Data Analysis: Statistical Modeling and Computation in Applications

Spatial and Environmental Data:
Introduction, Local Correlations

Overview

- Environmental data
- Modeling flows
- Short-range spatial correlations
  - intuition
  - 2 variables
  - multiple variables

Environmental Data – Examples

- Air Quality, Water Quality (pollutants)

Environmental Data – Examples

- Air Quality, Water Quality (pollutants)
- Weather & climate (temperature, winds, moisture, precipitation, extreme conditions, . . .)
- Storms

Environmental Data – Examples

- Air Quality, Water Quality (pollutants)
- Weather & climate (temperature, winds, moisture, precipitation, extreme conditions, . . .)
- Storms
- Ocean dynamics

Environmental Data – Examples

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

Environmental Data – Examples

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

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

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")

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

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

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, ...)

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, ...)
- resource/energy management, e.g., water, renewable energies

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, ...)
- resource/energy management, e.g., water, renewable energies
- use in planning, routing, backtracking, control (ships, airplanes, ...)

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, ...)
- resource/energy management, e.g., water, renewable energies
- use in planning, routing, backtracking, control (ships, airplanes, ...)

Questions

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, ...)
- resource/energy management, e.g., water, renewable energies
- use in planning, routing, backtracking, control (ships, airplanes, ...)

Questions

- relationships (correlations, association)
- trends; forecasting

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, ...)
- resource/energy management, e.g., water, 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?

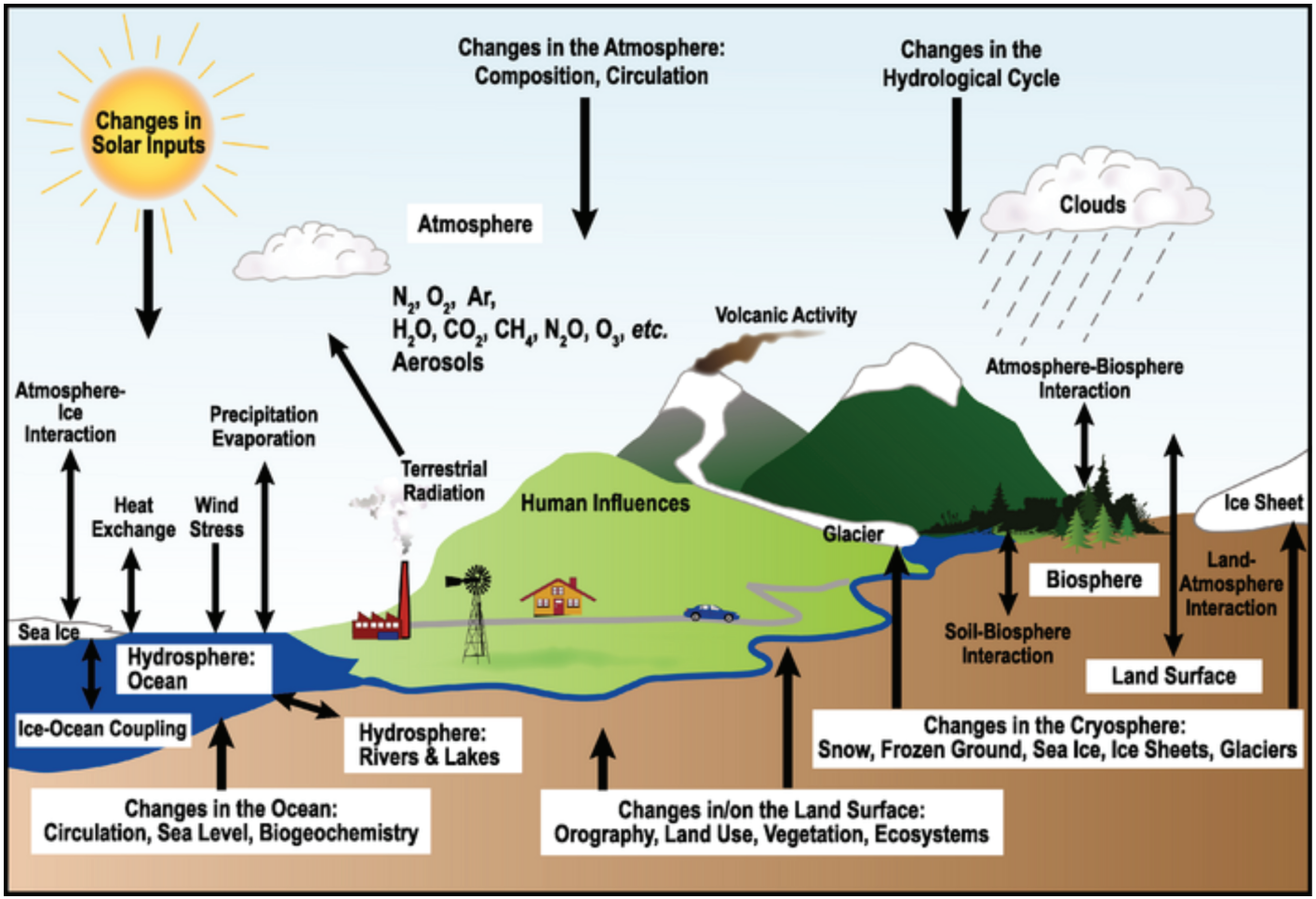
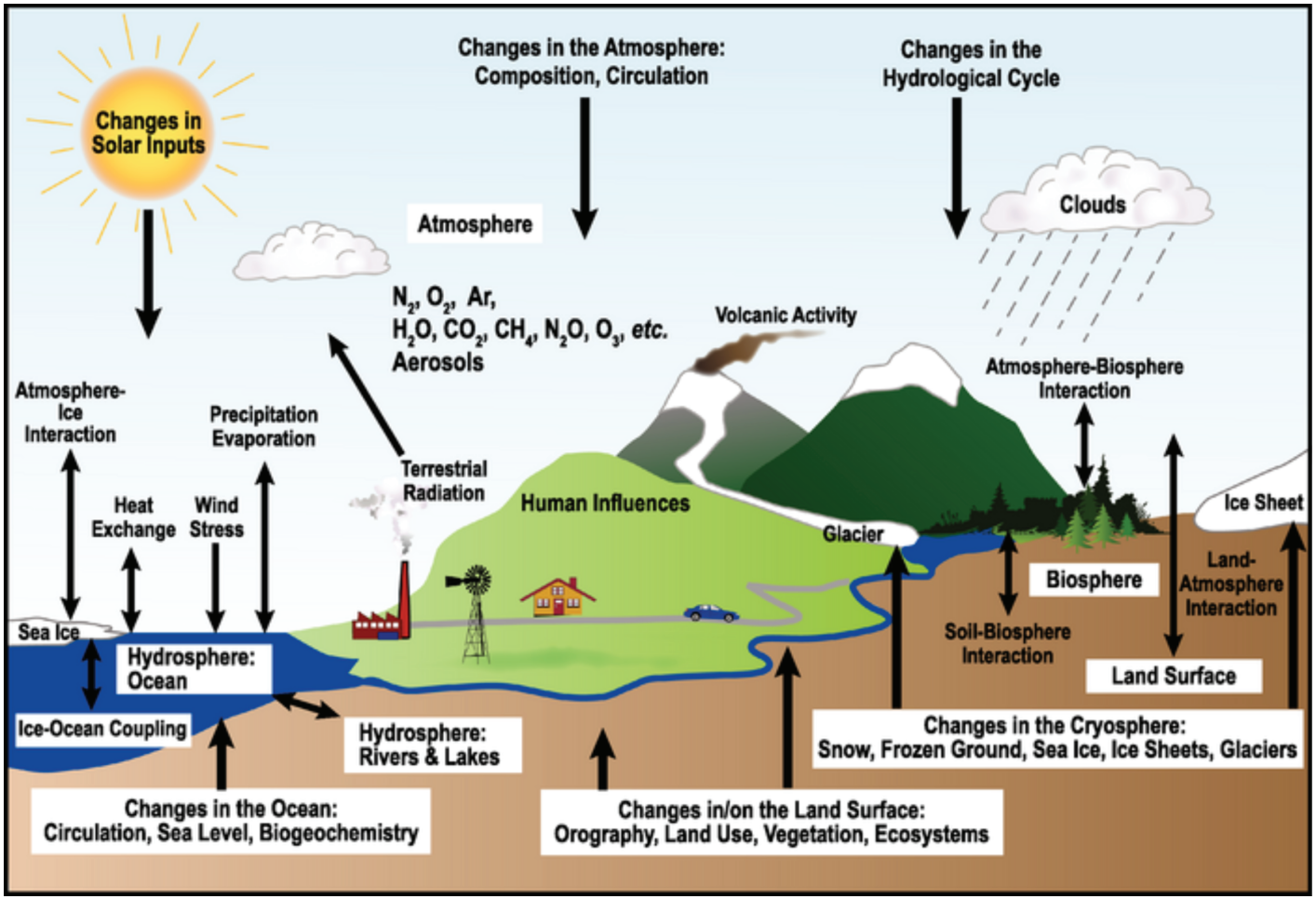


Image source: H. Kaper, H. Engler. Mathematics & Climate.

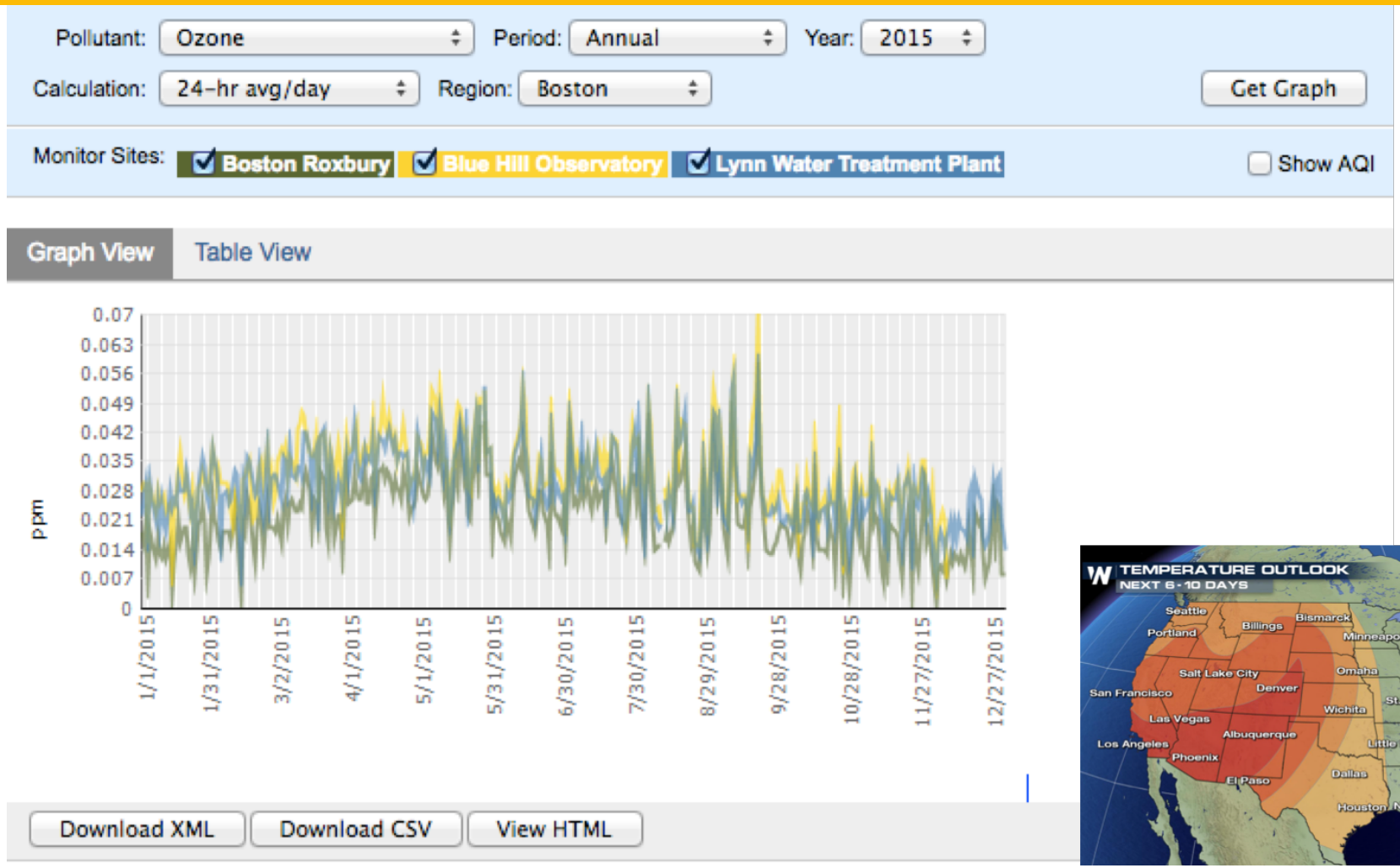
Environmental Data – What is special?



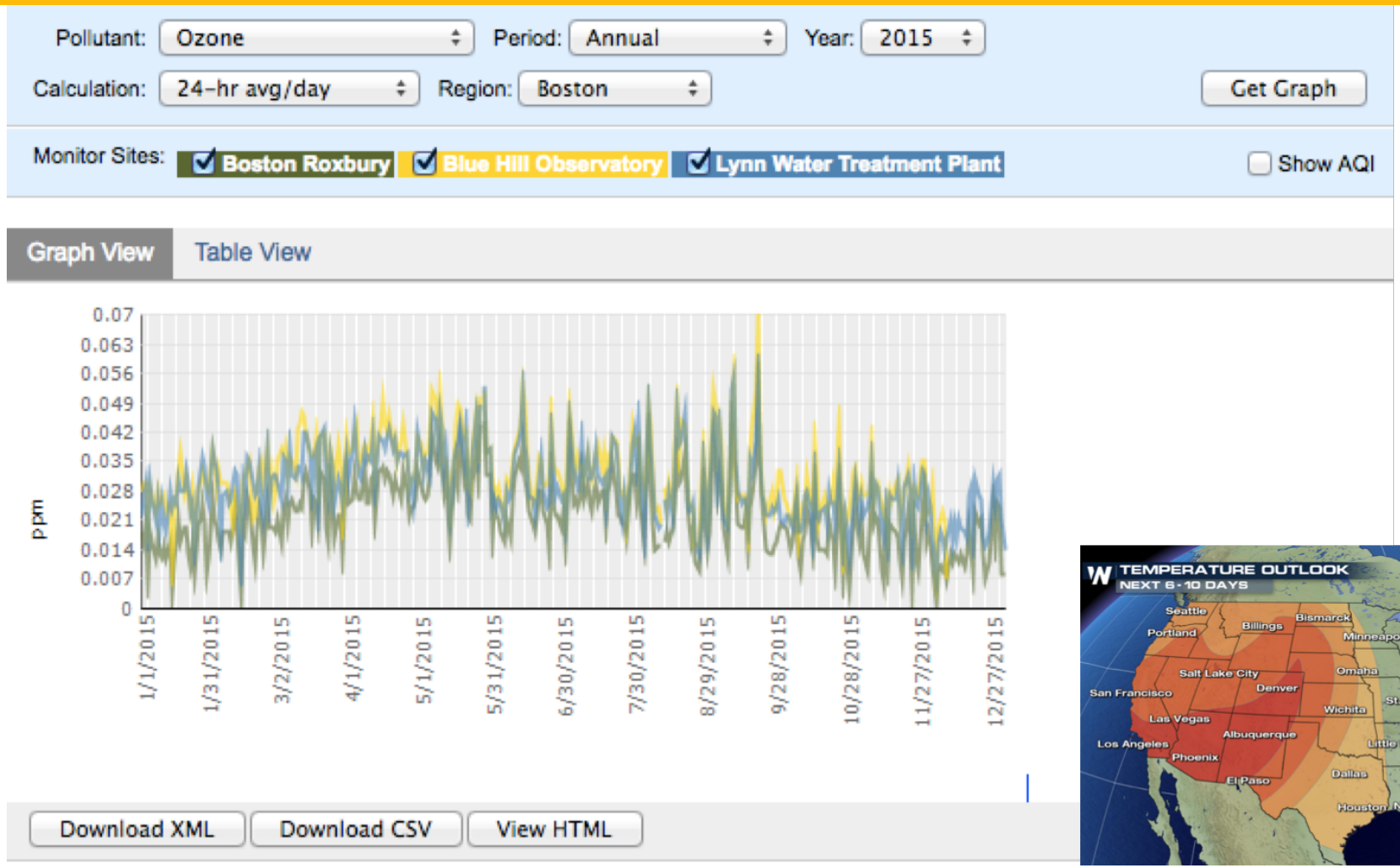
underlying physical processes; scientific models

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

Environmental Data – What is special?



Environmental Data – What is special?



spatial and temporal 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?

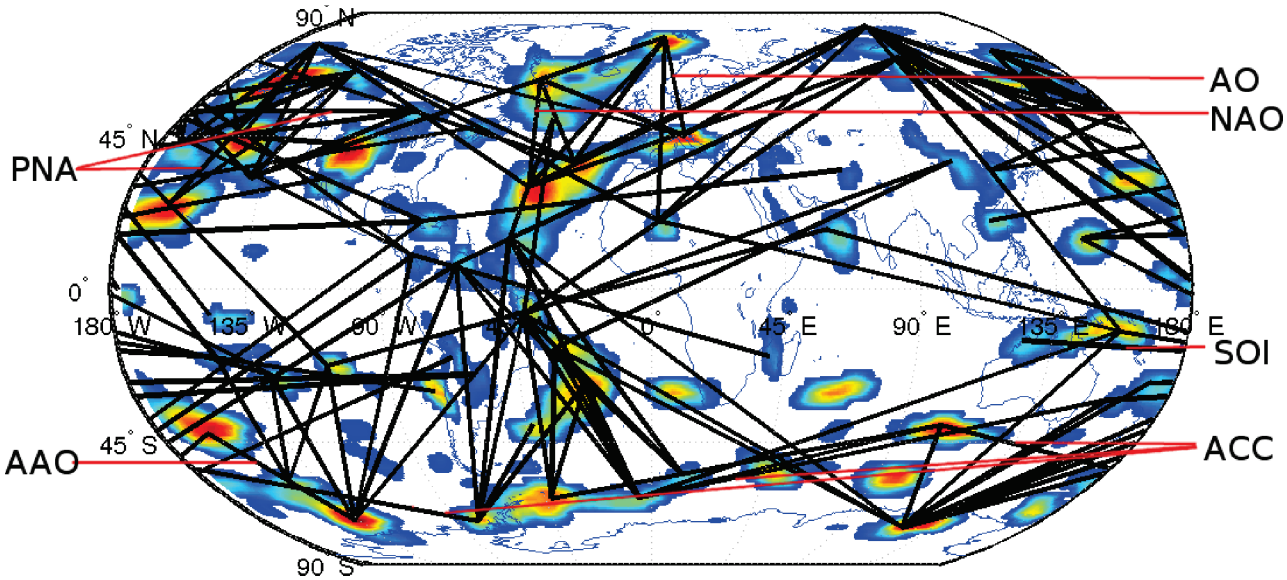


Figure 1.8. *Dipoles in NCEP sea-level data for the period* 1948–1967. *The color background shows the regions of high activity. The edges represent dipole connections between regions.*

Environmental Data – What is special?

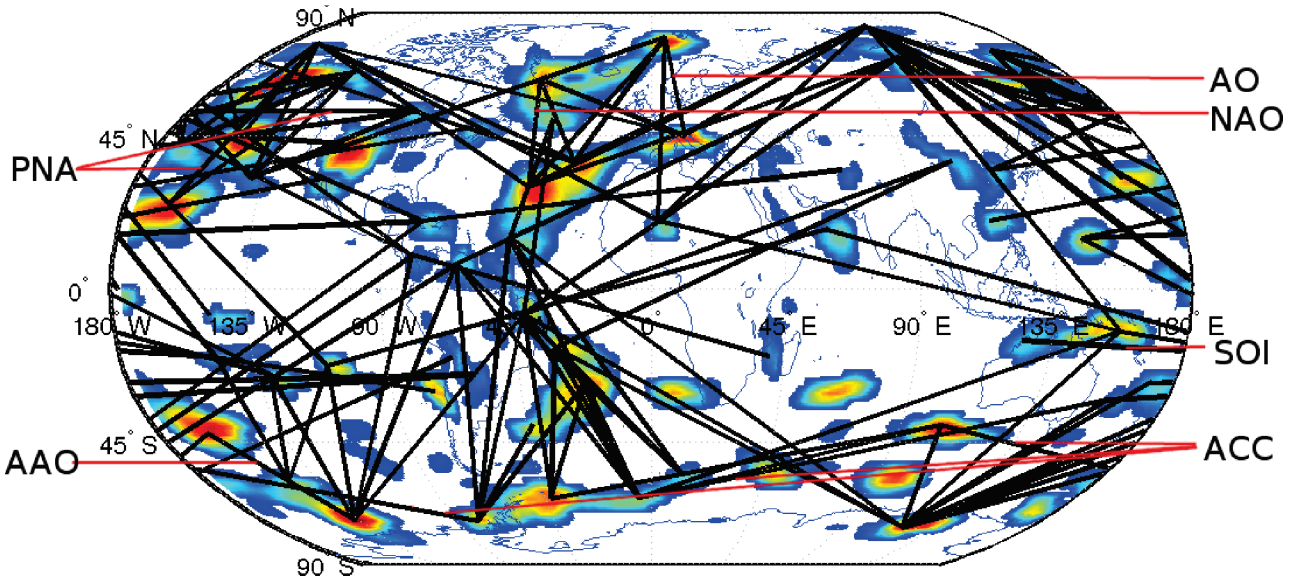


Figure 1.8. *Dipoles in NCEP sea-level data for the period* 1948–1967. *The color background shows the regions of high activity. The edges represent dipole connections between regions.*

long-range spatial (anti-)correlations

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

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

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

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

Environmental Data – What is special?

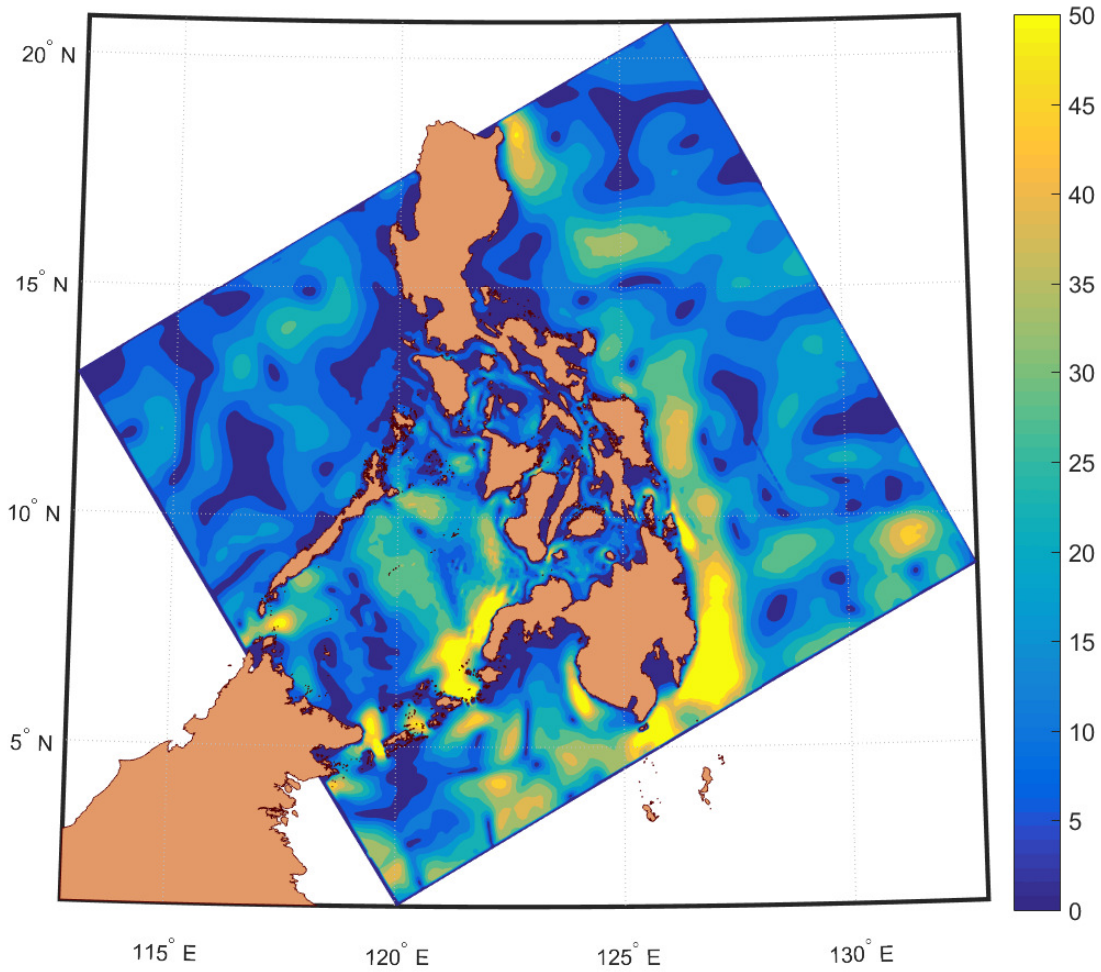
- correlations in time
- correlations in space (fields)
- scientific models + statistics; simulations
- methodological challenges for statistics
  - no controlled studies, only observational data
  - hypotheses from data
  - large data sets
  - heterogeneous data

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

Overview

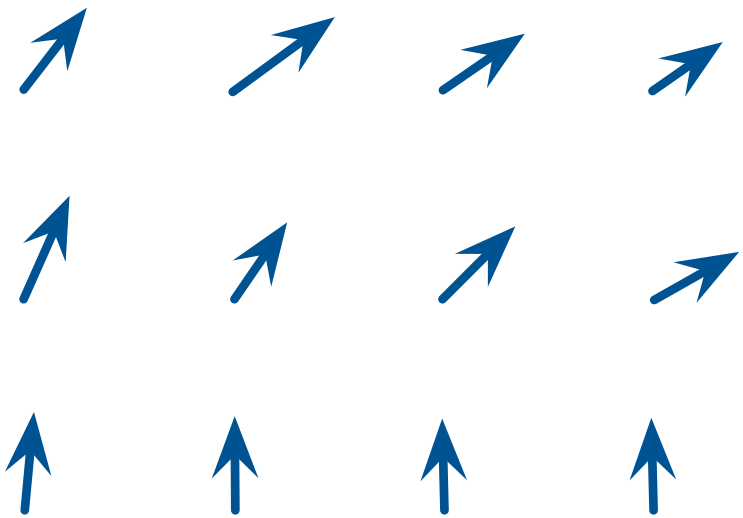
- Environmental data
- Modeling flows
- Short-range spatial correlations
  - intuition
  - 2 variables
  - multiple variables



Data for the problem set: flow in space & time

Flow fields

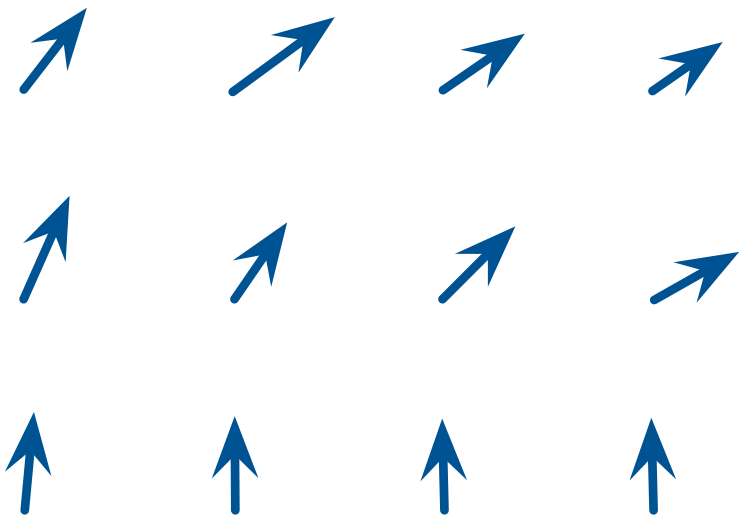
- $V(x, t)$: flow at location x at time t
  vector in 2d or 3d



$V(x, t)$ at a fixed time t

Flow fields

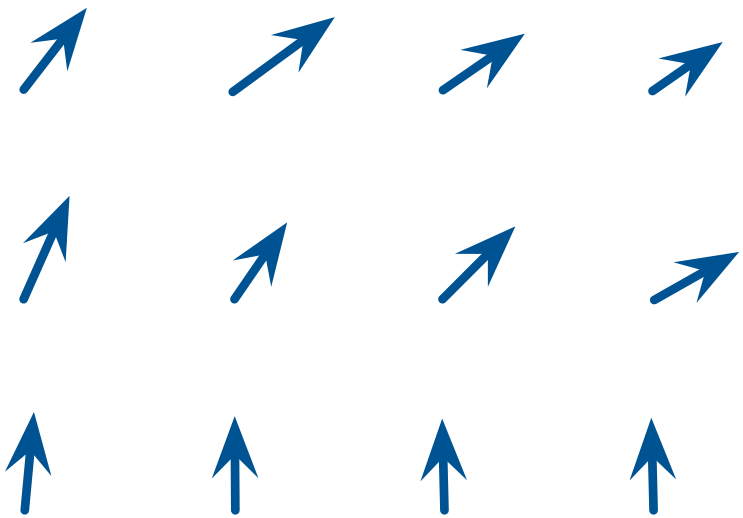
- $V(x, t)$: flow at location x at time t
  vector in 2d or 3d
- velocity: vector length



$V(x, t)$ at a fixed time t

Flow fields

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

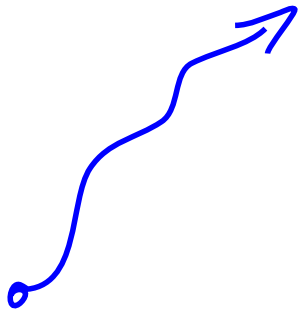


$V(x, t)$ at a fixed time t

What does variation in time mean?

Working with flow fields

- forward prediction

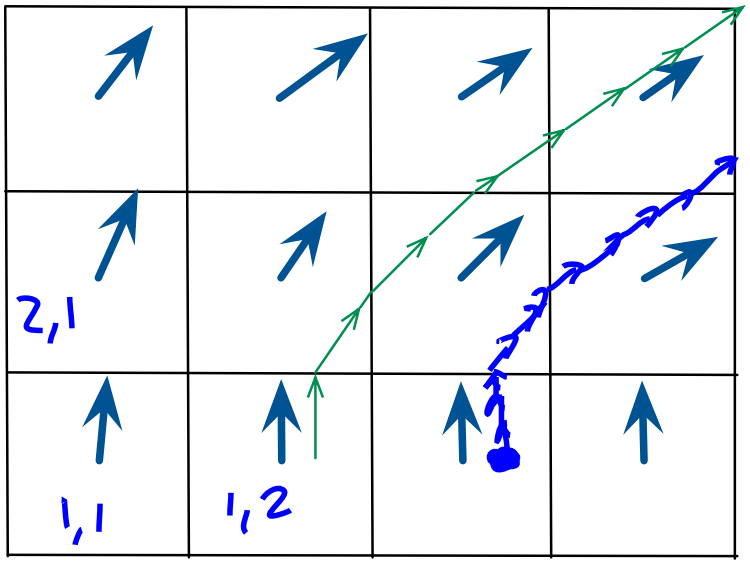


Working with flow fields

- forward prediction

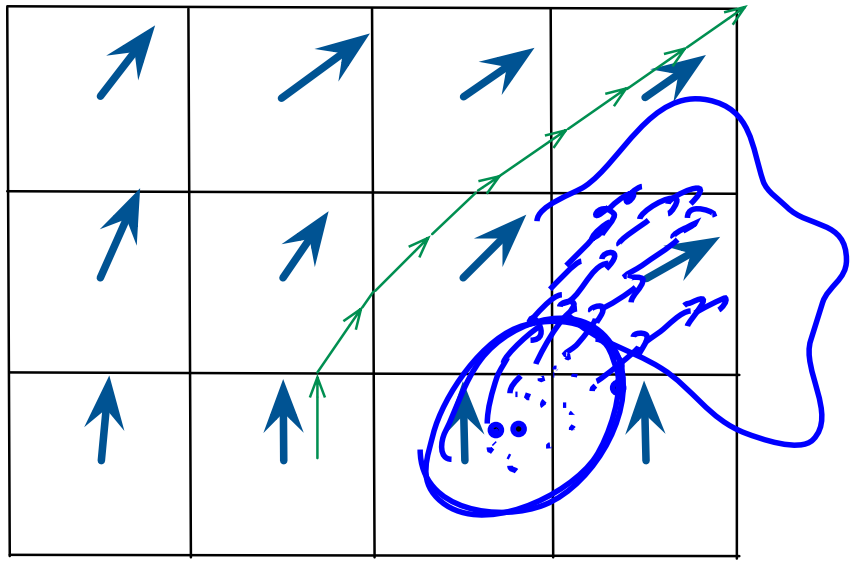
$$\frac{dx}{dt} = V(X(t), t)$$

$$X(t+\epsilon) = X(t) + \epsilon V(X(t), t)$$



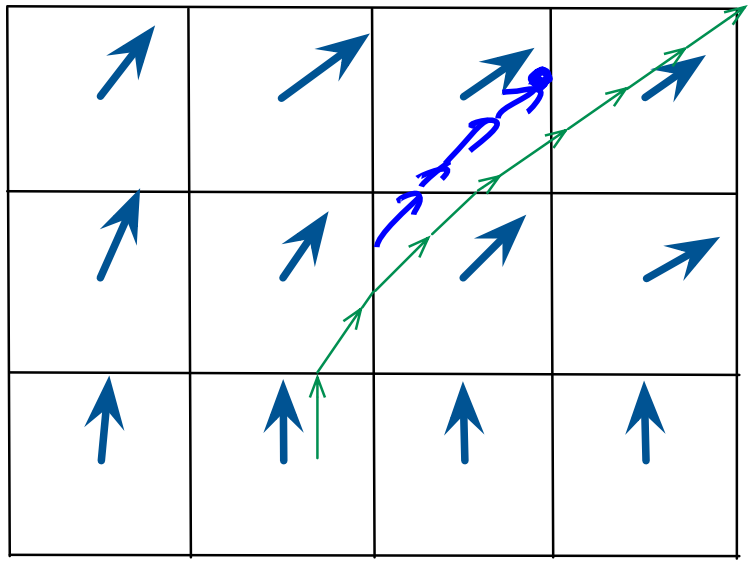
Working with flow fields

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

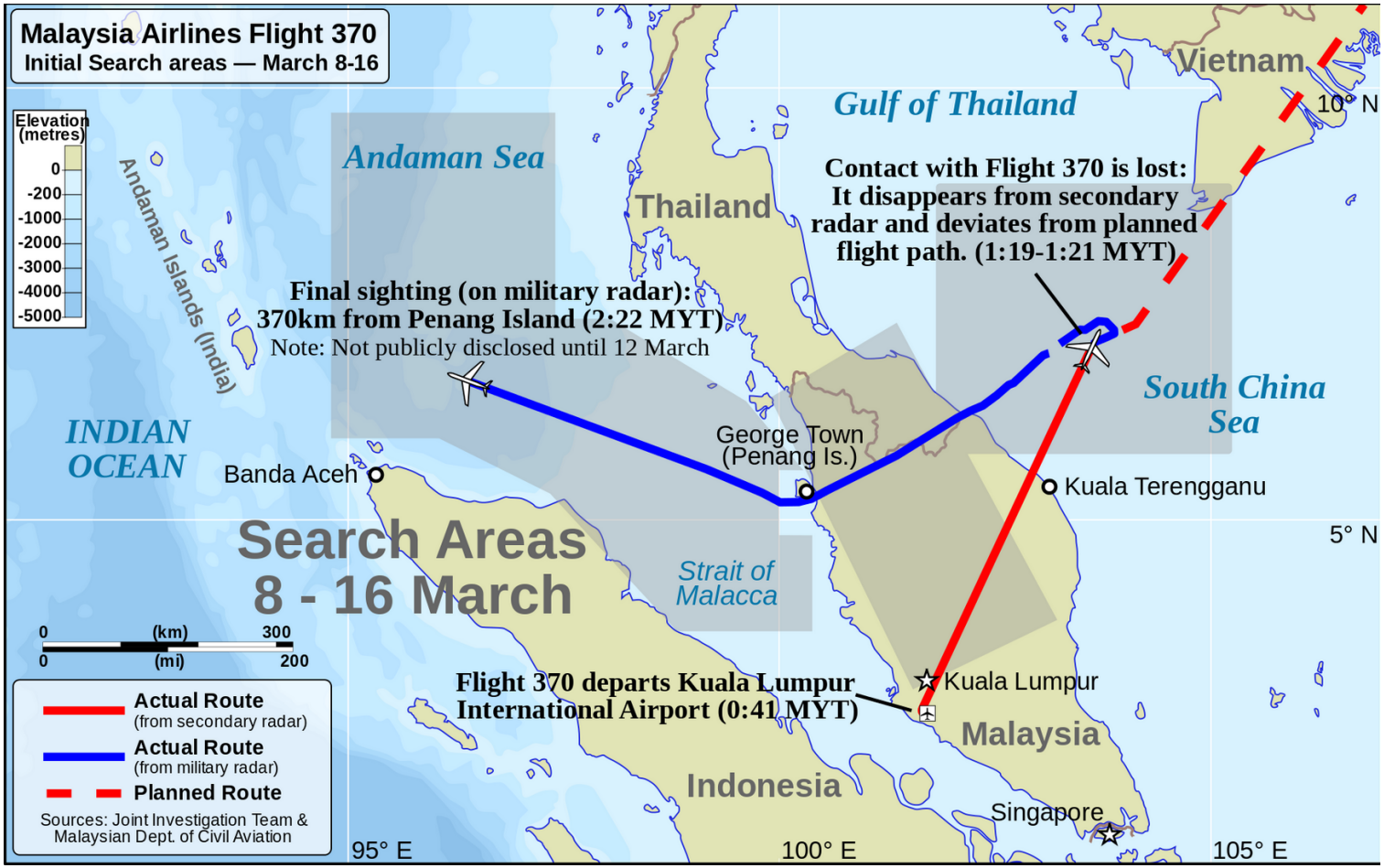


Working with flow fields

- 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

Airplane pieces found in multiple locations.

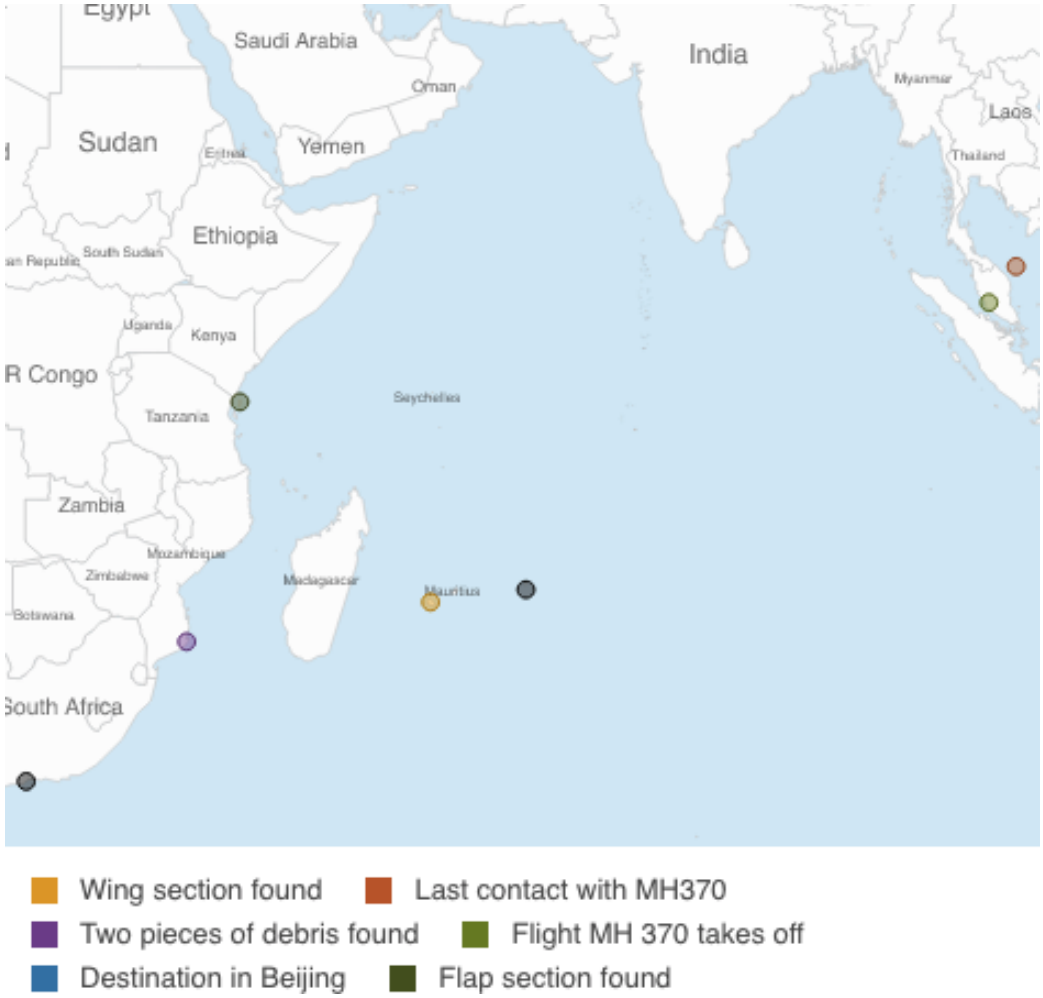
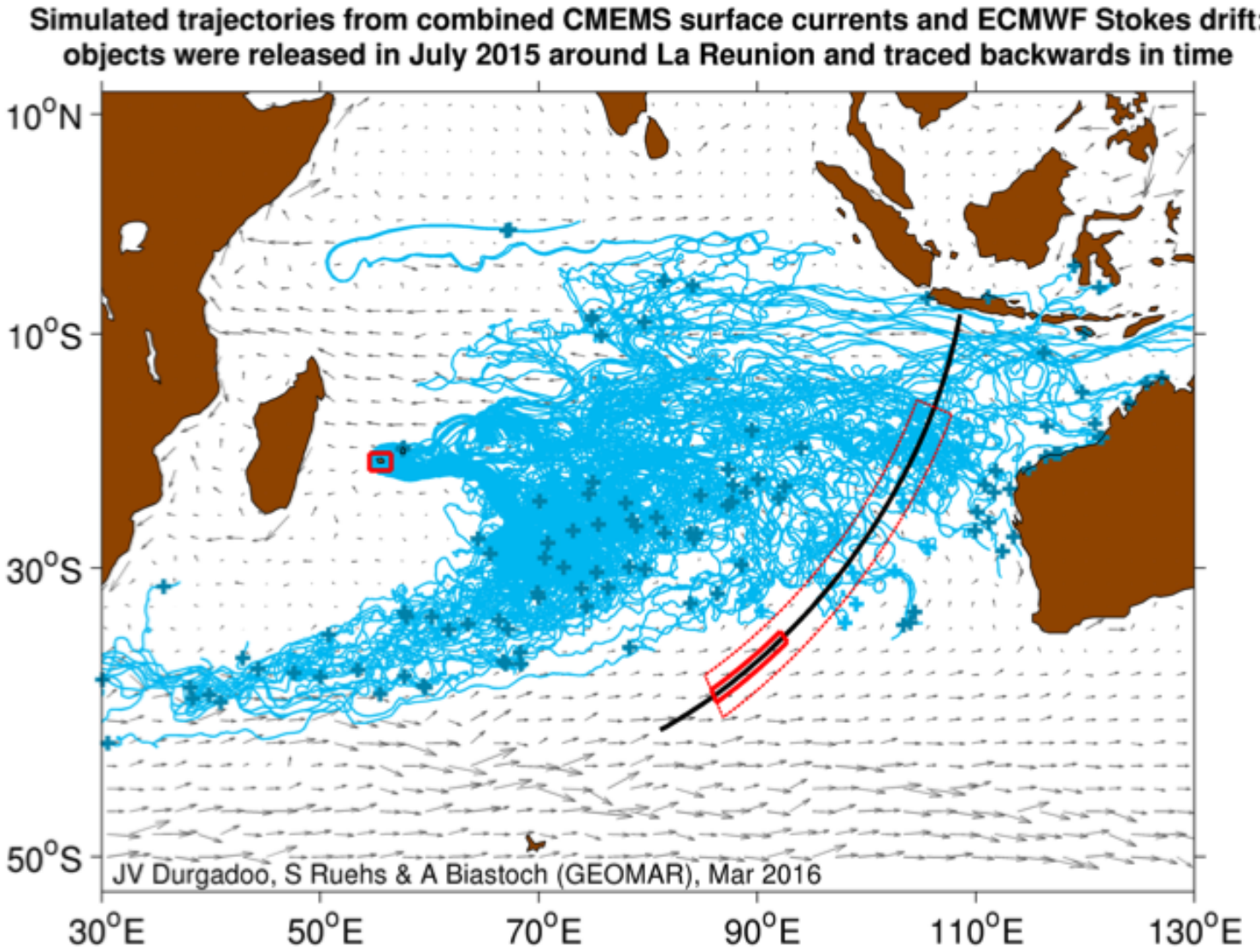
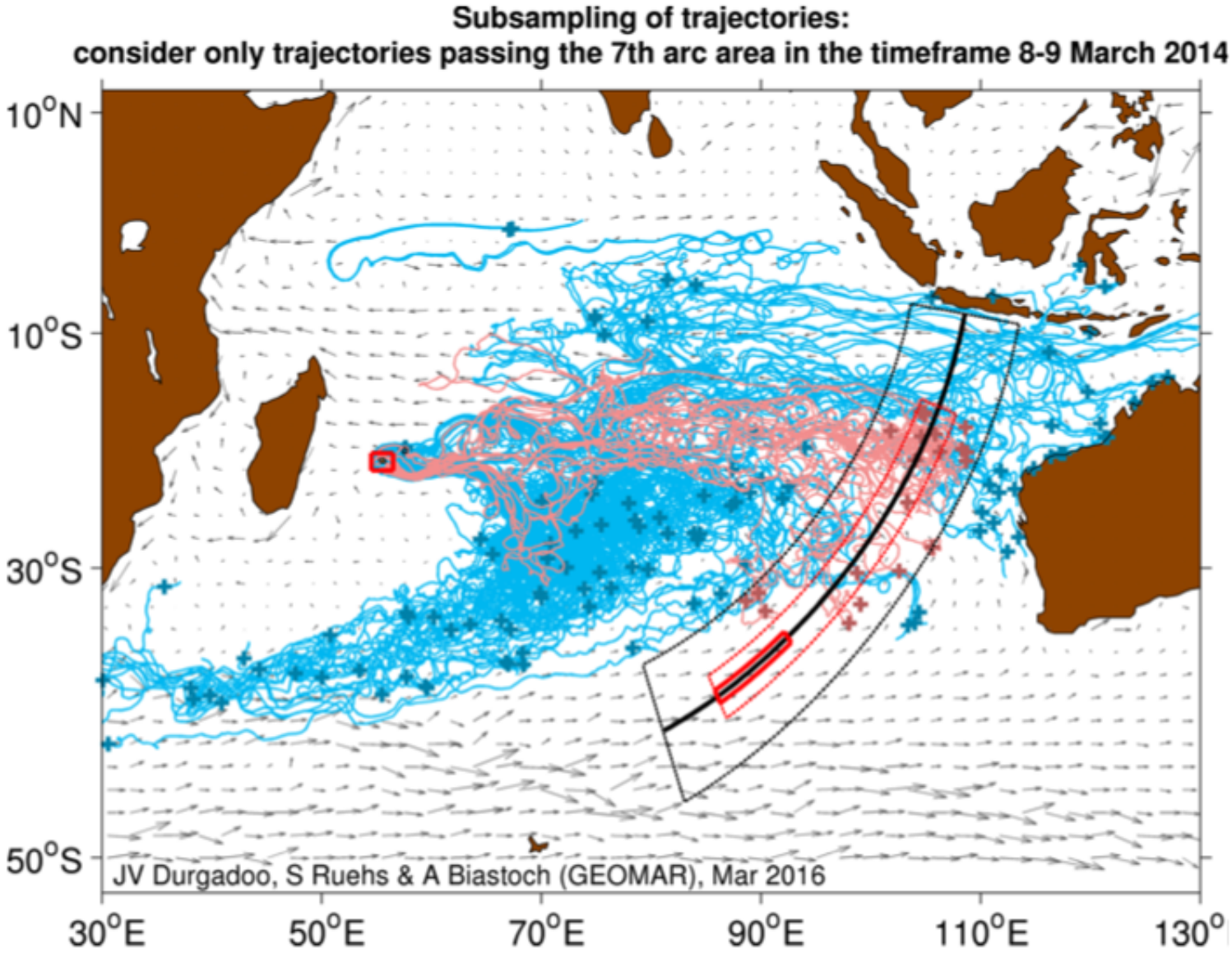


image source: BBC



full report:

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



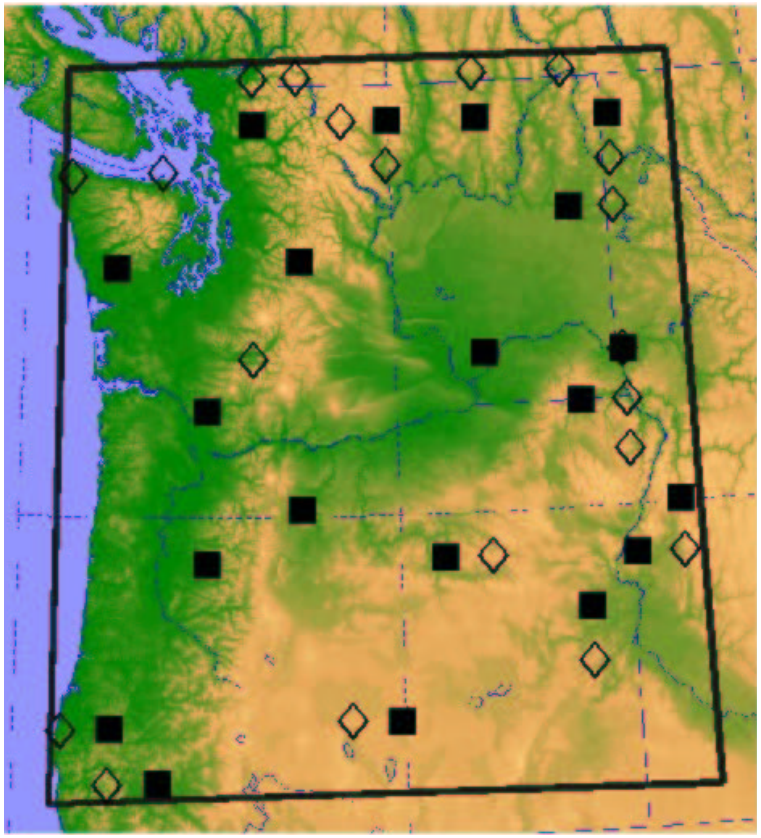
full report:

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

Overview

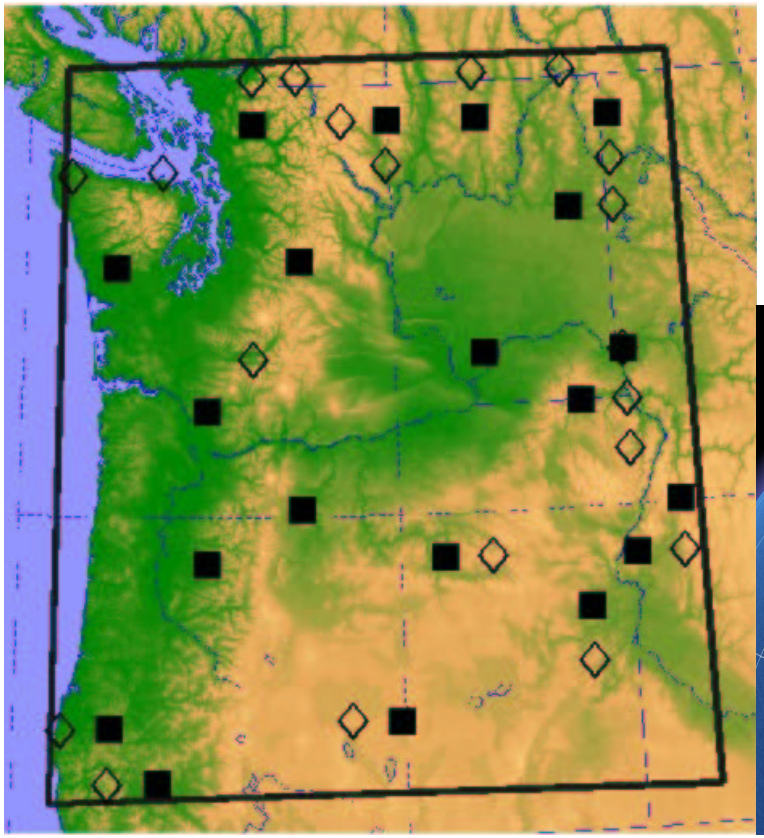
- Environmental data
- Modeling flows
- Short-range spatial correlations
  - intuition
  - 2 variables
  - multiple variables

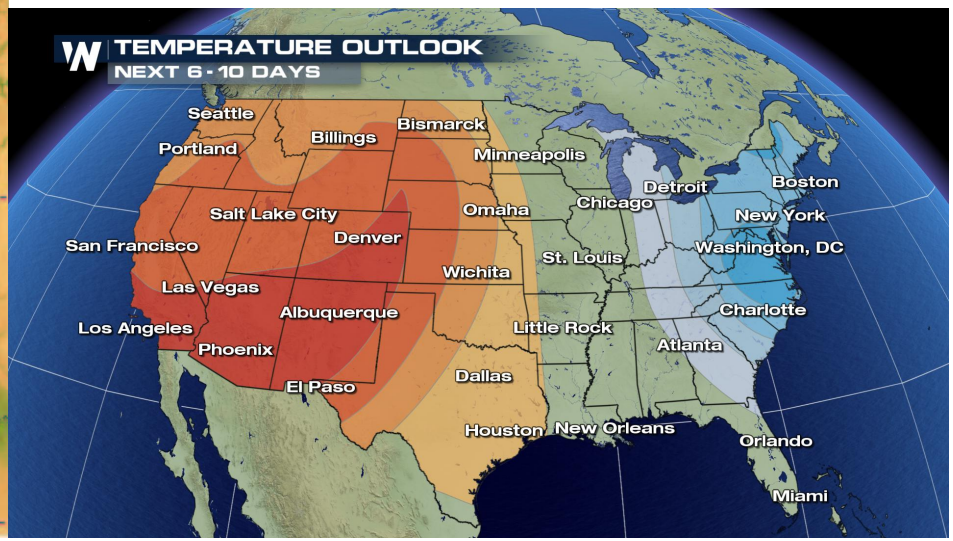
Sensing and correlations in space



- Measure & model correlations in space?
- Estimate temperature / rainfall / gold in other locations?

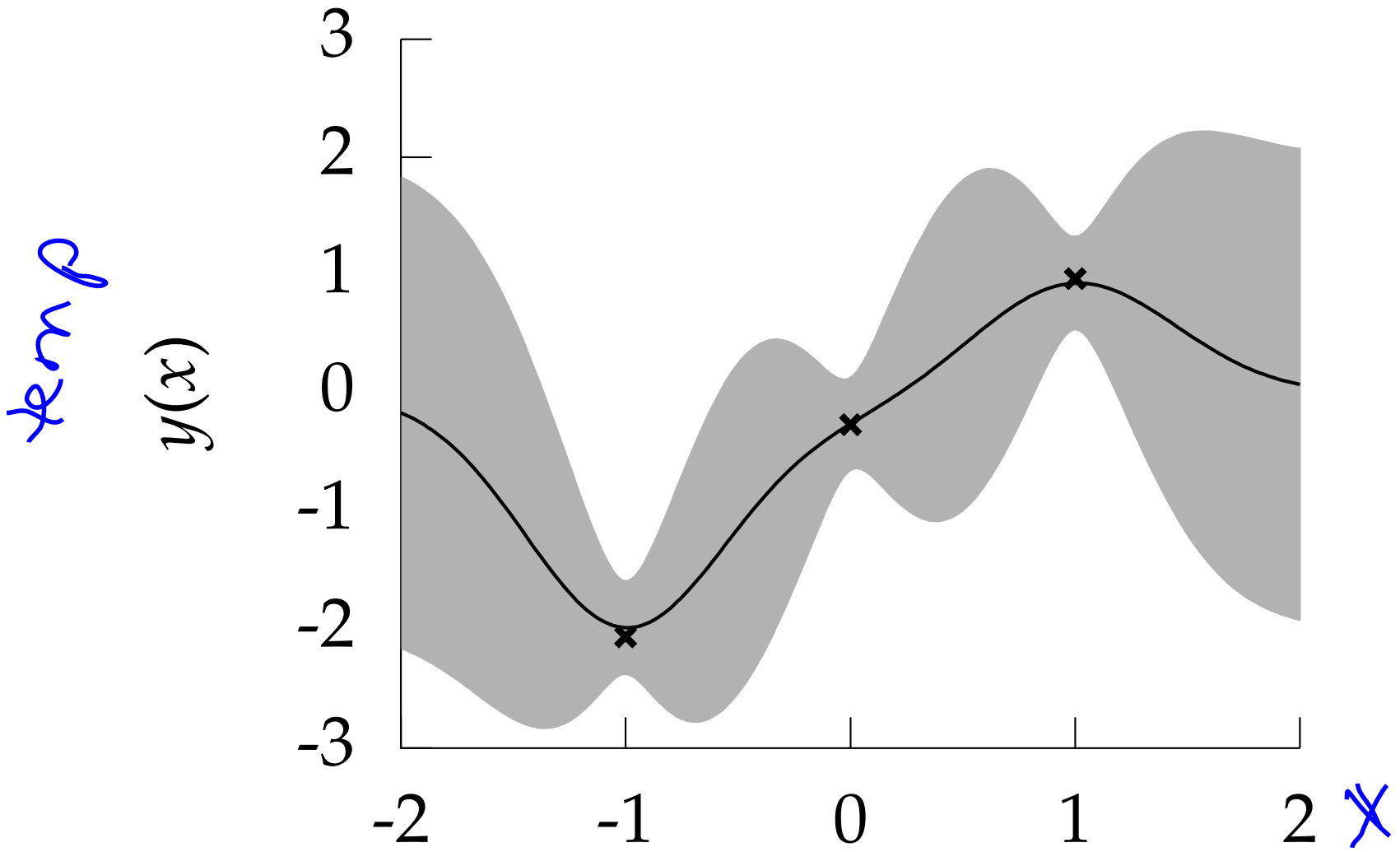
Sensing and correlations in space



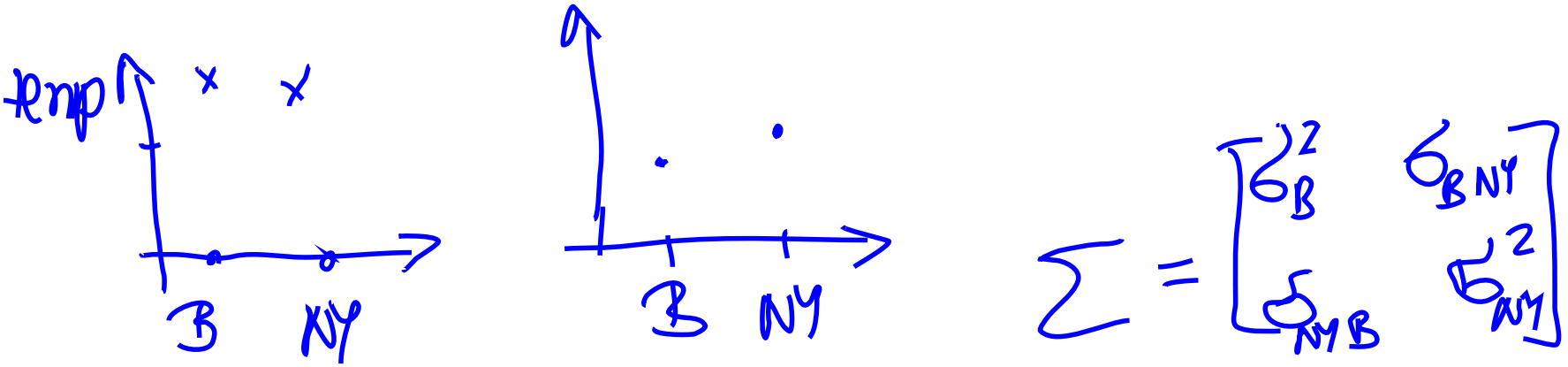


- Measure & model correlations in space?
- Estimate temperature / rainfall / gold in other locations?
- *Intuition: correlation is a function of distance*

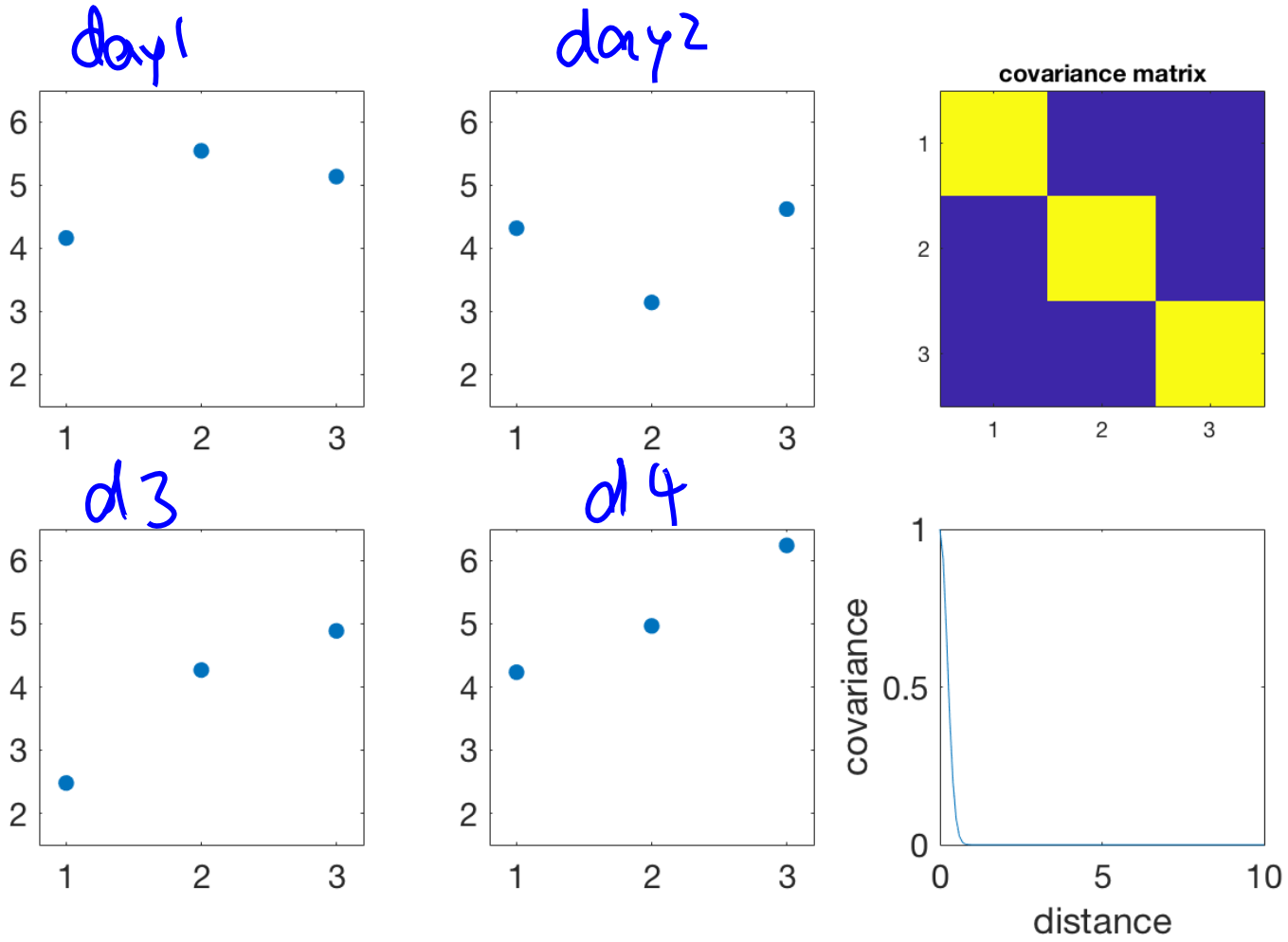
Where we are headed



A simple start

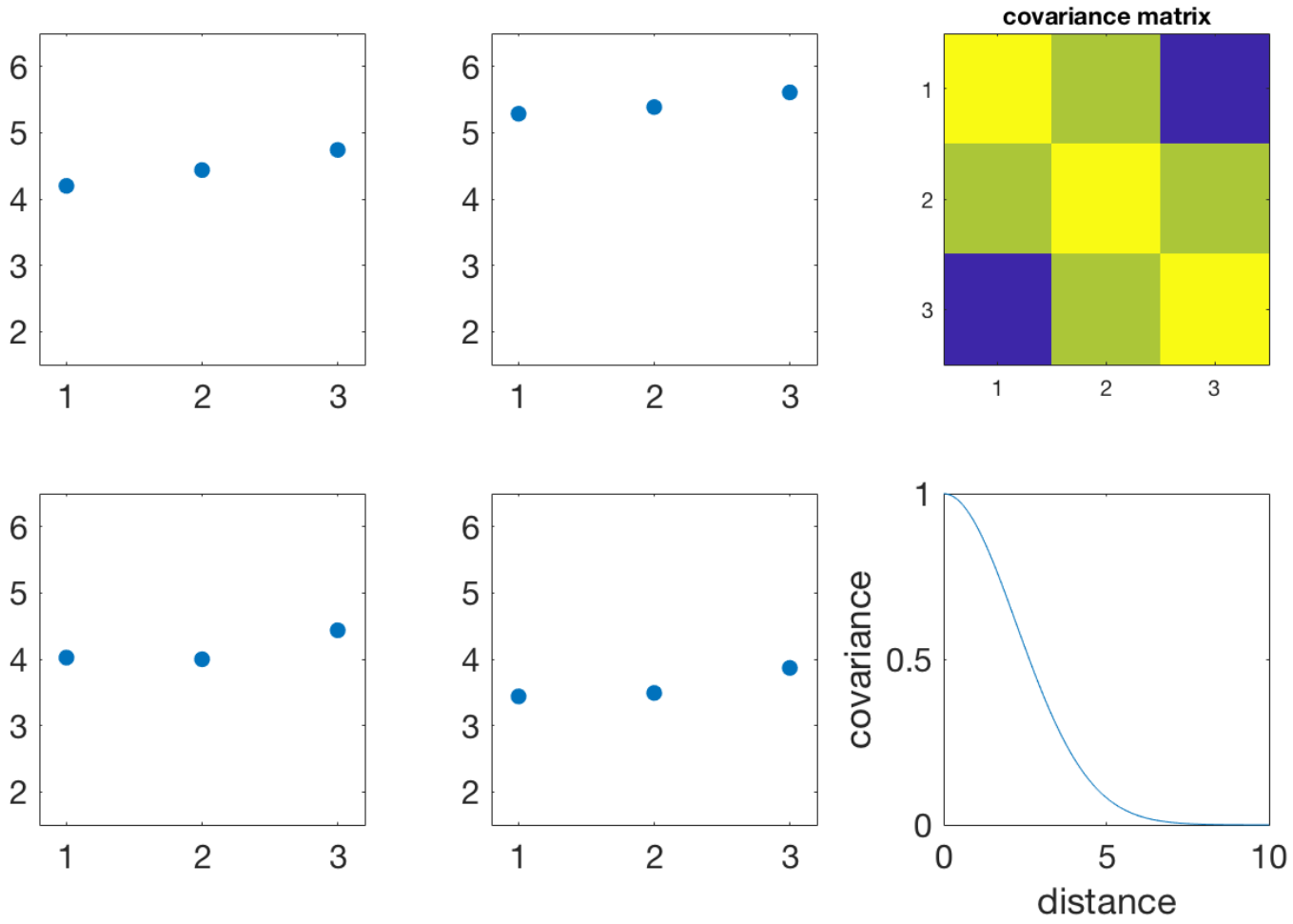


Intuition: correlated (Gaussian) random variables



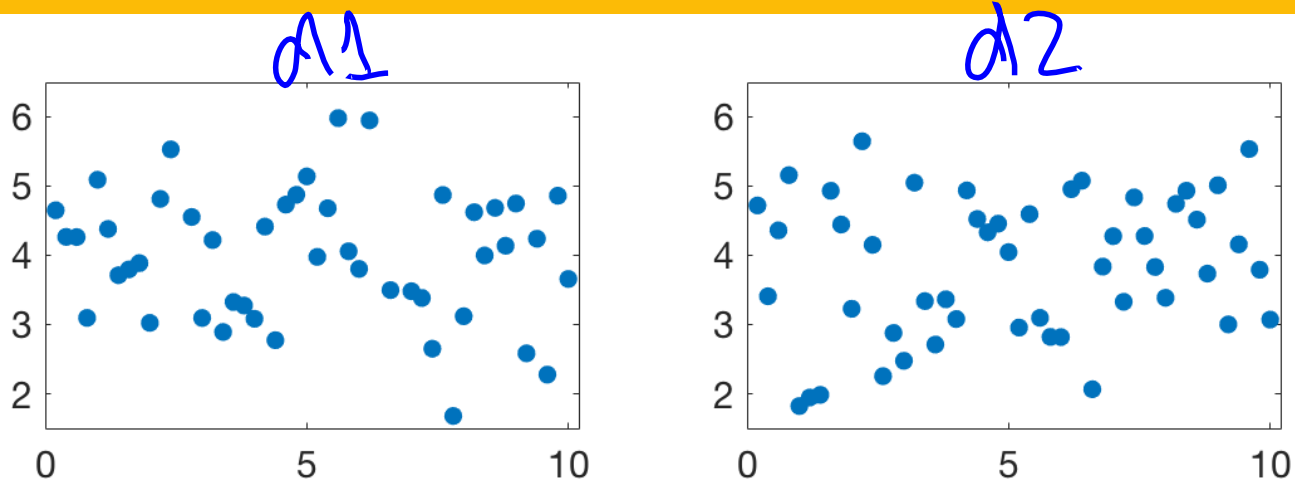
weaker correlation

Intuition: correlated (Gaussian) random variables

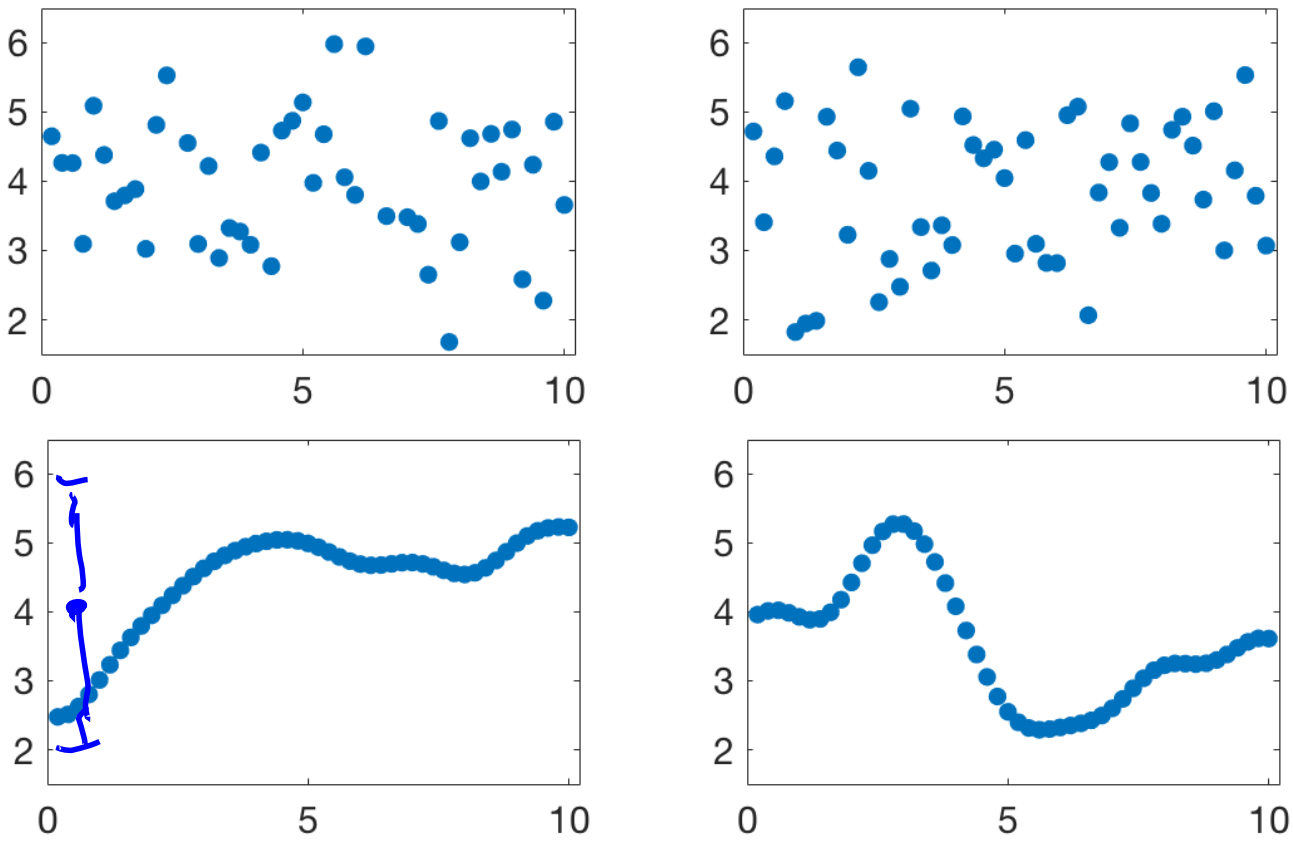


strong correlation

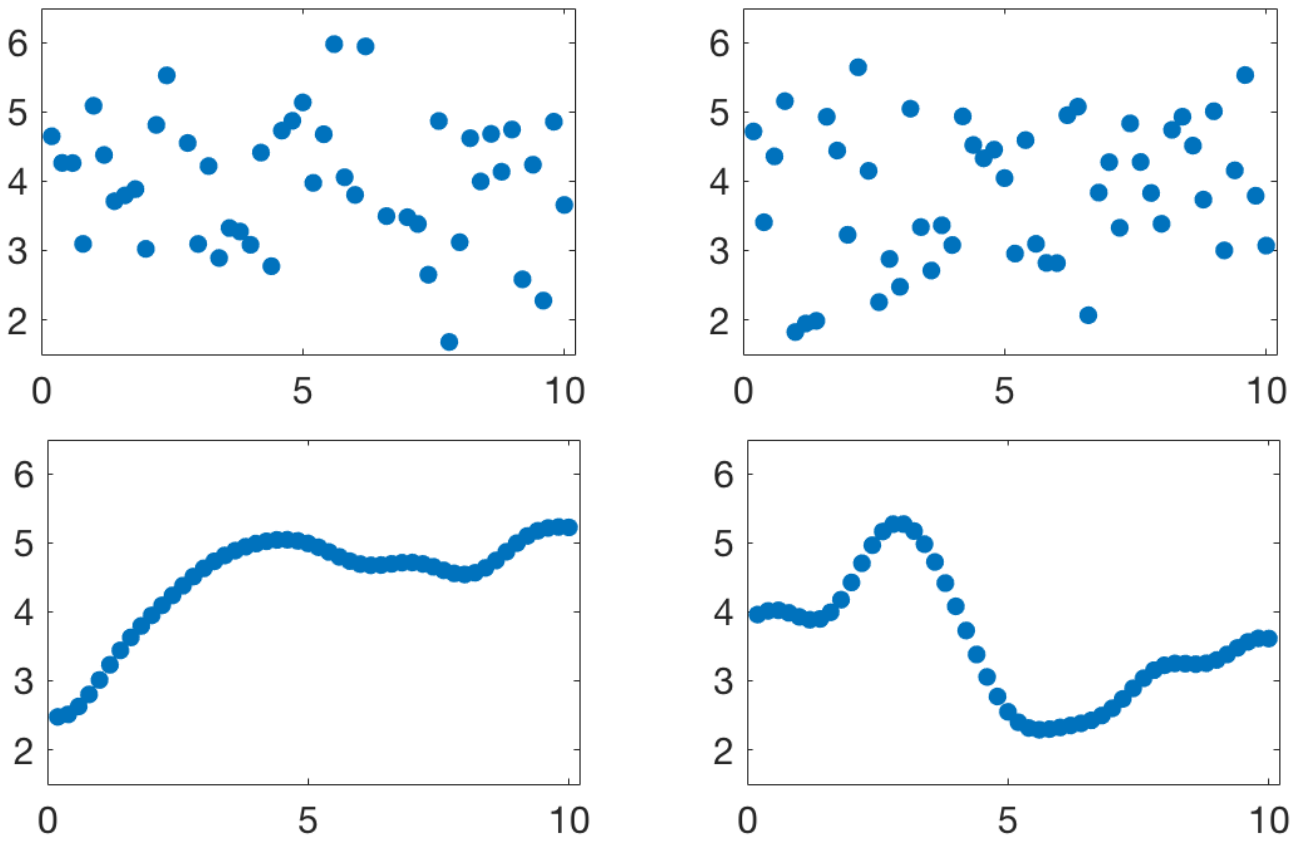
Intuition: 50 Gaussian random variables

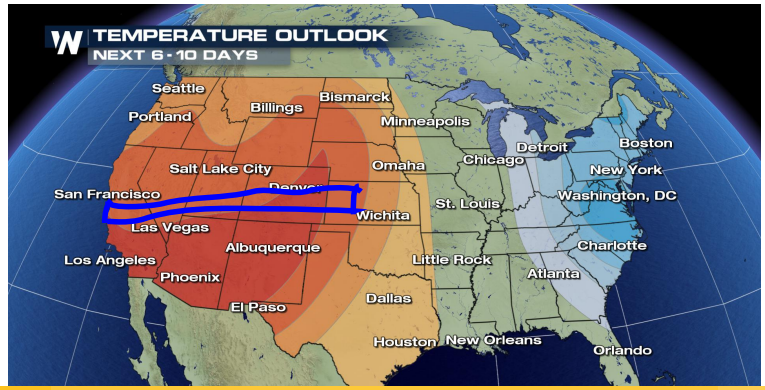


Intuition: 50 Gaussian random variables



Intuition: 50 Gaussian random variables

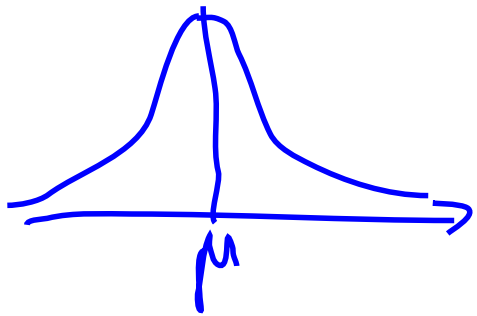




Multivariate Gaussian distribution

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

$y \sim \mathcal{N}(\mu, \sigma^2)$ $\quad p(y) \propto \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$



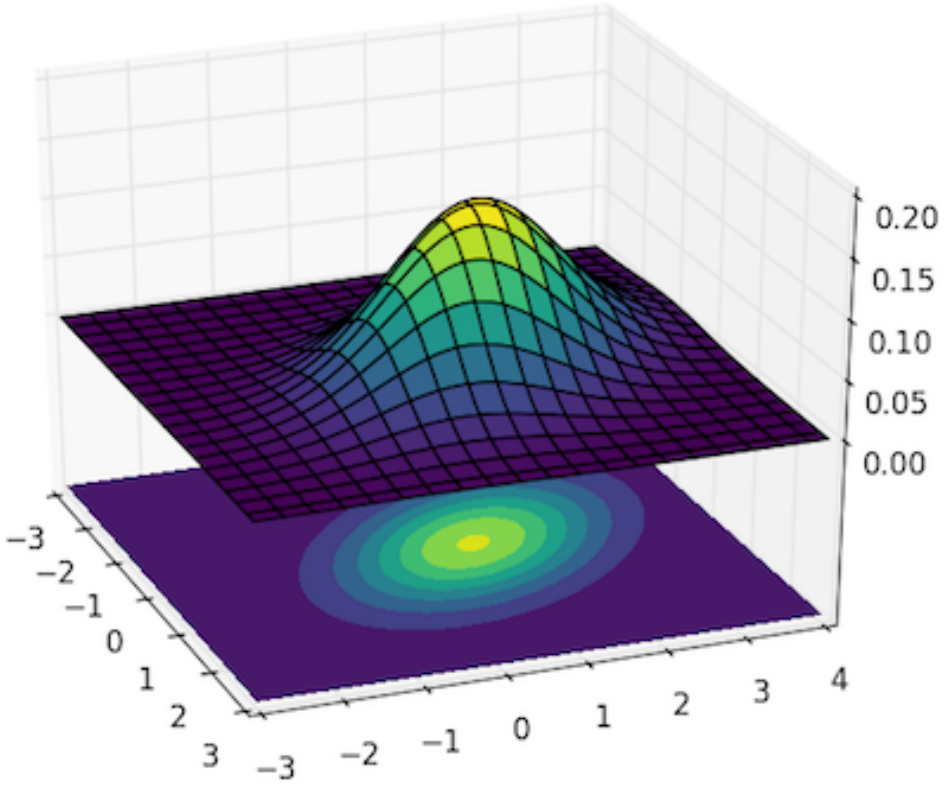
$$\begin{bmatrix} y_A \\ y_B \end{bmatrix}$$

$$\begin{bmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{bmatrix}$$

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

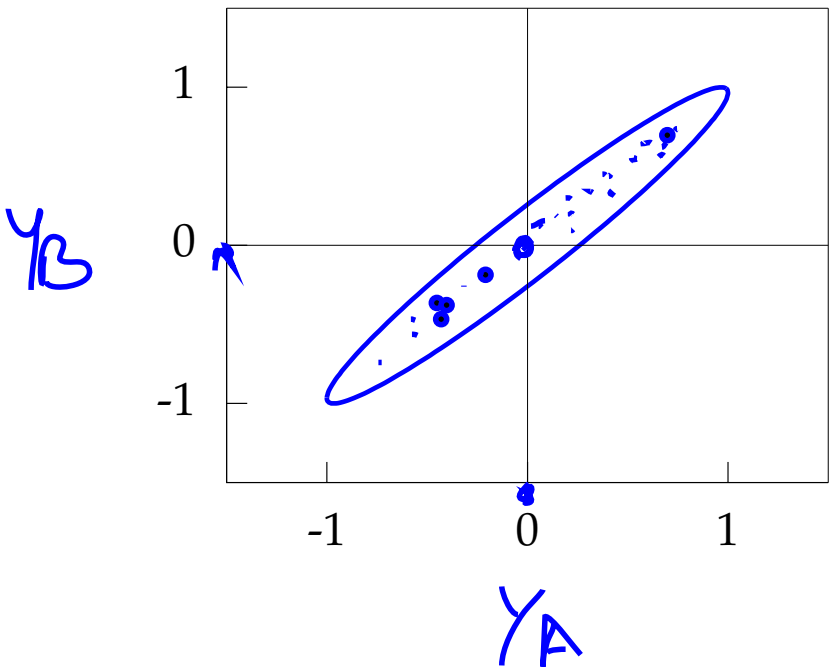
Multivariate Gaussian distribution

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



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

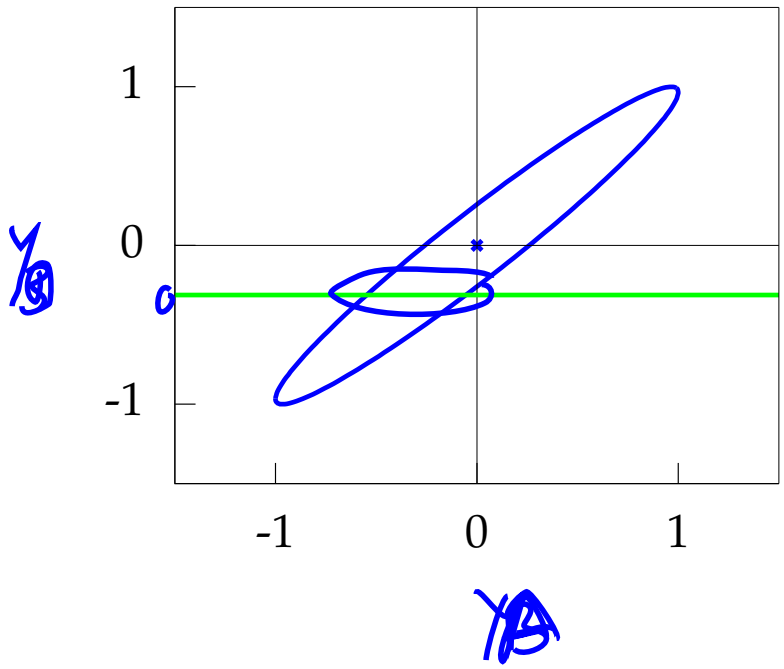
Covariance for 2 variables: intuition



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

Illustrations: Neil Lawrence

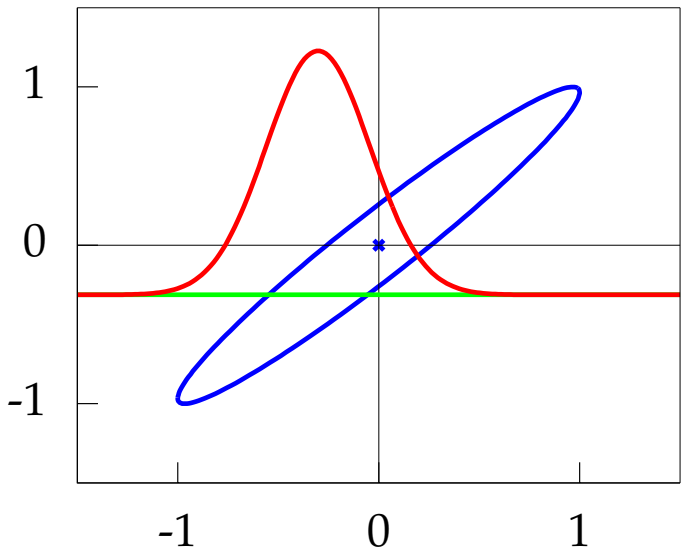
Covariance for 2 variables: intuition



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

Illustrations: Neil Lawrence

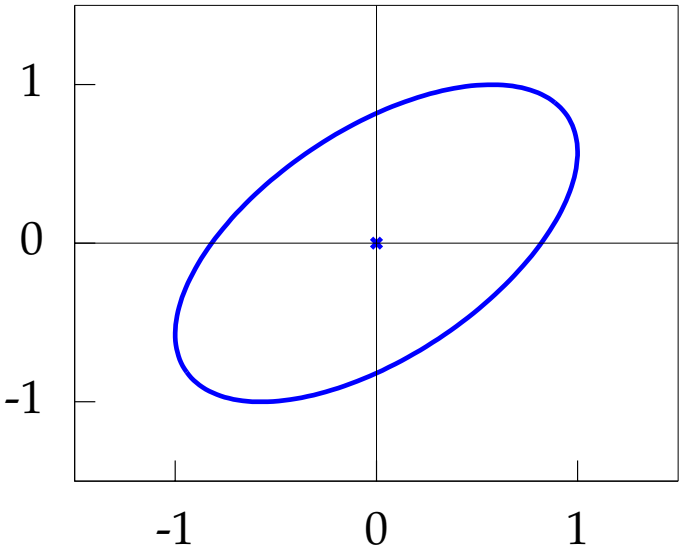
Covariance for 2 variables: intuition



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

Illustrations: Neil Lawrence

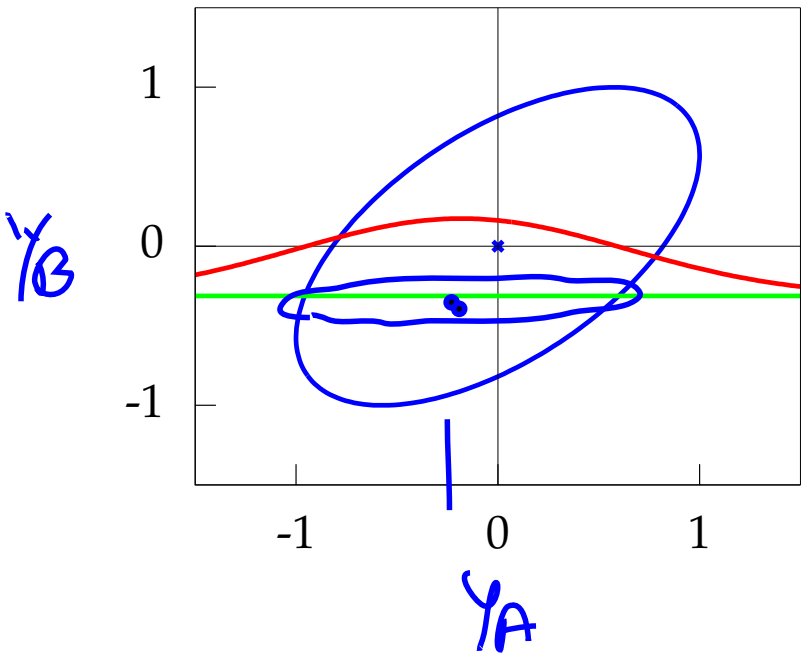
Covariance for 2 variables: intuition



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

Illustrations: Neil Lawrence

Covariance for 2 variables: intuition



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

Illustrations: Neil Lawrence

Overview

- Environmental data
- Modeling flows
- Short-range spatial correlations
  - intuition
  - 2 variables
  - multiple variables

Prediction: conditional probabilities

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

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

$Y_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

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

$$p(y_A | y_B) = \frac{p(y_A, y_B)}{p(y_B)} = \frac{(\quad) \exp(-(\quad)^T \Sigma^{-1} (\quad))}{p(y_B)}$$

Prediction: conditional probabilities

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

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

$Y_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$

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

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

Prediction: conditional probabilities

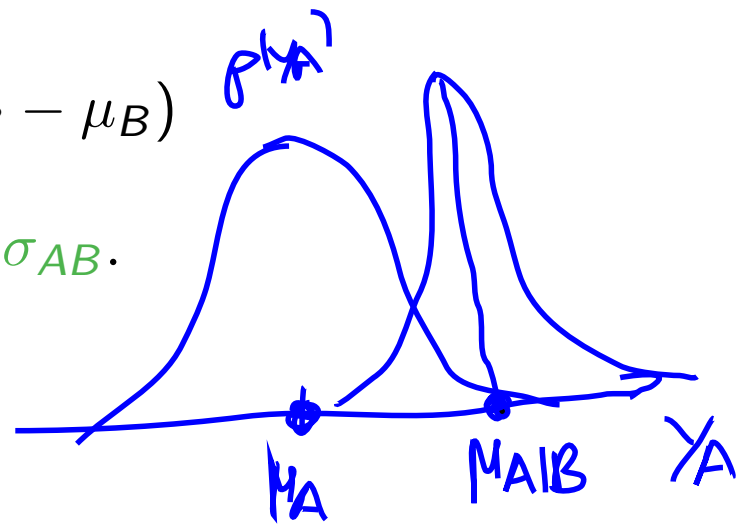
- Y_A, Y_B Gaussian random variables. We observe $Y_B = y_B$.

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

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

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

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



Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{pmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{pmatrix}$$

ij-th entry: $\mathrm{cov}(Y_i, Y_j)$

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{pmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{pmatrix}$$

$$\sigma_{N*} = [\operatorname{cov}(Y_*, Y_1), \ldots, \operatorname{cov}(Y_*, Y_N)]$$

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{bmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{bmatrix}$$

$\sigma_{N*} = [\mathrm{cov}(Y_*, Y_1), \ldots, \mathrm{cov}(Y_*, Y_N)]$

$\sigma_*^2 = \mathrm{var}(Y_*)$

$p(y_*) = \mathcal{N}(0, \sigma_*^2)$

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{pmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{pmatrix}$$

$$\sigma_{N*} = [\mathrm{cov}(Y_*, Y_1), \ldots, \mathrm{cov}(Y_*, Y_N)]$$
$$\sigma_*^2 = \mathrm{var}(Y_*)$$

- Observe $y_1, \ldots, y_N$. Conditioning yields a Gaussian $p(Y_*|y_1, \ldots y_N)$ with parameters

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{pmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{pmatrix}$$

$$\sigma_{N*} = [\mathrm{cov}(Y_*, Y_1), \ldots, \mathrm{cov}(Y_*, Y_N)]$$
$$\sigma_*^2 = \mathrm{var}(Y_*)$$

- Observe $y_1, \ldots, y_N$. Conditioning yields a Gaussian $p(Y_*|y_1, \ldots y_N)$ with parameters

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

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{bmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{bmatrix}$$

$$\sigma_{N*} = [\mathrm{cov}(Y_*, Y_1), \ldots, \mathrm{cov}(Y_*, Y_N)]$$
$$\sigma_*^2 = \mathrm{var}(Y_*)$$

- Observe $y_1, \ldots, y_N$. Conditioning yields a Gaussian $p(Y_*|y_1, \ldots y_N)$ with parameters

$$\mu_{A|B} = \mu_A + \frac{\sigma_{AB}}{\sigma_B^2}(y_B - \mu_B) \quad \Rightarrow \quad \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1}(y_{1:N} - \mu_{1:N})$$

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{pmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{pmatrix}$$

$$\sigma_{N*} = [\mathrm{cov}(Y_*, Y_1), \ldots, \mathrm{cov}(Y_*, Y_N)]$$
$$\sigma_*^2 = \mathrm{var}(Y_*)$$

- Observe $y_1, \ldots, y_N$. Conditioning yields a Gaussian $p(Y_*|y_1, \ldots y_N)$ with parameters

$$\mu_{A|B} = \mu_A + \frac{\sigma_{AB}}{\sigma_B^2}(y_B - \mu_B) \quad \Rightarrow \quad \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1}(y_{1:N} - \mu_{1:N})$$

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

Conditioning on multiple variables / partial observation

- Assume $Y_1, \ldots, Y_N, Y_*$ are jointly Gaussian with $\mu = 0$ and

$$\Sigma = \begin{bmatrix} \Sigma_N & \sigma_{N*}^\top \\ \sigma_{N*} & \sigma_*^2 \end{bmatrix}$$

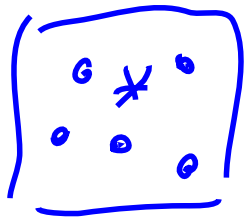
$$\sigma_{N*} = [\mathrm{cov}(Y_*, Y_1), \ldots, \mathrm{cov}(Y_*, Y_N)]$$
$$\sigma_*^2 = \mathrm{var}(Y_*)$$

- Observe $y_1, \ldots, y_N$. Conditioning yields a Gaussian $p(Y_*|y_1, \ldots y_N)$ with parameters

$$\mu_{A|B} = \mu_A + \frac{\sigma_{AB}}{\sigma_B^2}(y_B - \mu_B) \quad \Rightarrow \quad \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1}(y_{1:N} - \mu_{1:N})$$

$$\sigma^2_{A|B} = \sigma_A^2 - \sigma_{AB}\sigma_B^{-2}\sigma_{AB} \quad \Rightarrow \quad \sigma^2_{*|1:N} = \sigma_*^2 - \sigma_{N*}^\top \Sigma_N^{-1}\sigma_{N*}.$$

Comments



- prediction: mean/mode $\hat{y}_* = \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1} y_{1:N}$ (for $\mu = 0$)

Comments

- prediction: mean/mode $\hat{y}_* = \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1} y_{1:N}$ (for $\mu = 0$)
- variance shrinks as we observe more data: $\sigma^2_{*|1:N} \leq \sigma^2_*$

Comments

- prediction: mean/mode $\hat{y}_* = \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1} y_{1:N}$ (for $\mu = 0$)
- variance shrinks as we observe more data: $\sigma^2_{*|1:N} \leq \sigma^2_*$
- example of *Bayesian Inference*.

$p(y_*) = (\mu_*, \sigma_*^2)$ ←

$p(y_* | y_1 \dots y_N)$

Comments

- prediction: mean/mode $\hat{y}_* = \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1} y_{1:N}$ (for $\mu = 0$)
- variance shrinks as we observe more data: $\sigma^2_{*|1:N} \leq \sigma^2_*$
- example of *Bayesian Inference*.

- Just need the *covariance* ...

Comments

- prediction: mean/mode $\hat{y}_* = \mu_{*|1:N} = \mu_* + \sigma_{N*}^\top \Sigma_N^{-1} y_{1:N}$ (for $\mu = 0$)
- variance shrinks as we observe more data: $\sigma^2_{*|1:N} \leq \sigma^2_*$
- example of *Bayesian Inference*.

- Just need the *covariance* ...
- Idea: covariance is a function of distance!

Summary

- Specifics of environmental data: spatio-temporal dependencies, physical models, simulations
- Modeling flows via discretization
- Modeling short-range spatial correlations with Gaussians

References and Further Reading

Environmental Data, Climate Informatics, Comp. Sustainability:

- H. Kaper, H. Engler. *Mathematics & Climate*. Chapters 1, 8.
- W. Menke, J. Menke. *Environmental Data Analysis with Matlab*
- Tackling Climate Change with Machine Learning. `https://arxiv.org/abs/1906.05433` *list of high impact problems*
- Computational Sustainability: Computing for a Better World and a Sustainable Future. *Communications of the ACM*, Sep 2019. `https://cacm.acm.org/magazines/2019/9/238970-computational-sustainability/fulltext`
- `http://www.climateinformatics.org/`