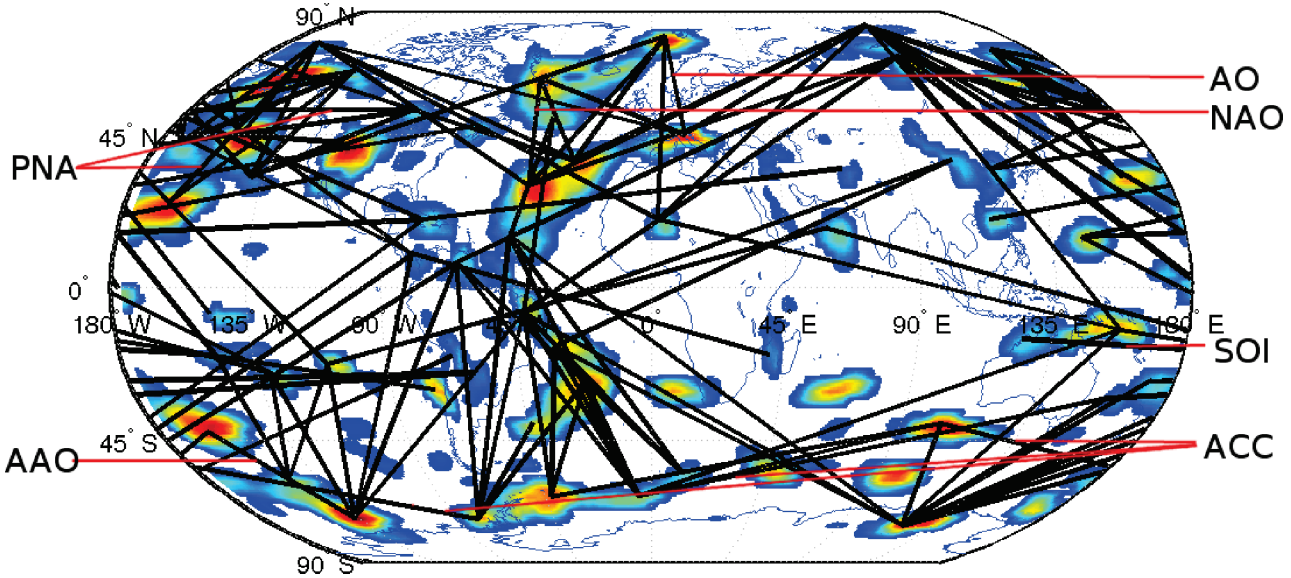
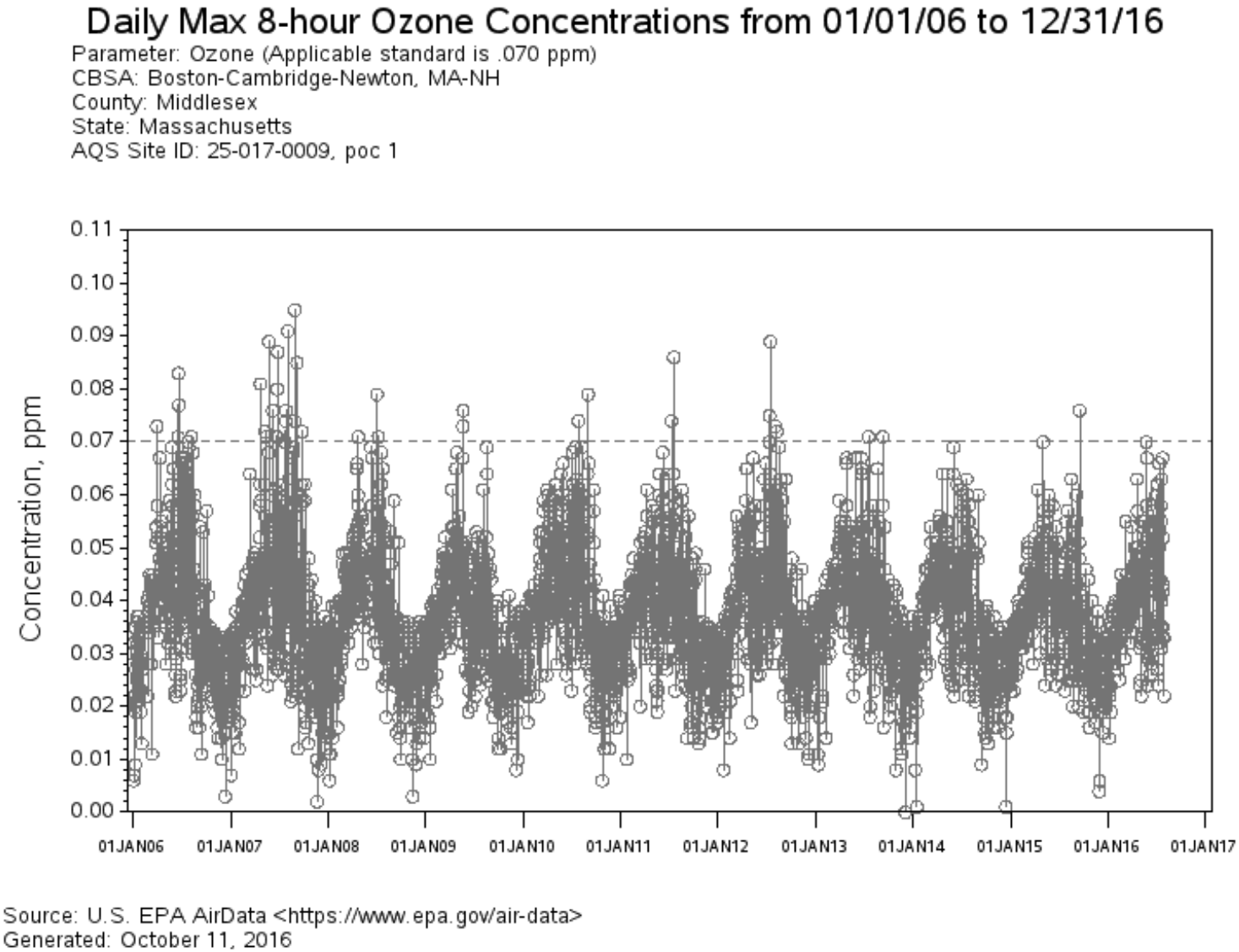


Figure 1.8. *Dipoles in NCEP sea-level data for the period* 1948–1967. *The color background shows the regions of high activity. The edges represent dipole connections between regions.*

Middlesex ozone 2006-2016



temporal correlation (periodicity): seasons

Environmental Data – What is special?

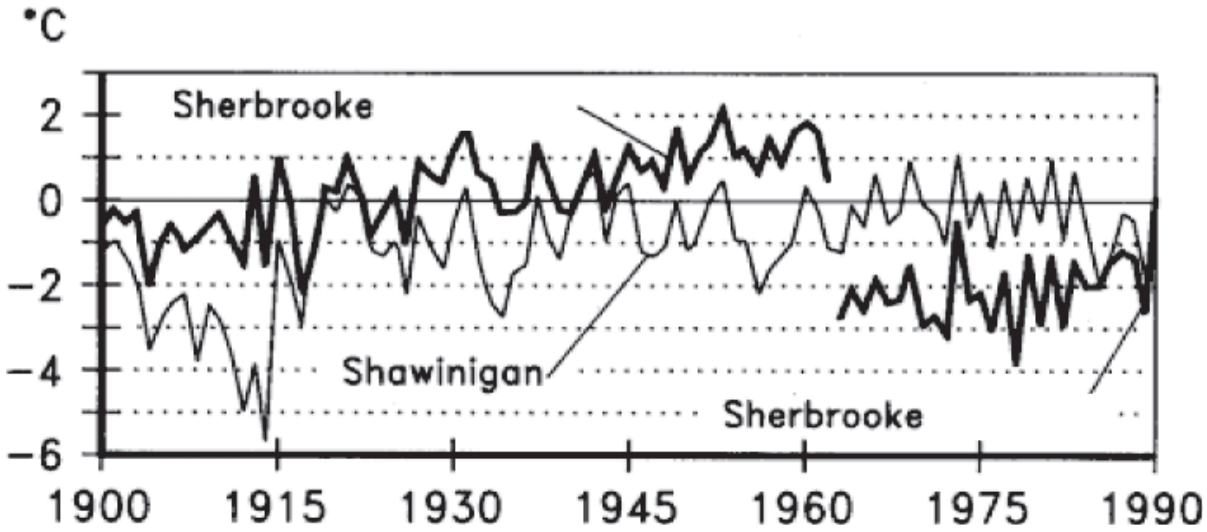


Figure 8.1. *Temperature records for two towns in Québec* [117].

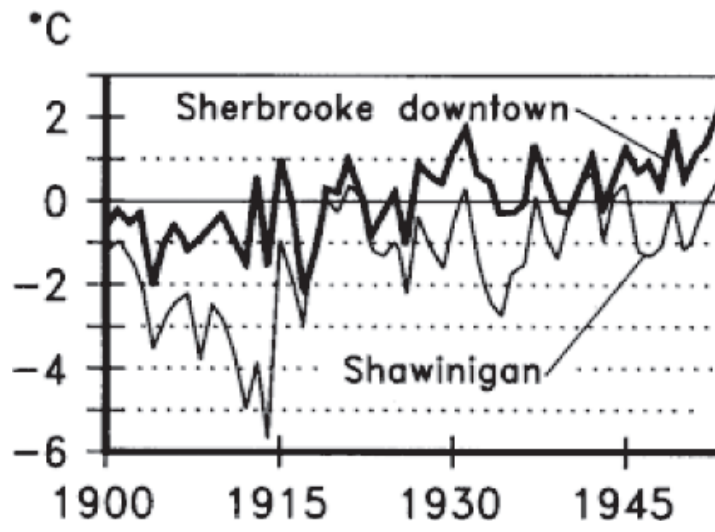


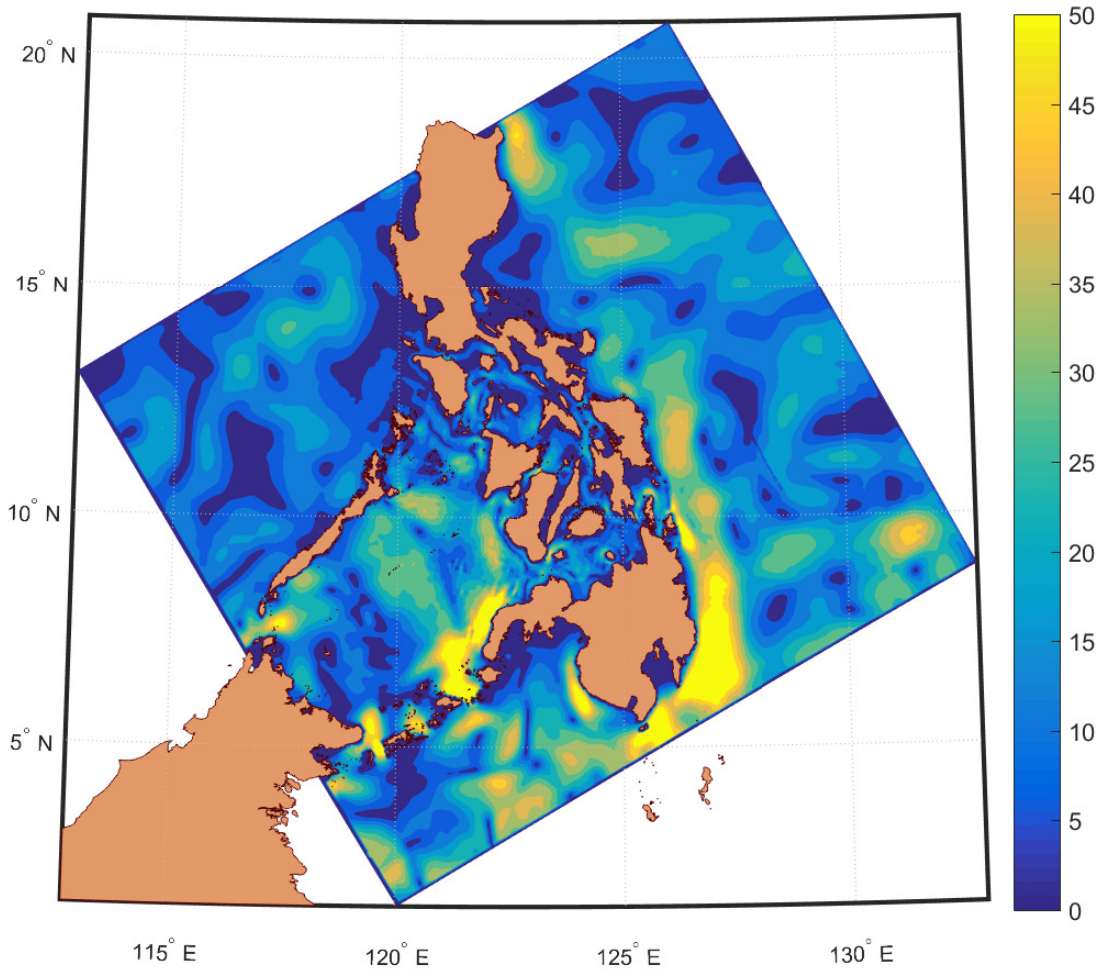
Figure 8.1. *Temperature records for tw*

heterogeneous data (different measurement stations' biases, e.g., urban heat island effect)

image source: H. Kaper, H. Engler. Mathematics & Climate

Environmental Data – What is special?

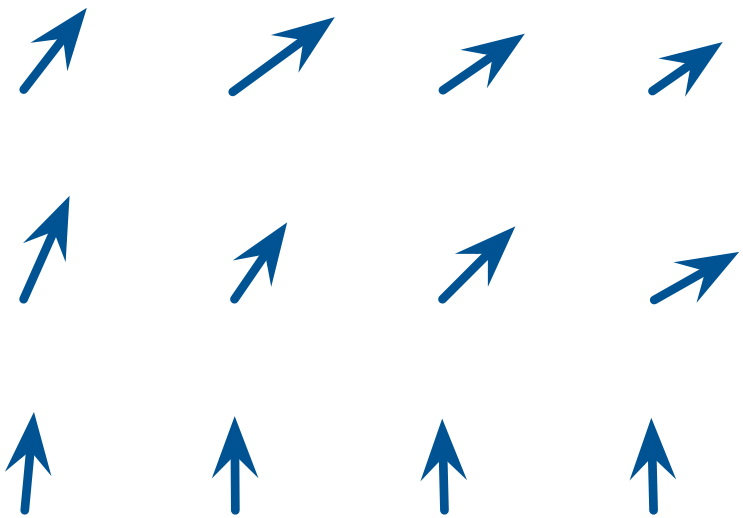
- correlations in time
- correlations in space (fields)
- scientific models + statistics; simulations
- methodological challenges for statistics
  - no controlled studies, only observational data
  - hypotheses from data
  - large data sets
  - heterogeneous data
  - proxy data



Our data: flow in space & time

Flow fields

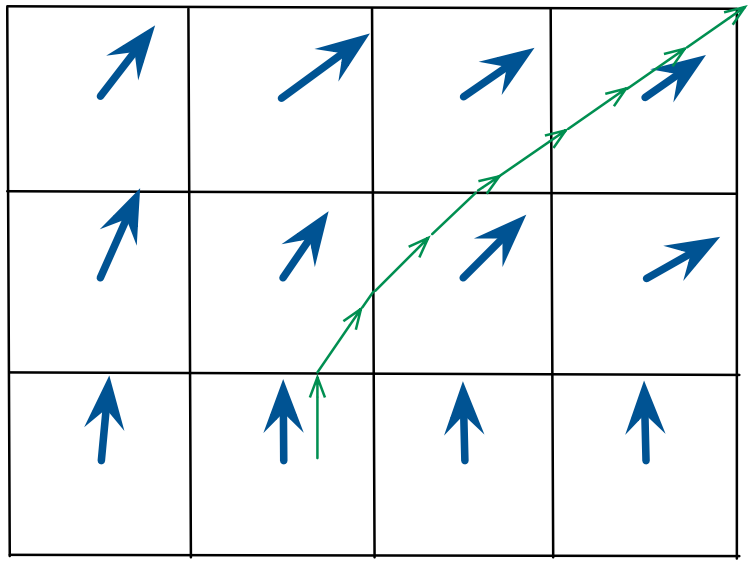
- $V(x, t)$: flow at location x at time t
  vector in 2d or 3d
- velocity: vector length



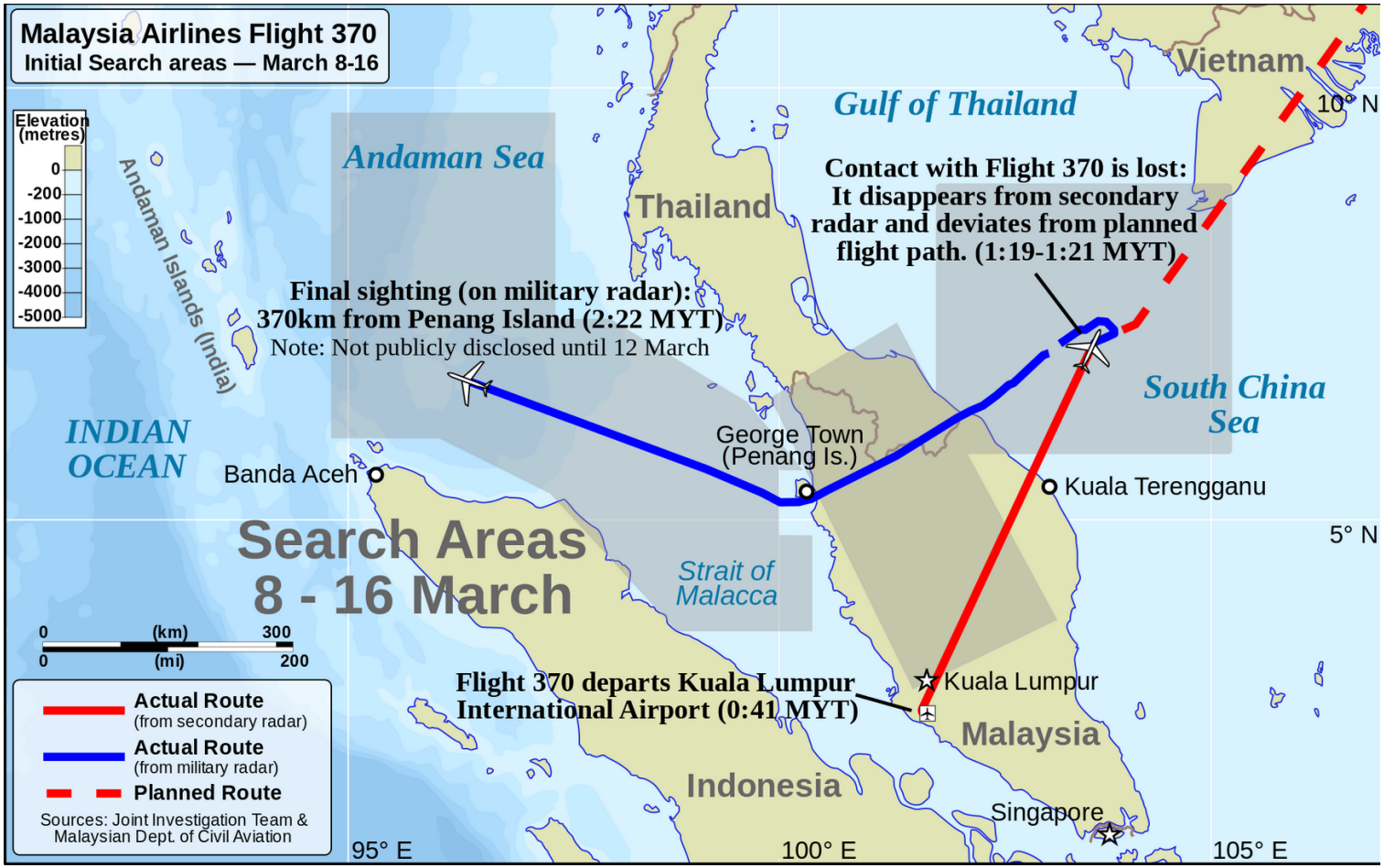
$V(x, t)$ at a fixed time t
What does variation in time mean?

Working with flow fields

- forward prediction
  simulate (propagate) distributions, include variation in time
- hindcasting



8 March 2014: Malaysian Airlines Flight 370 (MH370) disappeared on its way from Kuala Lumpur to Beijing.



source: Wikipedia

8 March 2014: Malaysian Airlines Flight 370 (MH370) disappeared on its way from Kuala Lumpur to Beijing.

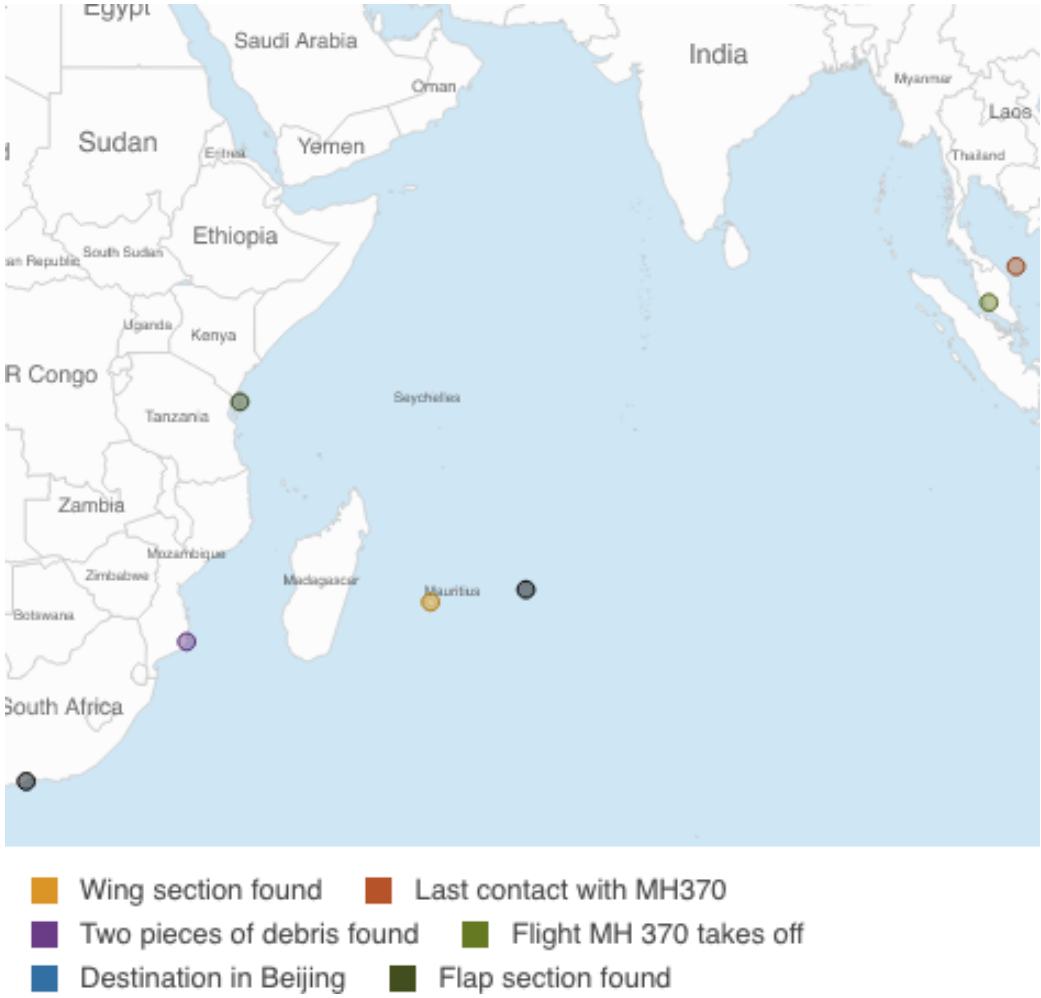
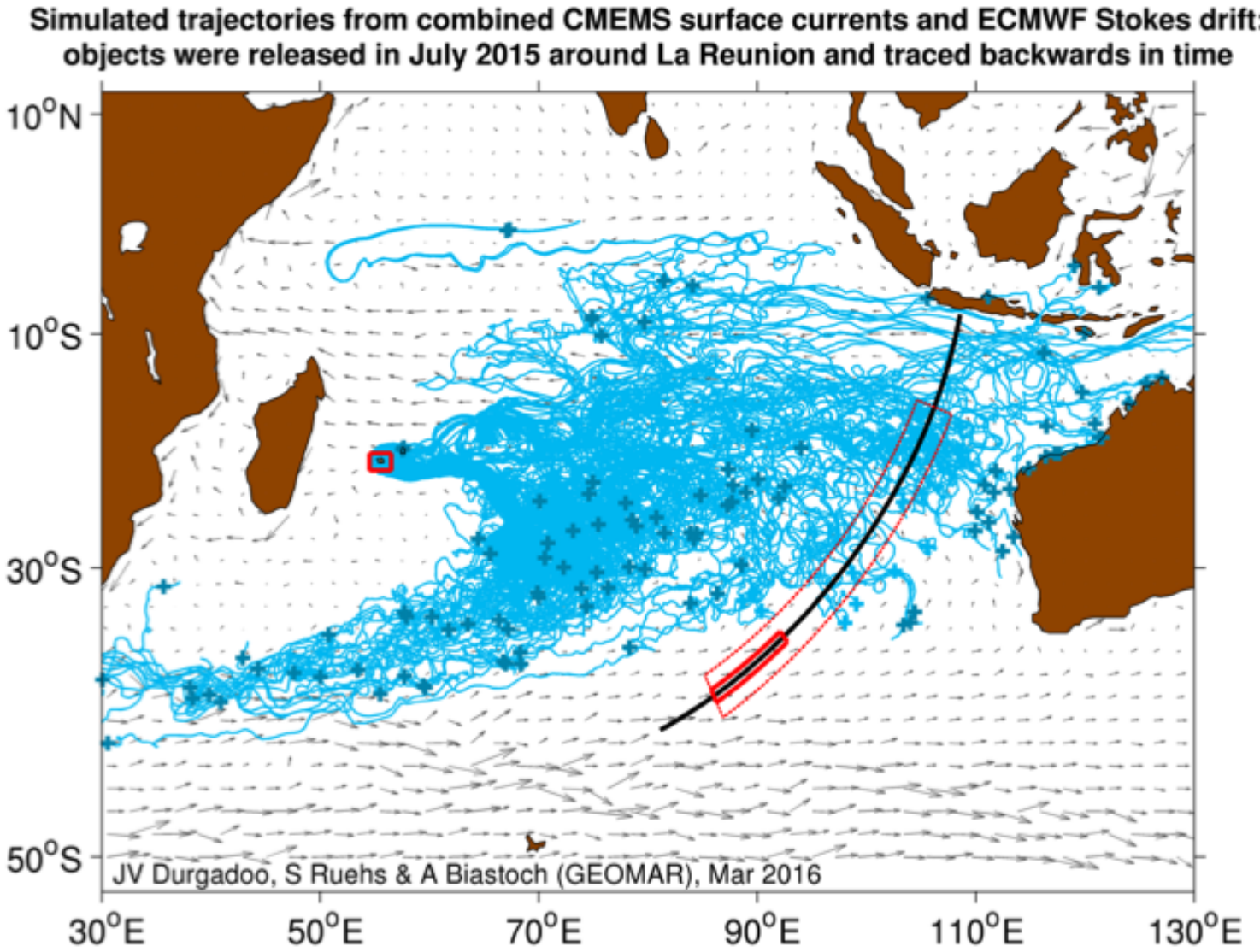
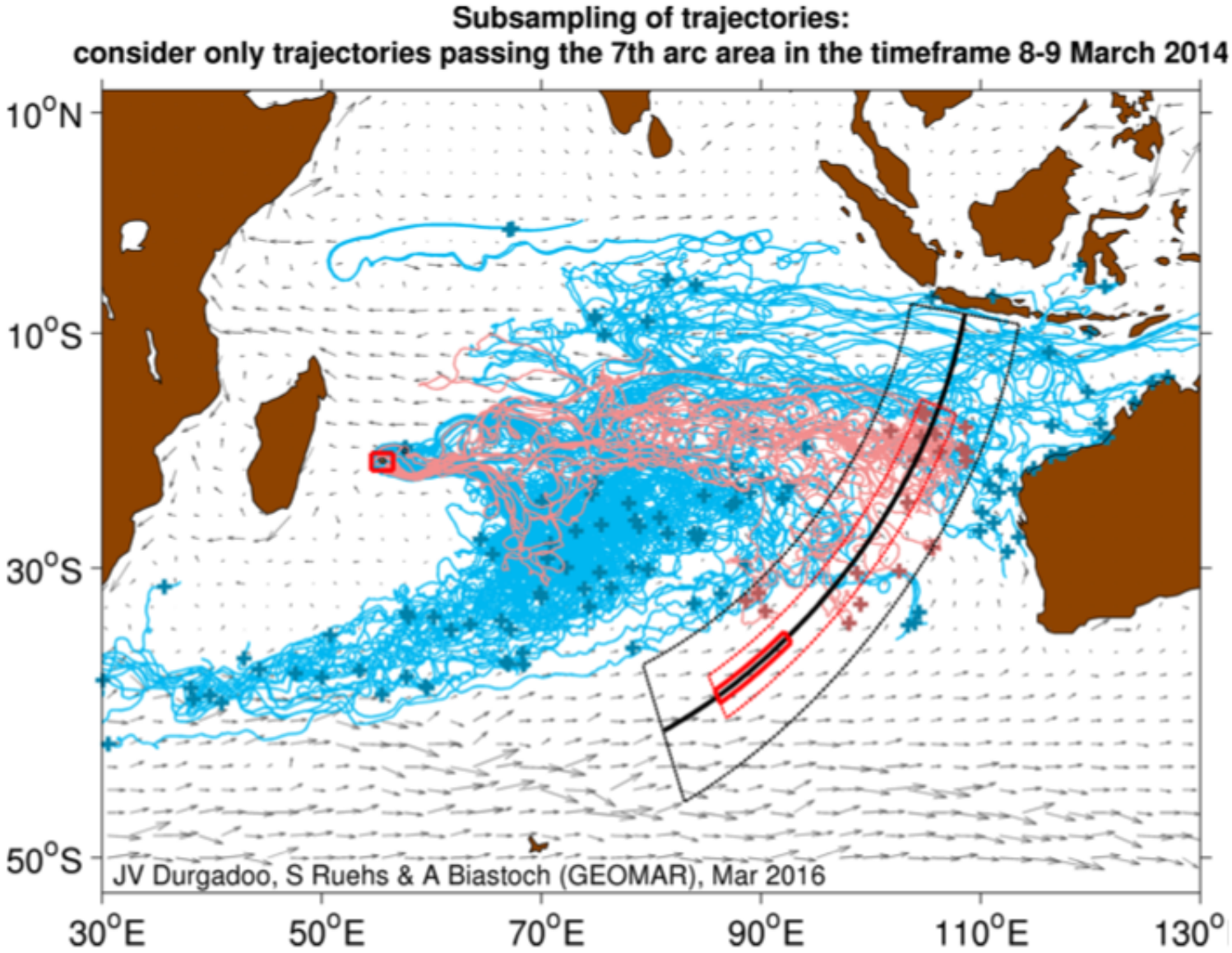


image source: BBC



full report:

`http://www.geomar.de/fileadmin/content/service/presse/Pressemitteilungen/2016/MH370_Report_May2016.pdf`

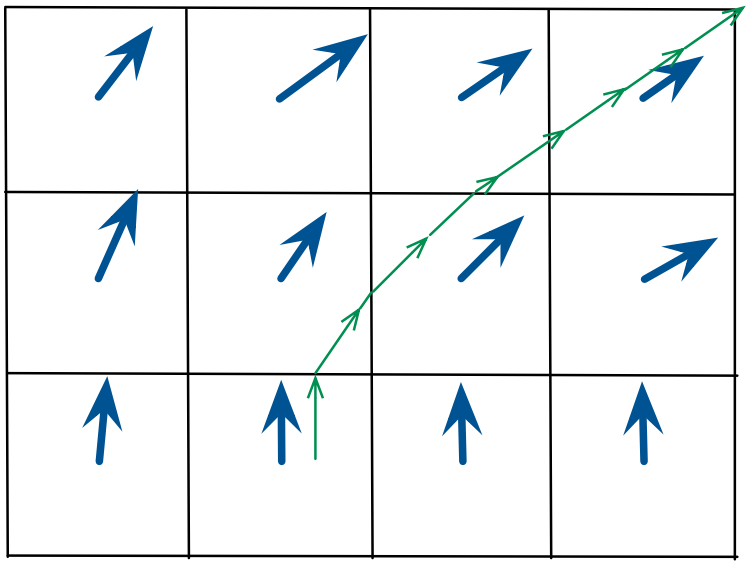


full report:

`http://www.geomar.de/fileadmin/content/service/presse/Pressemitteilungen/2016/MH370_Report_May2016.pdf`

Working with flow fields

- forward prediction
  simulate (propagate) distributions, include variation in time
- hindcasting
- vehicles with engine?



Path planning

- displacement: force from current and steering

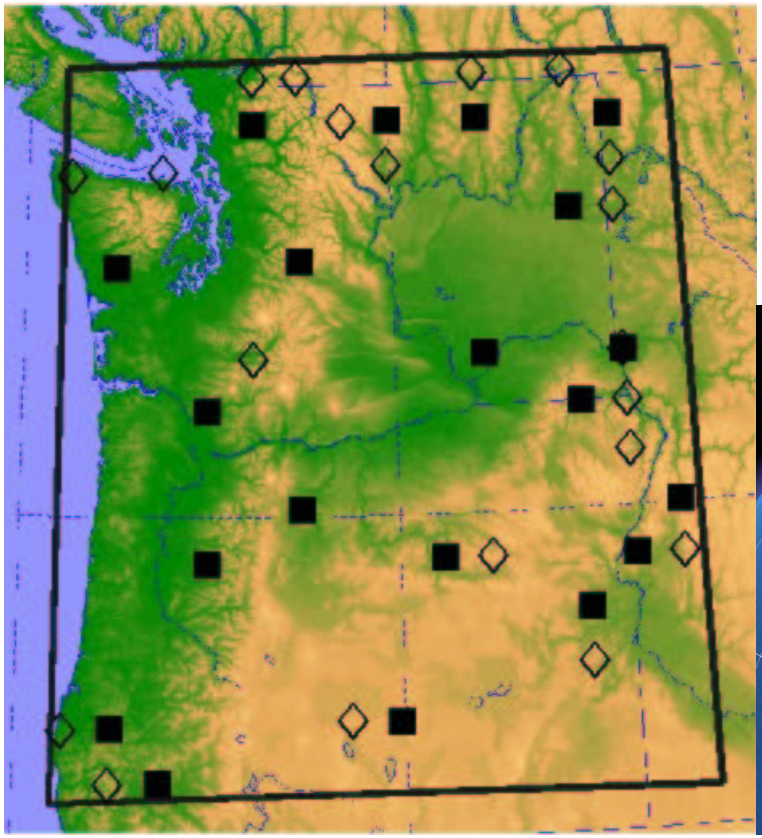
$$X(t+\epsilon) \approx X(t) + \epsilon V(X(t), t) + \epsilon F(t) d(t)$$

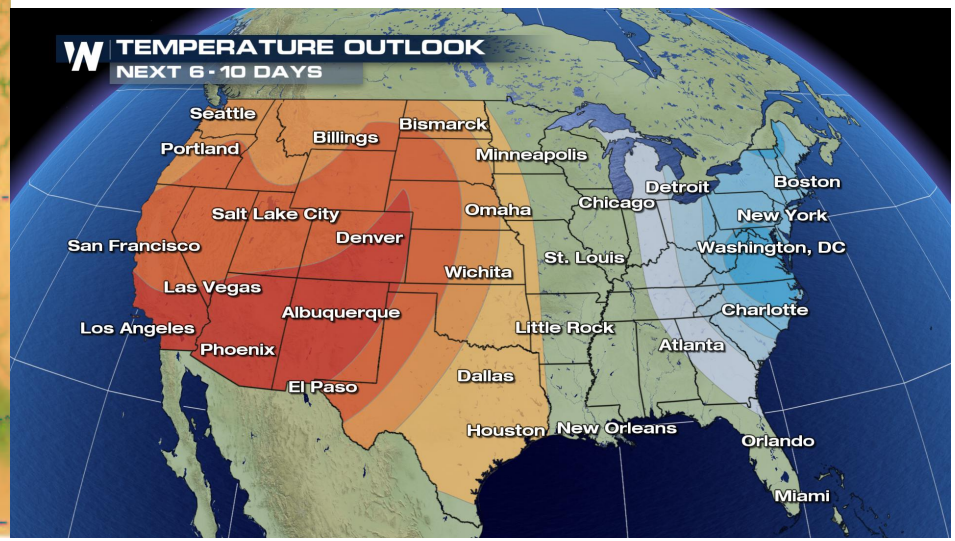
- strictly speaking, $X(t)$ is a function of starting point and time: write $X(t; y^0, t^0)$ instead.
- differential equation:

$$\frac{dX}{dt} = F(t)d(t) + V(X(t; y^0, t^0), t)$$

- current can help or hinder

Sensing and correlations in space

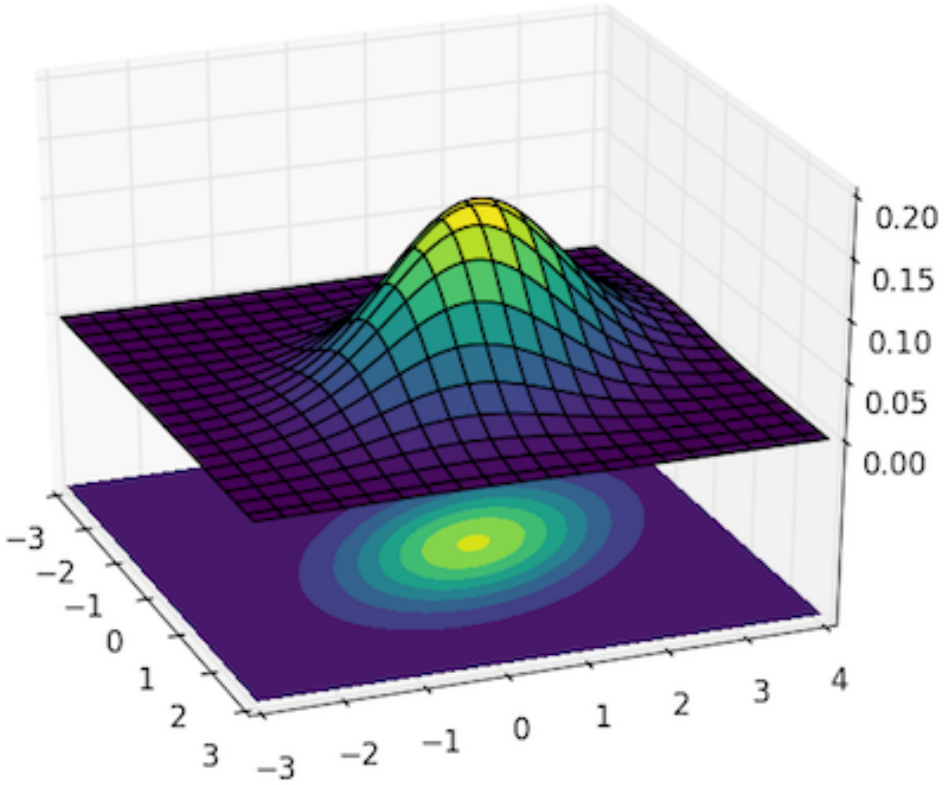




- Measure & model correlations in space?
- Estimate temperature / rainfall / gold in other locations?
- *Intuition: correlation is a function of distance*

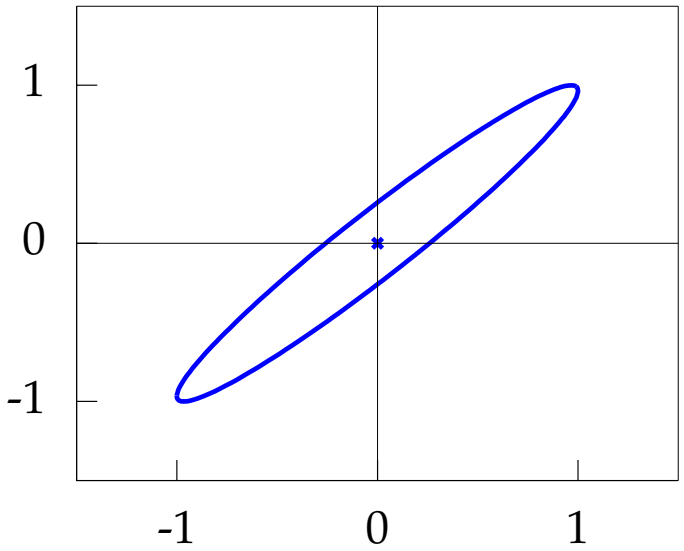
Multivariate Gaussian distribution

$$y \sim \mathcal{N}(\mu, \Sigma) \qquad p(y) = \frac{1}{(2\pi)^{-d/2}|\Sigma|^{1/2}} \exp\Big(-\tfrac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\Big)$$



$$\Sigma_{ij} = \mathrm{cov}(y_i, y_j)$$

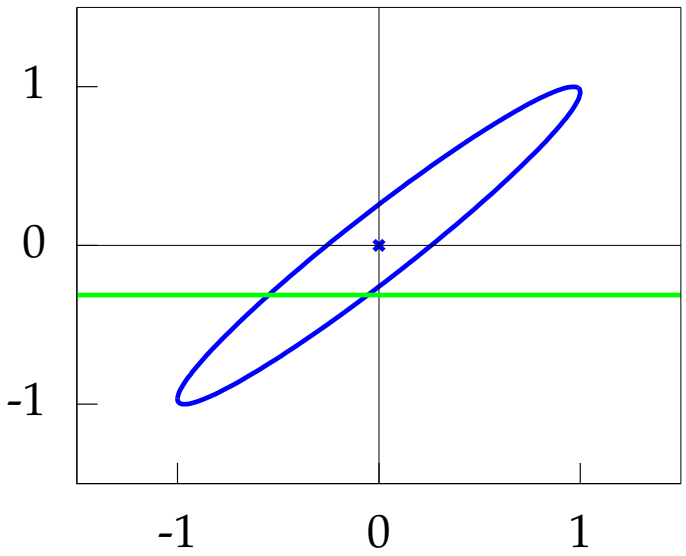
Covariance for 2 variables: intuition



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

Illustrations: Neil Lawrence

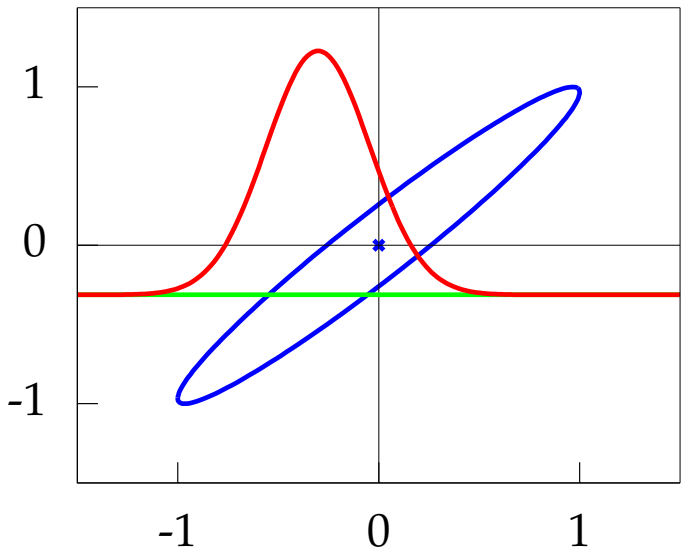
Covariance for 2 variables: intuition



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

Illustrations: Neil Lawrence

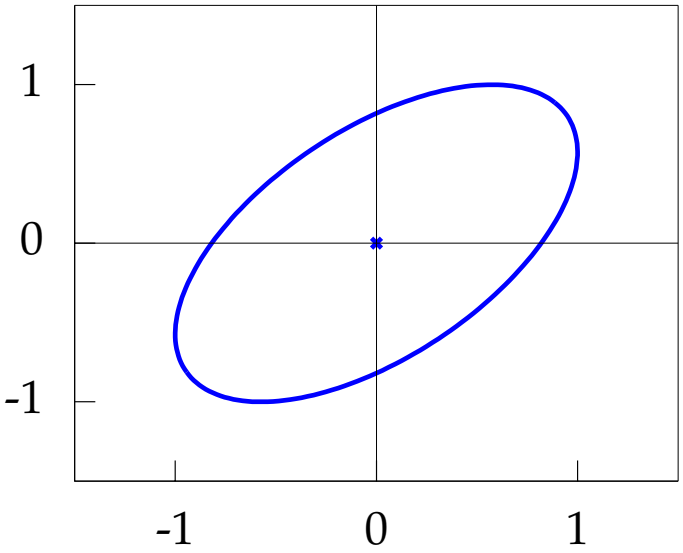
Covariance for 2 variables: intuition



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

Illustrations: Neil Lawrence

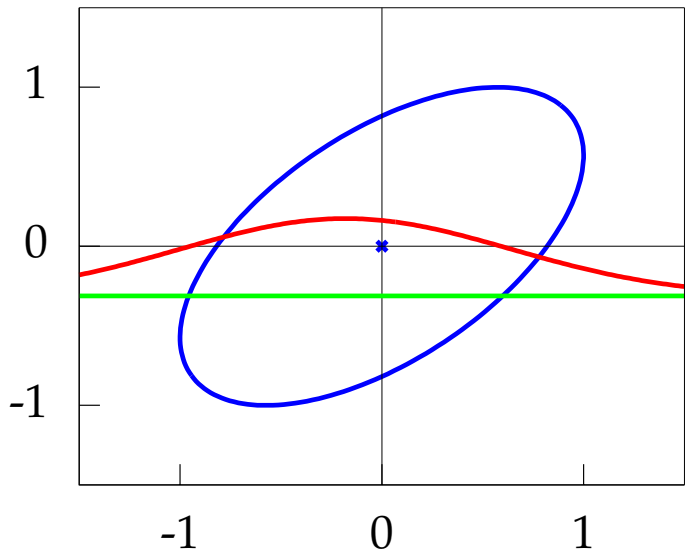
Covariance for 2 variables: intuition



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

Illustrations: Neil Lawrence

Covariance for 2 variables: intuition



$$\begin{bmatrix} 1 & 0.57375 \\ 0.57375 & 1 \end{bmatrix}$$

Illustrations: Neil Lawrence

Sensing and correlations in space

- Idea: model measurements Y_i at locations x_i as Gaussian, $1 \leq i \leq N$
- *covariance is a function of locations* (kernel function), e.g. RBF (squared exponential) kernel:

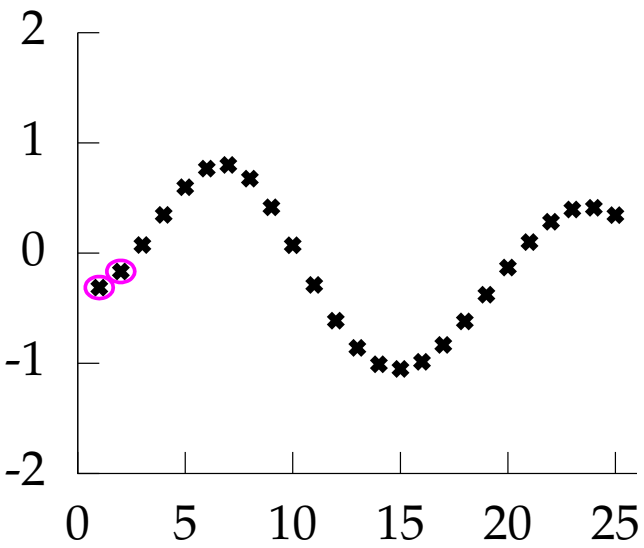
$$\mathrm{cov}(Y_i, Y_j) = k(x_i, x_j) = \exp\Big(-\frac{\|x_i - x_j\|^2}{2\ell^2}\Big)$$

- covariance: $\Sigma_N =$

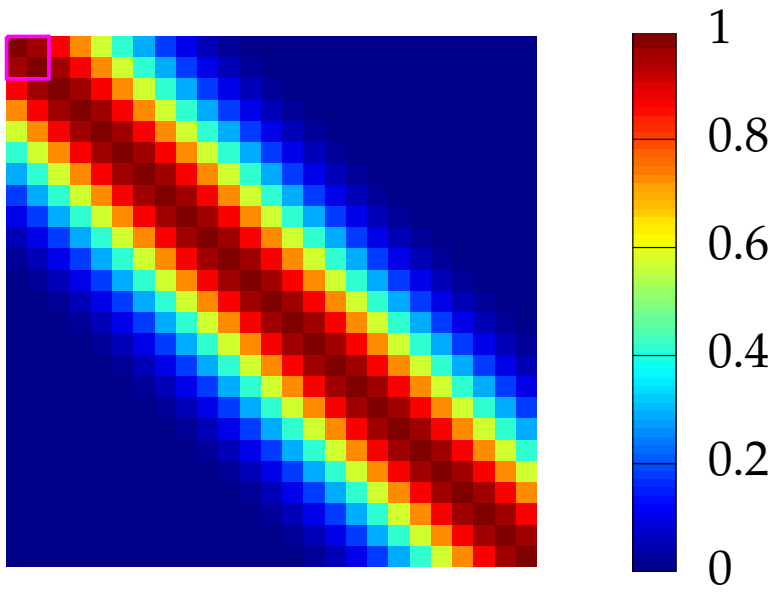
$$\begin{bmatrix} \mathrm{cov}(Y_1, Y_1) & \dots & \mathrm{cov}(Y_1, Y_N) \\ \mathrm{cov}(Y_2, Y_1) & \dots & \vdots \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_N, Y_1) & \dots & \mathrm{cov}(x_N, x_N) \end{bmatrix} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1) & \dots & & k(x_N, x_N) \end{bmatrix}$$

Covariance as a function of distance

- demo
- draw $(y_1, \ldots, y_{25})$ jointly Gaussian, with covariance decaying with distance in 1D.



(a) A 25 dimensional correlated random variable (values ploted against index)



(b) colormap showing correlations between dimensions.

Illustrations: Neil Lawrence

Covariance matrix and Gaussian Processes

- express covariance by *covariance function**

$$\text{cov}(y_i, y_j) = k(x_i, x_j)$$

  e.g., a function of $x_i - x_j$

- this generalizes to *all* points in our space!

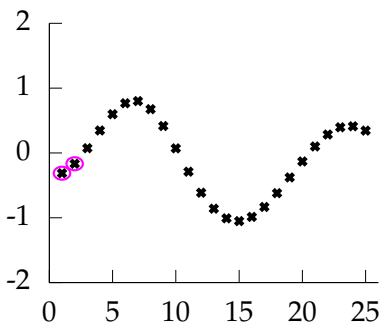
> A *Gaussian Process* is a collection of random variables, any finite number of which are Gaussian.

- GP is fully specified by mean and covariance functions

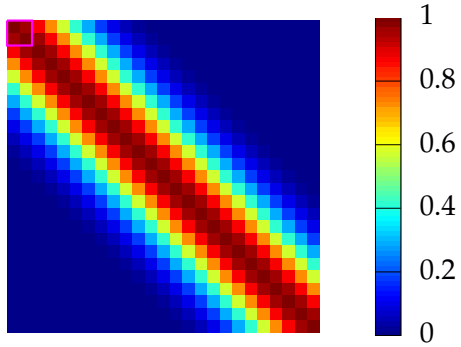
$$m(x) = \mathbb{E}[f(x)] \qquad k(x, x') = \text{cov}(f(x), f(x')).$$

- *smallprint: function $k(x_i, x_j)$ must generate a valid covariance matrix: symmetric, and resulting covariance matrix must be positive semidefinite

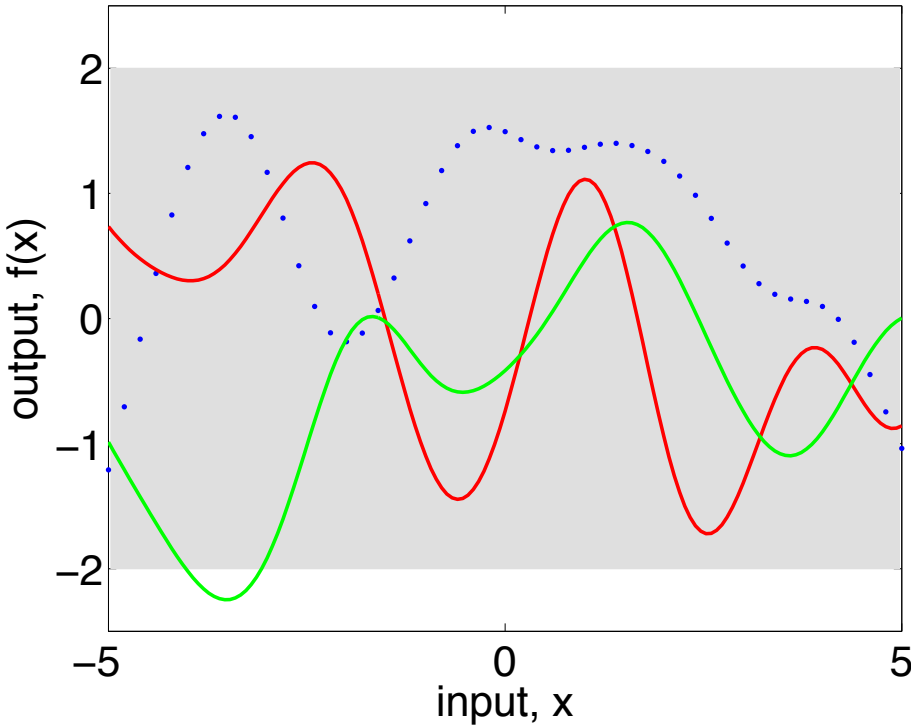
Distribution over functions



(a) A 25 dimensional correlated random variable (values ploted against index)



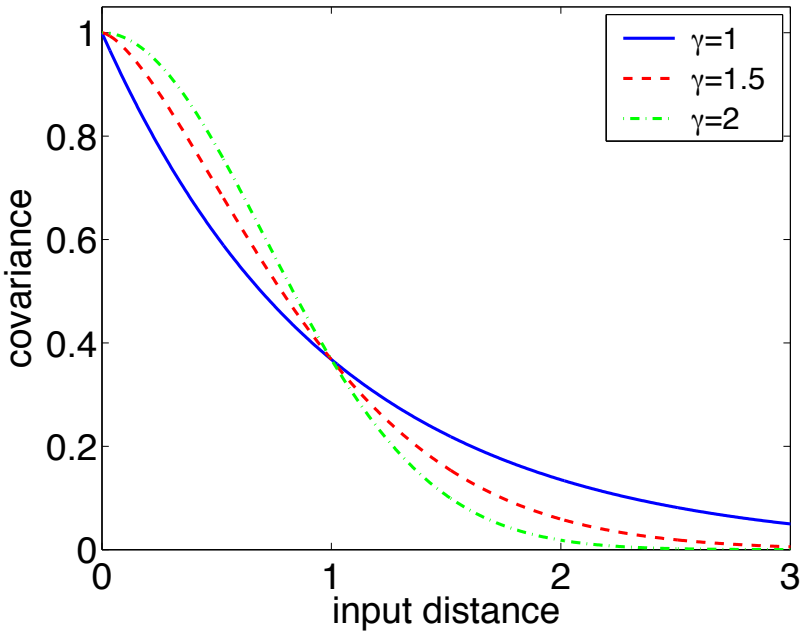
(b) colormap showing correlations between dimensions.

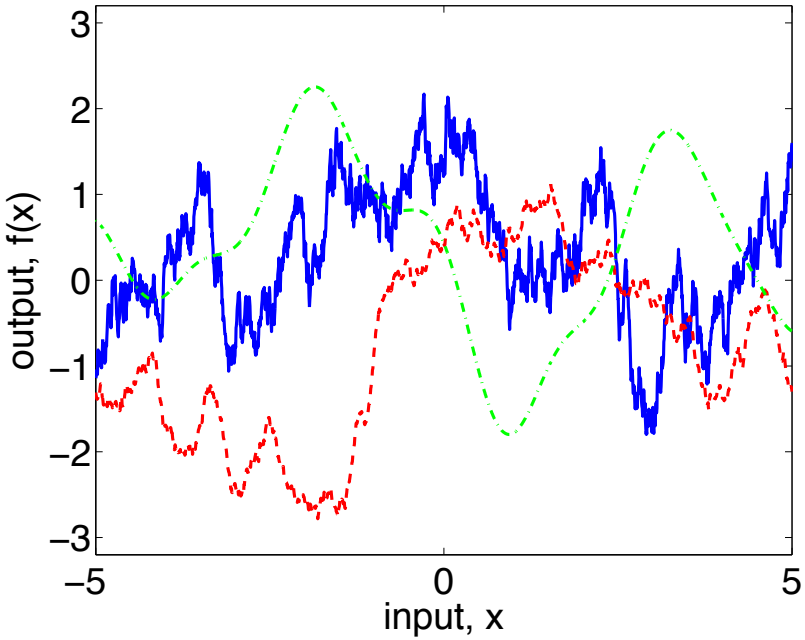


Influence of the covariance function

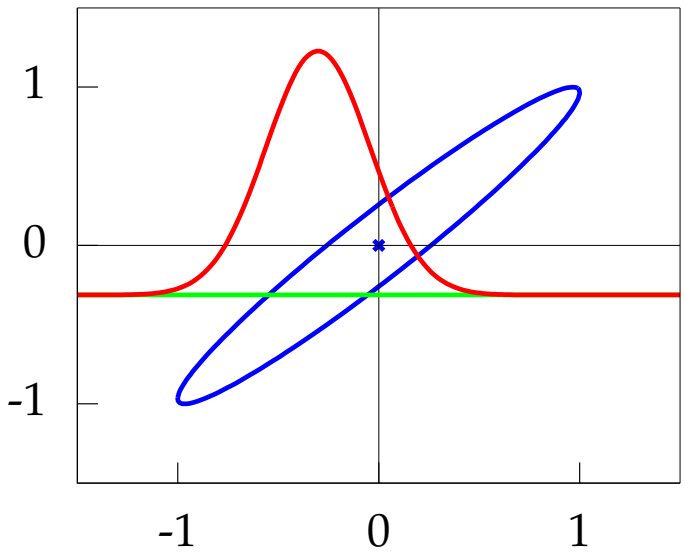
Example: $k(x, x') = \exp(-\left(\frac{\|x-x'\|}{\ell}\right)^{\gamma})$

(Gamma-exponential kernel) What happens as we vary γ?





Spatial Prediction: intuition



$$\begin{bmatrix} 1 & 0.96587 \\ 0.96587 & 1 \end{bmatrix}$$

use conditioning: $p(y_A|y_B)$

Prediction: conditional probabilities

- Y_A, Y_B Gaussian random variables. We observe $Y_B = y_B$.

$$\begin{bmatrix} Y_A \\ Y_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{bmatrix} \right)$$

- *Conditioning:* $p(Y_A | Y_B = y_B)$ is also Gaussian with mean and variance

$$\mu_{A|B} = \mu_A + \frac{\sigma_{AB}}{\sigma_B^2}(y_B - \mu_B)$$

$$\sigma^2_{A|B} = \sigma_A^2 - \sigma_{AB}\sigma_B^{-2}\sigma_{AB}.$$

References and Further Reading

- H. Kaper, H. Engler. *Mathematics & Climate*. Chapters 1, 8.
- W. Menke, J. Menke. *Environmental Data Analysis with Matlab*
- `http://www.climateinformatics.org/`
- C. Rasmussen, S. Williams. Gaussian Processes for Machine Learning. `http://www.gaussianprocess.org/gpml/`, Chapters 2.2, 2.3, 4.2, 5.4
- More references, tutorials, code: `http://www.gaussianprocess.org/`