DATA ANALYSIS
Year 2019–2020

Statistical Methods

Master M1–Marine Physics
Université de Bretagne Occidentale

Jonathan GULA
[gula@univ-brest.fr]

Statistical Analysis

Jonathan GULA
gula@univ-brest.fr

Objectif

Basic knowledge of statistical data analysis methods and application to geophysics data

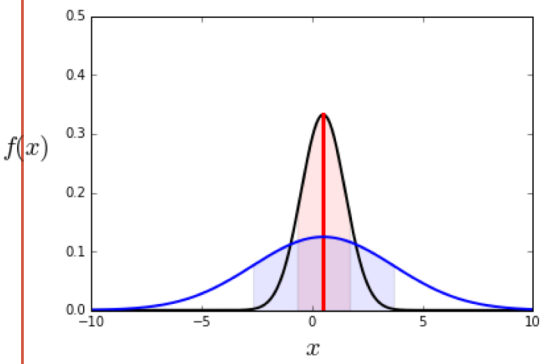
Plan

- Basic notions: random variables, PDF, CDF, distributions, moments, estimators, the central limit theorem.
- Resampling methods (bootstrap, jackknife, Monte Carlo method), construction of confidence intervals, hypothesis testing, Bayesian statistics
- Extreme value theory , generalized Pareto distributions

Statistical Analysis

Jonathan GULA
gula@univ-brest.fr

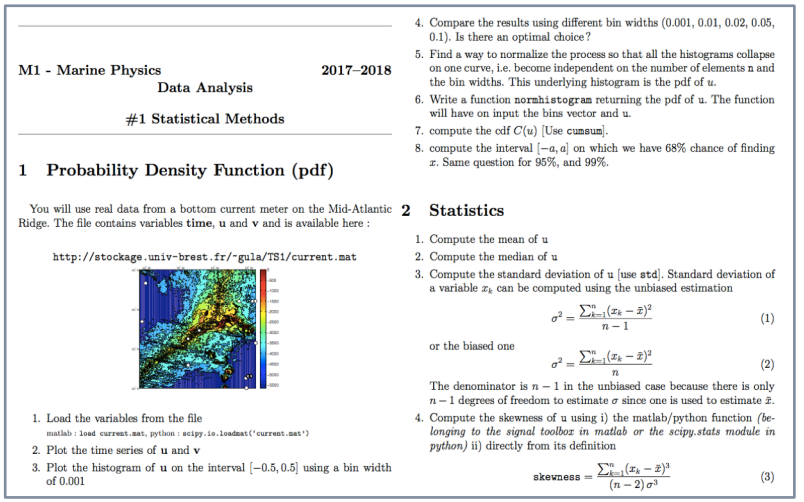
Lessons (basic theory)



$$\sigma^2 = \int (x-\mu)^2 f(x)\, dx$$

Activities

- Applications to geophysics data (mooring, argo, model, etc.) using **Python**



M1 - Marine Physics 2017–2018

Data Analysis

#1 Statistical Methods

1 Probability Density Function (pdf)

You will use real data from a bottom current meter on the Mid-Atlantic Ridge. The file contains variables **time**, **u** and **v** and is available here :

`http://stockage.univ-brest.fr/~gula/TS1/current.mat`

1. Load the variables from the file
   matlab : `load current.mat`, python : `scipy.io.loadmat('current.mat')`
2. Plot the time series of **u** and **v**
3. Plot the histogram of **u** on the interval $[-0.5, 0.5]$ using a bin width of 0.001
4. Compare the results using different bin widths (0.001, 0.01, 0.02, 0.05, 0.1). Is there an optimal choice ?
5. Find a way to normalize the process so that all the histograms collapse on one curve, i.e. become independent on the number of elements n and the bin widths. This underlying histogram is the pdf of u.
6. Write a function `normhistogram` returning the pdf of u. The function will have on input the bins vector and u.
7. compute the cdf $C(u)$ [Use `cumsum`].
8. compute the interval $[-a, a]$ on which we have 68% chance of finding x. Same question for 95%, and 99%.

2 Statistics

1. Compute the mean of u
2. Compute the median of u
3. Compute the standard deviation of u [use `std`]. Standard deviation of a variable x_k can be computed using the unbiased estimation

$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n-1} \quad (1)$$

or the biased one

$$\sigma^2 = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n} \quad (2)$$

The denominator is $n-1$ in the unbiased case because there is only $n-1$ degrees of freedom to estimate σ since one is used to estimate $\bar{x}$.

4. Compute the skewness of u using i) the matlab/python function *(belonging to the signal toolbox in matlab or the scipy.stats module in python)* ii) directly from its definition

$$\text{skewness} = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^3}{(n-2)\,\sigma^3} \quad (3)$$

Grade

1 homework + 1 computer exam

Introduction

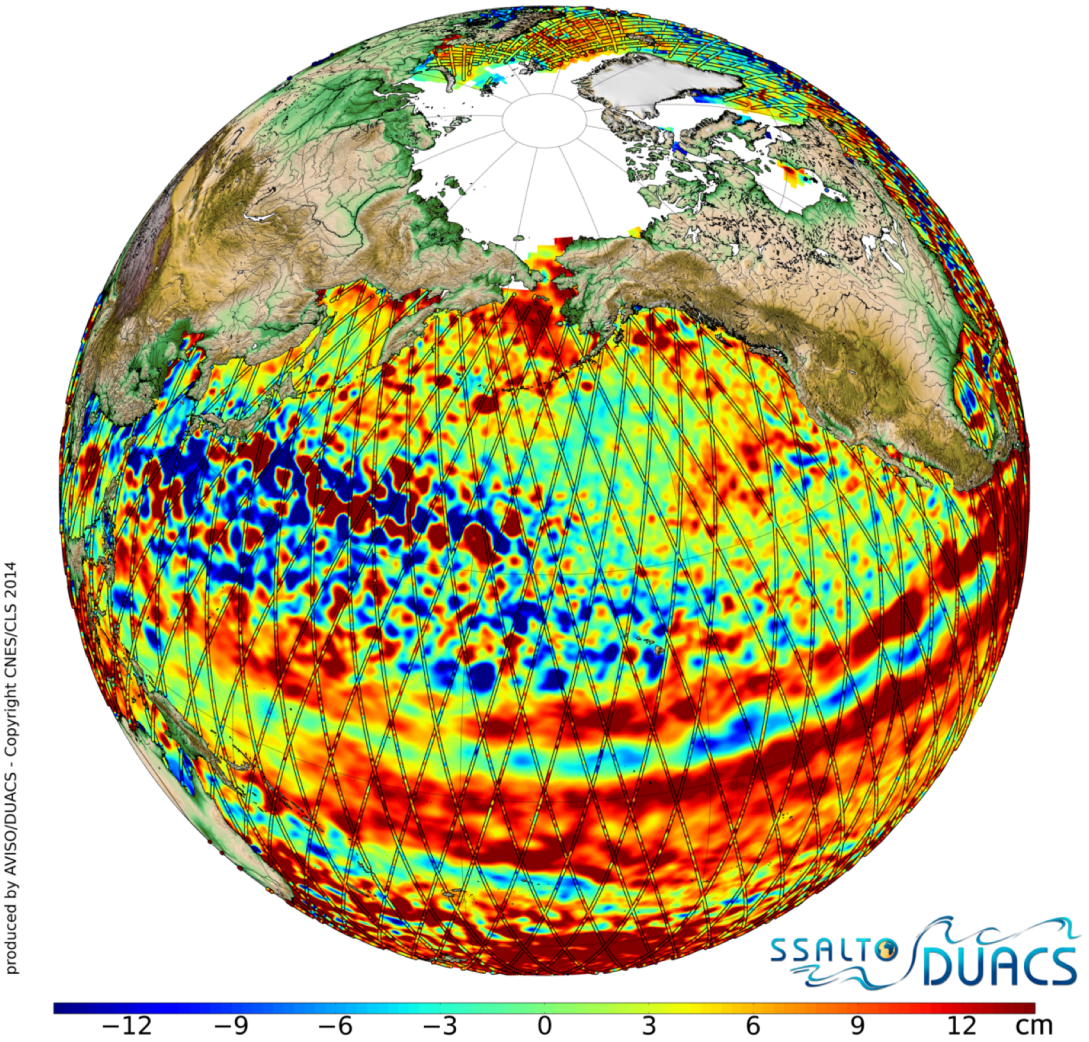
What kind of data are we dealing with?

How do we observe the ocean?

Sampling the ocean

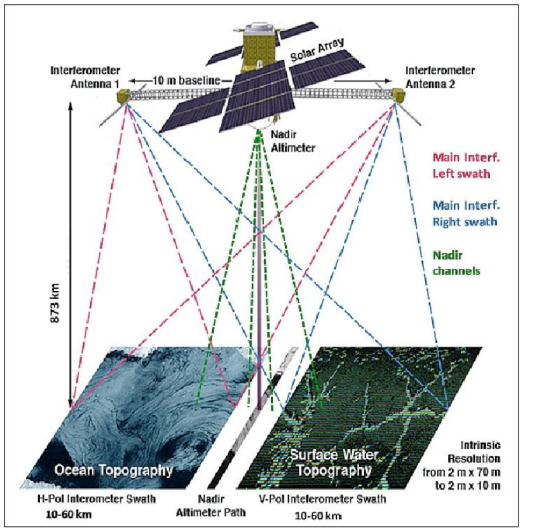
Altimetric measurements from satellite:

- **Ssalto/Duacs products**
  - Resolution = ¼ degree,
  - Sampling interval = 2/3 days,



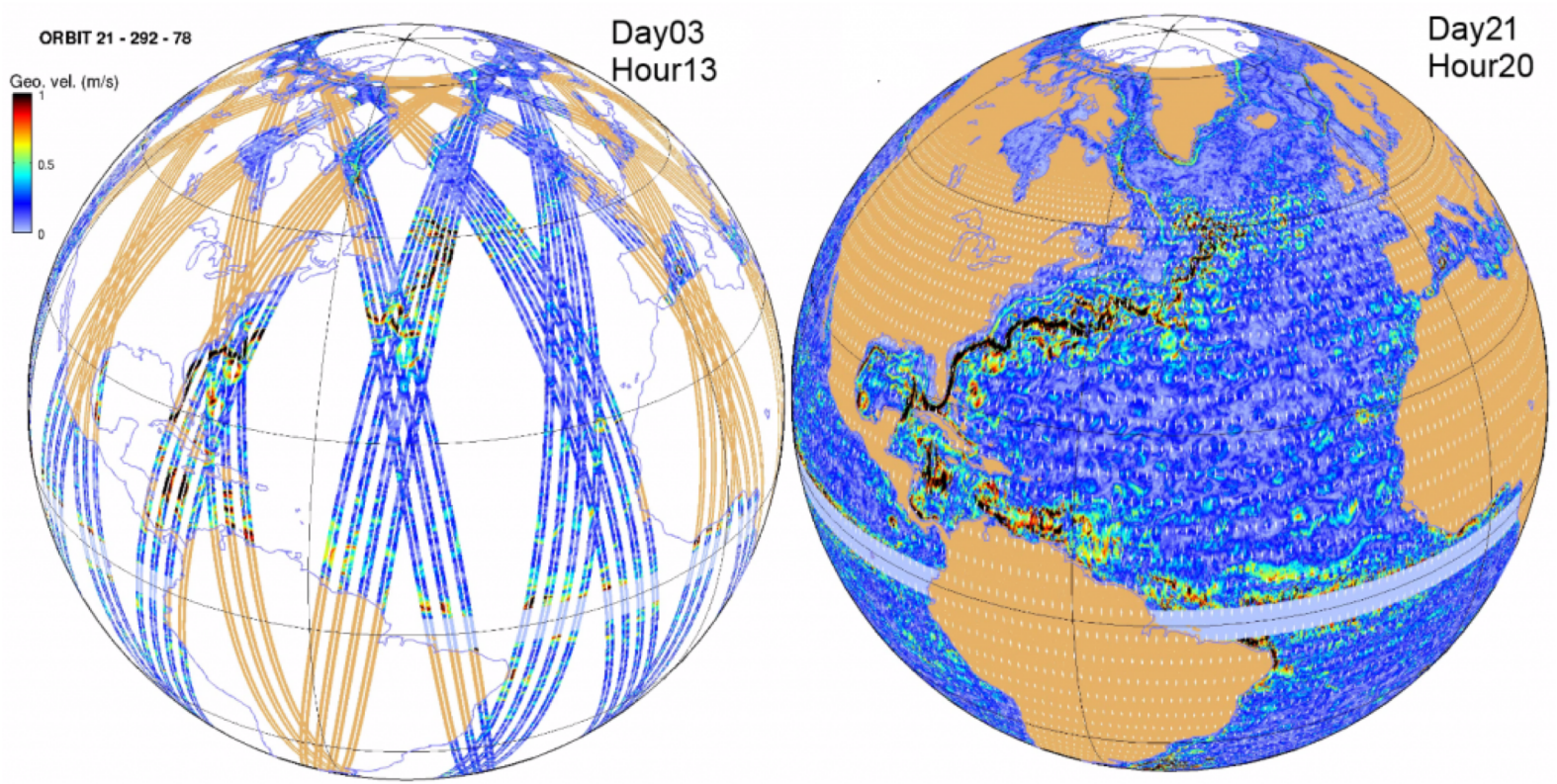
http://www.aviso.altimetry.fr/en/home.html

Sampling the ocean



Altimetric measurements from satellite:

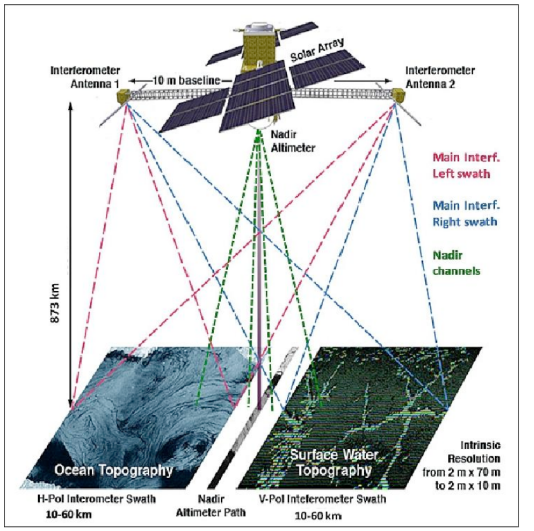
- **SWOT (2020)**
  - Resolution = 1-10 km,
  - Sampling interval ≈ 5 - 21 days,

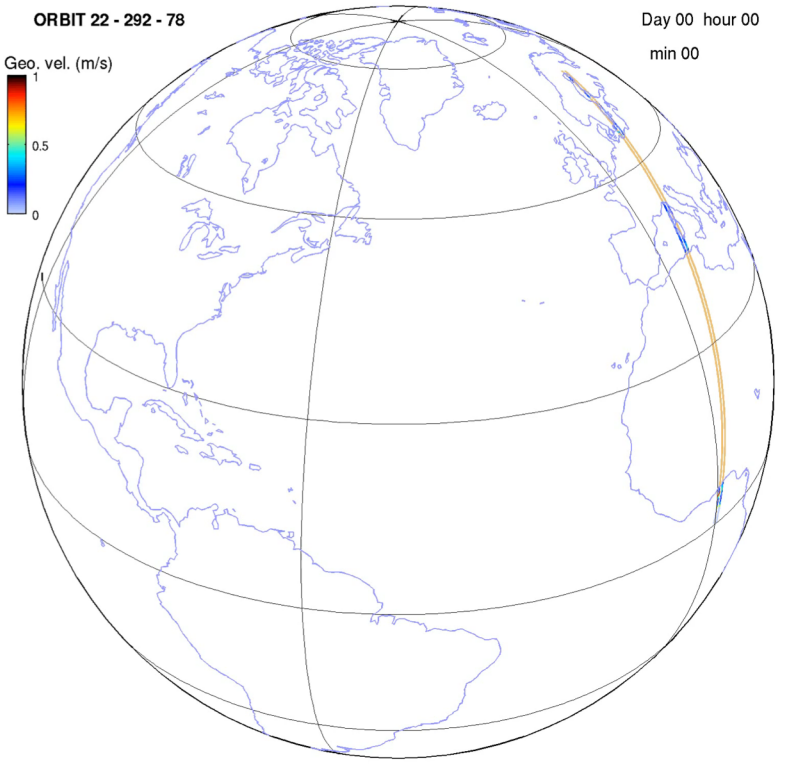


Sampling the ocean

Altimetric measurements from satellite:

- **SWOT (2020)**
  - Resolution = 1-10 km,
  - Sampling interval ≈ 5 - 21 days,

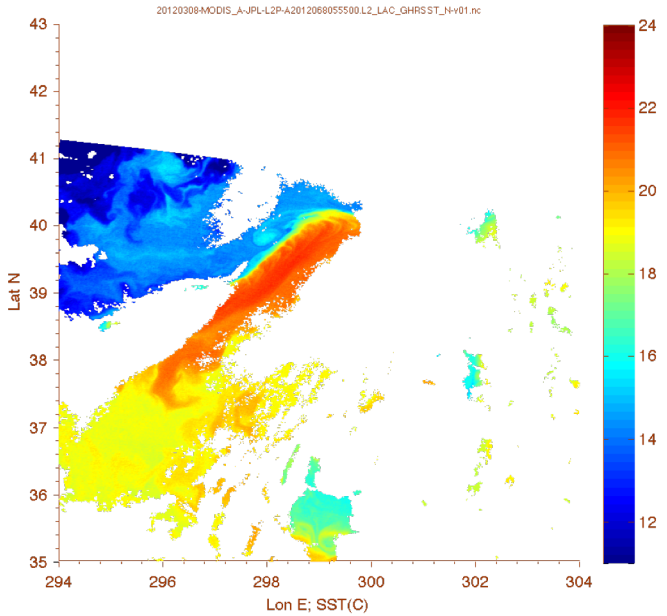


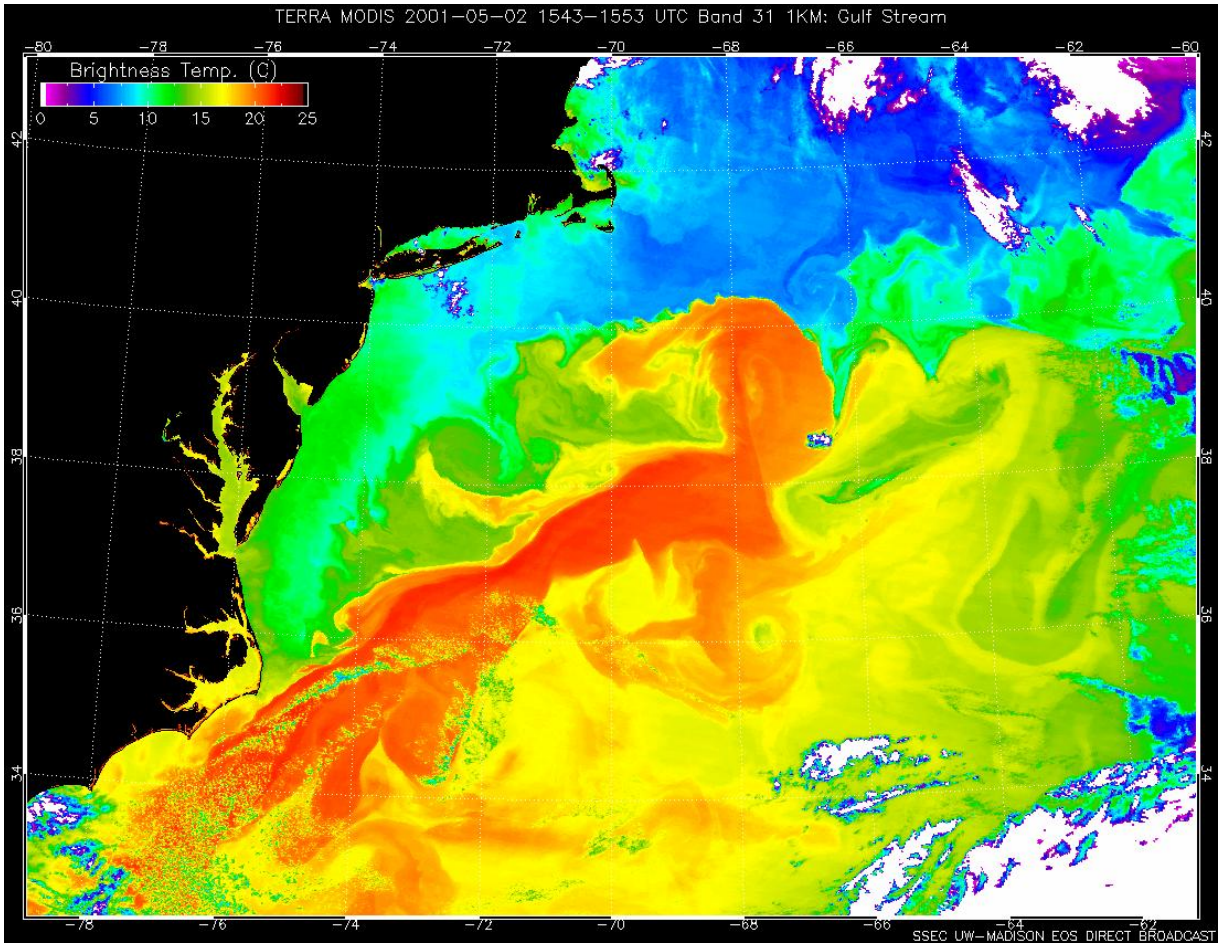


Sampling the ocean

SST/chlorophyll measurements from satellite:

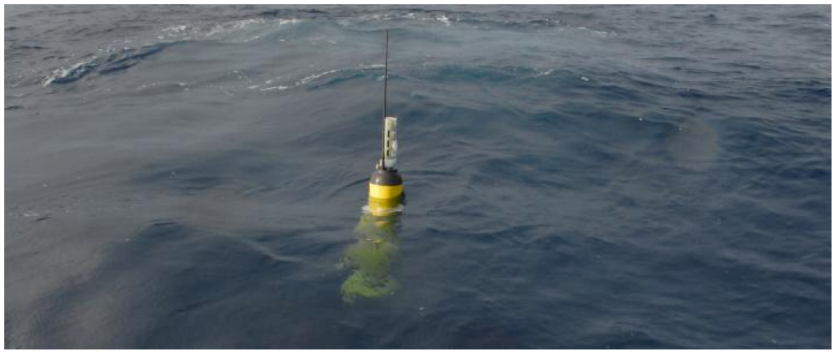
- **MODIS** **(Moderate Resolution Imaging Spectroradiometer)**
  - Resolution = 1 km,
  - Sampling interval ≈ 1 - 2 days,





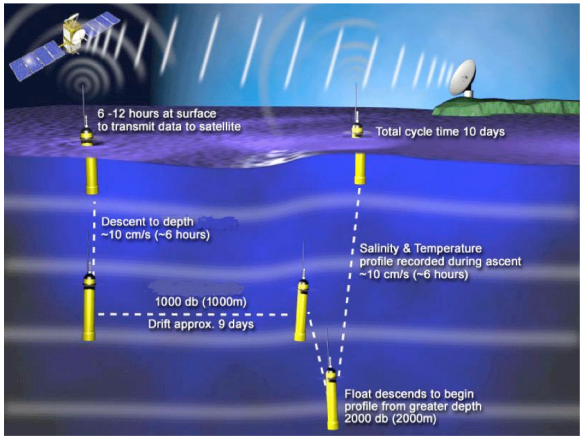
https://marine.rutgers.edu/cool/sat_data/?nothumbs=0

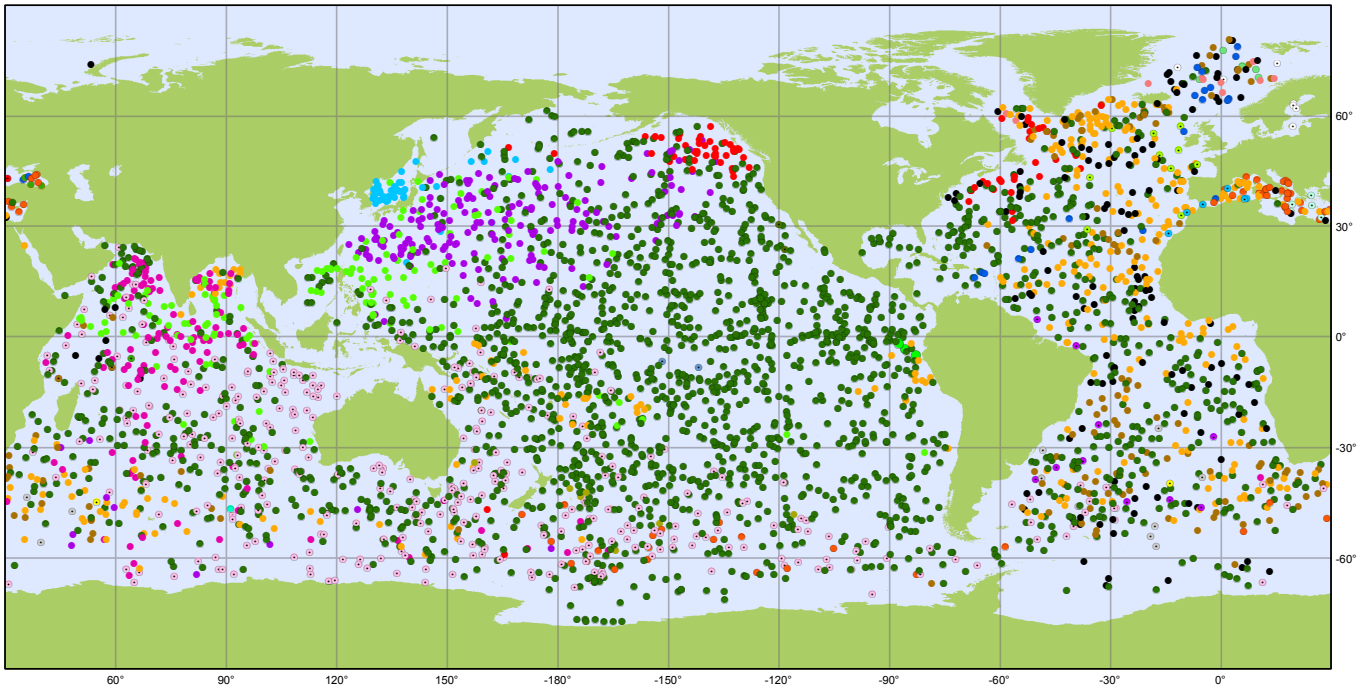
Sampling the ocean



In-situ floats:

- **Argo**
  - Vertical resolution = few meters
  - Sampling interval = 1 profile every 10 days,





Argo **National contributions - 3804 Operational Floats** **September 2016**

Latest location of operational floats (data distributed within the last 30 days)

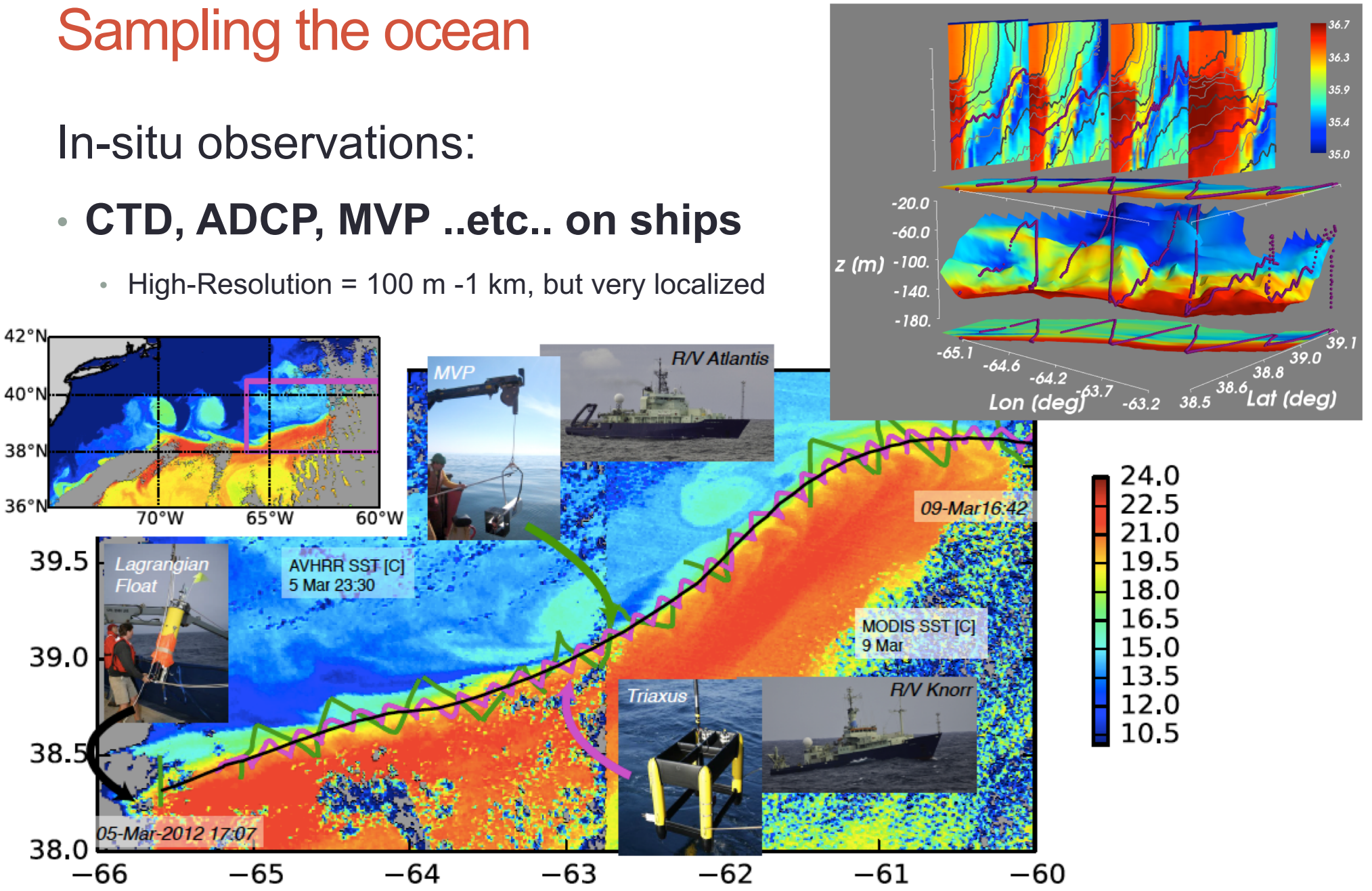
OTHER (3)	CANADA (74)	FRANCE (336)	ITALY (61)	MEXICO (2)	SOUTH AFRICA (1)
ARGENTINA (2)	CHINA (139)	GERMANY (141)	JAPAN (187)	NETHERLANDS (12)	SPAIN (6)
AUSTRALIA (363)	ECUADOR (2)	GREECE (6)	KENYA (1)	NEW ZEALAND (9)	TURKEY (1)
BRAZIL (10)	EUROPE (21)	INDIA (129)	KOREA, REPUBLIC OF (44)	NORWAY (10)	UK (135)
BULGARIA (3)	FINLAND (7)	IRELAND (10)	MAURITIUS (2)	POLAND (5)	USA (2078)

Generated by www.jcommops.org, 08/10/2016

Sampling the ocean

In-situ observations:

- **CTD, ADCP, MVP ..etc.. on ships**
  - High-Resolution = 100 m -1 km, but very localized



LatMix Campaign at the Gulf Stream North Wall

Sampling the ocean

In-situ observations:

- **Moorings**
  - Very High vertical Resolution and sampling,

    but very localized

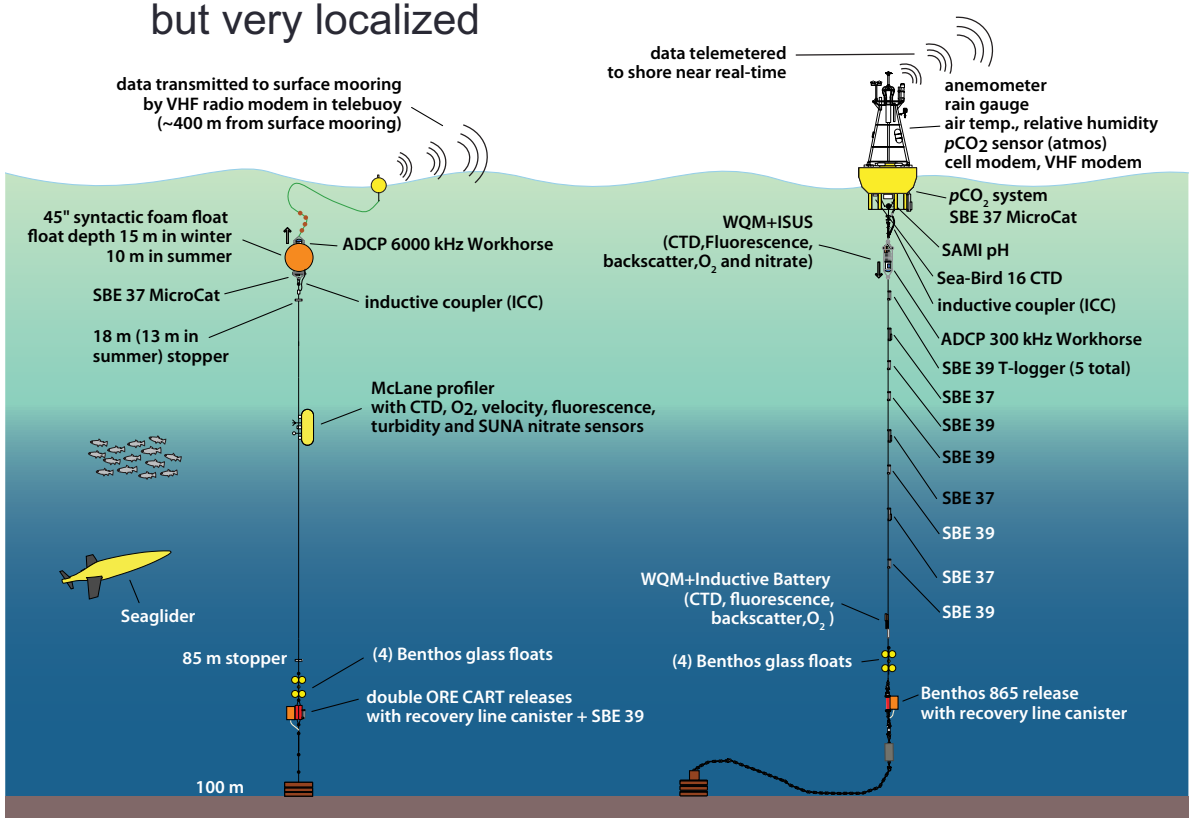


Figure 2. Schematic of the NEMO system showing the surface mooring (right), the subsurface profiling mooring (left), and the glider.

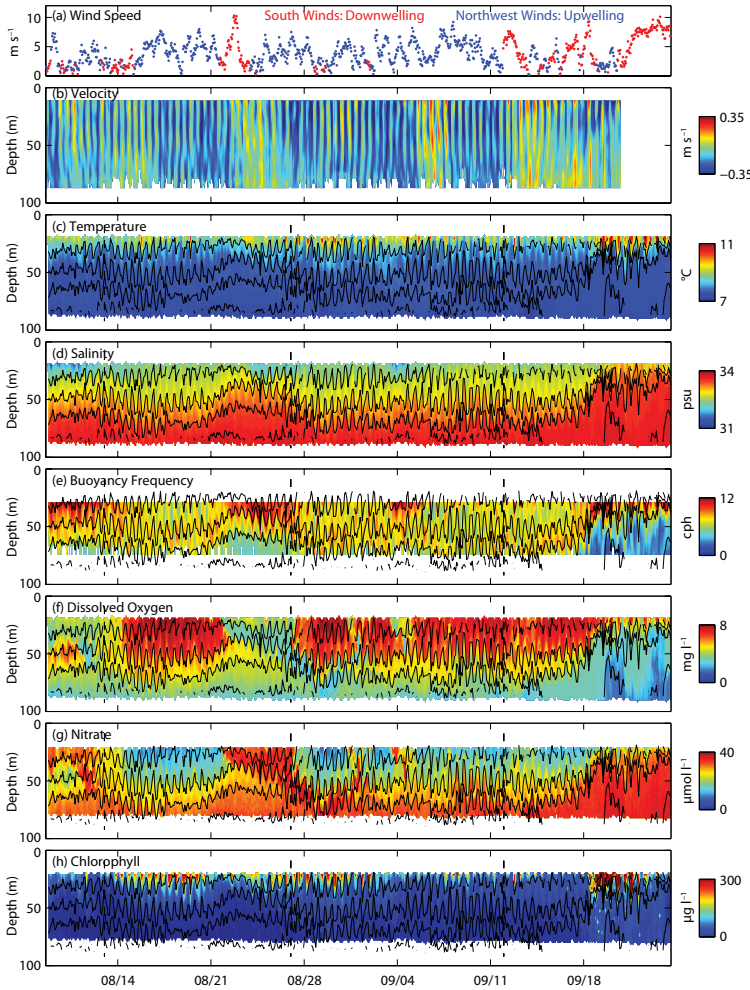


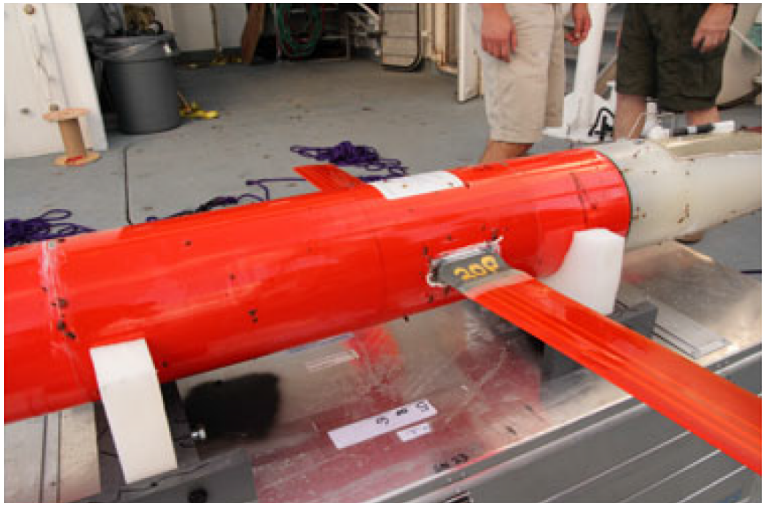
Figure 4. Time series of data from the subsurface mooring, corresponding to the last 46 days of the period plotted in the previous figure. Panels are wind speed colored by (a) direction as in Figure 3, (b) velocity toward 315° true, (c) temperature, (d) salinity, (e) buoyancy frequency, (f) dissolved oxygen, (g) nitrate, and (h) chlorophyll. Isopycnals whose mean spacing is 10 m are over-plotted in each panel in black.

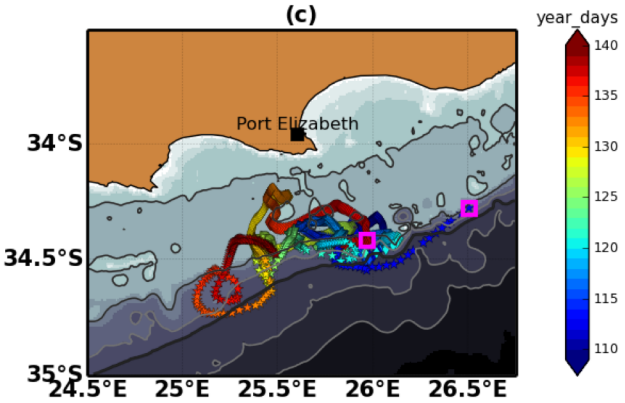
Alford, M.H., J.B. Mickett, S. Zhang, P. MacCready, Z. Zhao, and J. Newton. **2012.** Internal waves on the Washington continental shelf. *Oceanography*. 25(2):66–79.

Sampling the ocean

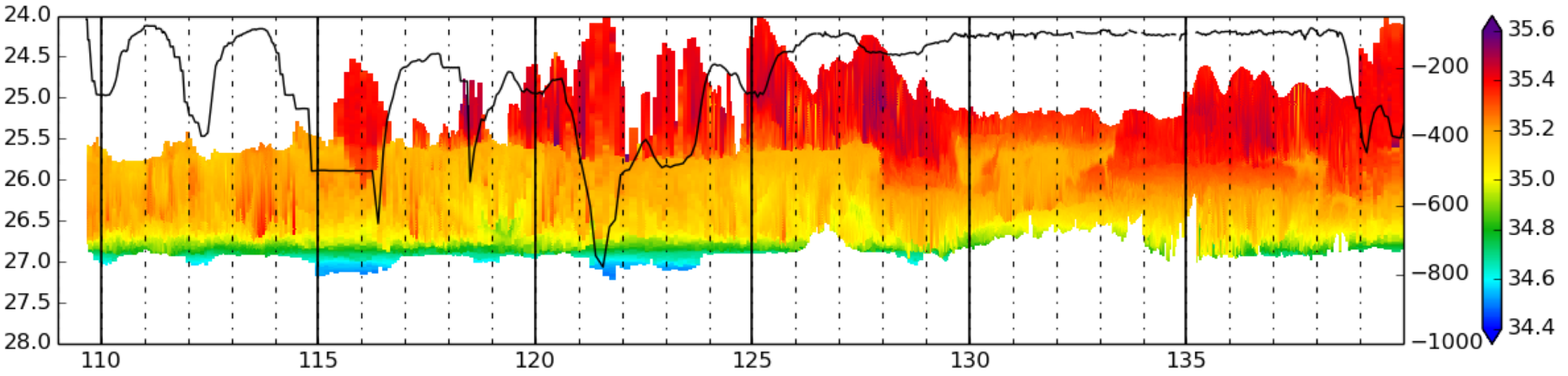
In-situ observations:

- **Gliders**
  - Very High vertical Resolution and sampling,





Salinity from Glider sections in the Aghulas Current

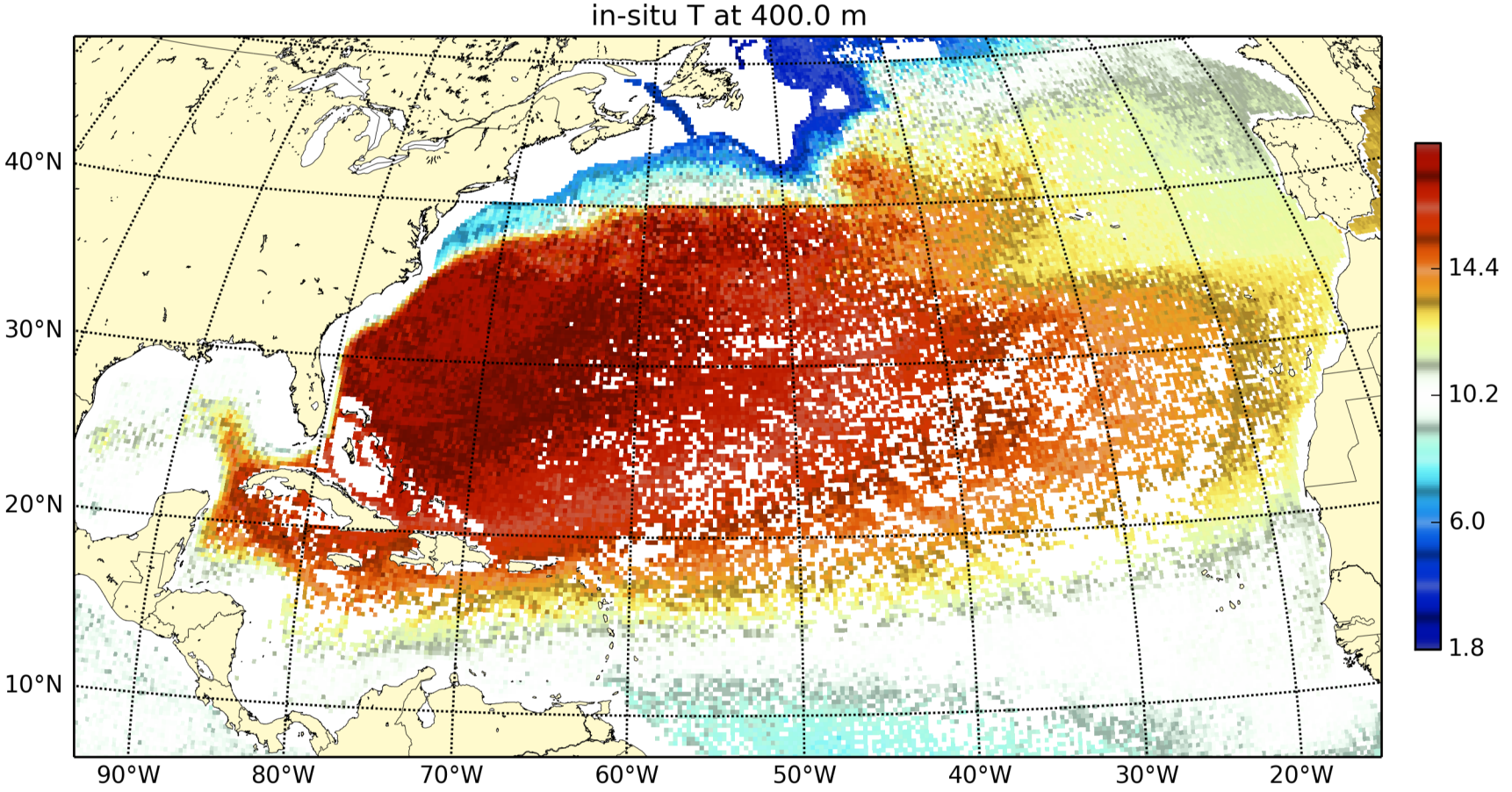


Sampling the ocean

In-situ observations:

- **World Ocean Atlas:**

All temperature data from 1955 to 2012
binned to ¼ degree grid



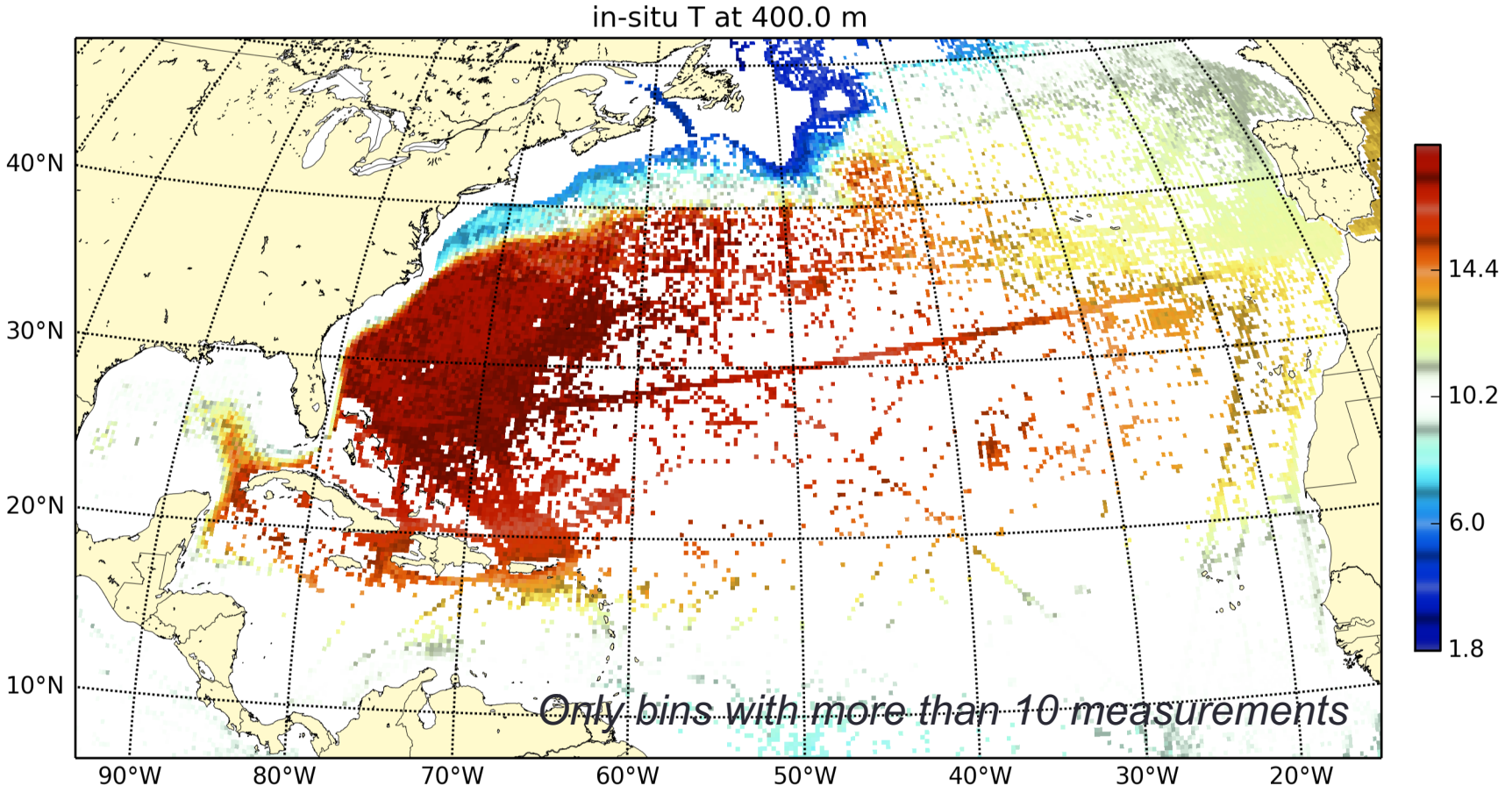
https://www.nodc.noaa.gov/OC5/SELECT/woaselect/woaselect.html

Sampling the ocean

In-situ observations:

- **World Ocean Atlas:**

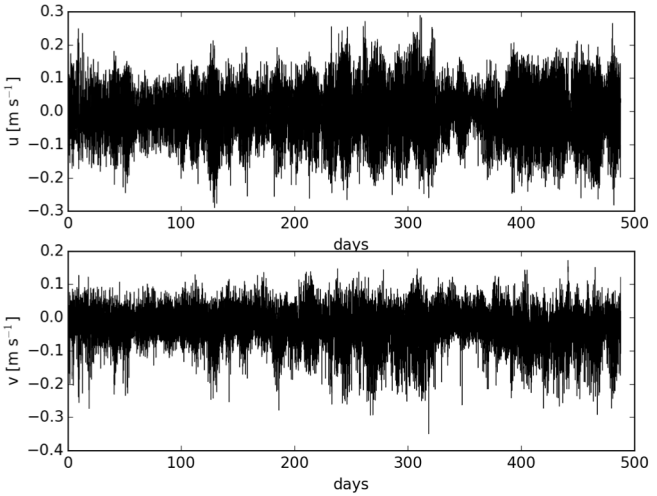
All temperature data from 1955 to 2012
binned to ¼ degree grid

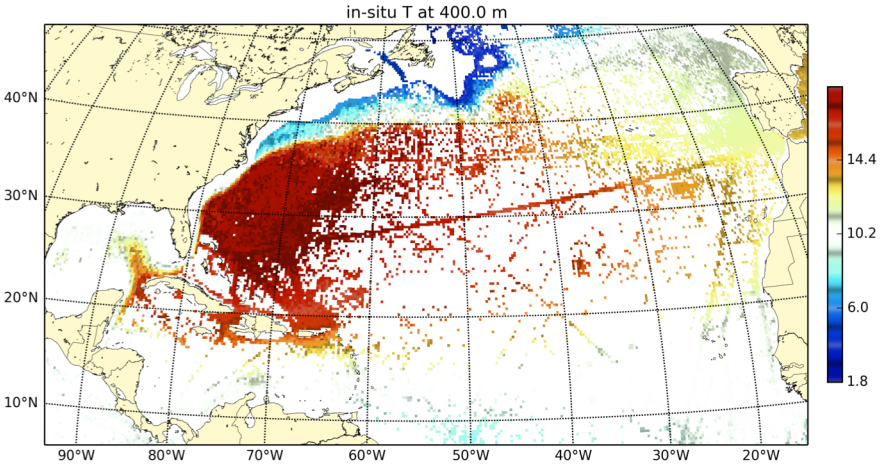


https://www.nodc.noaa.gov/OC5/SELECT/woaselect/woaselect.html

Objective of statistical methods

We have only a sample of data drawn from a much larger population.





Objective of statistical methods

We have only a sample of data drawn from a much larger population = **random variables**

Statistical methods can help you answer typical questions like:

- how reliable is your observation? how long should observations last so that all information is retained?
- Are your results statistically significant? To what extent a result happens by chance (random variation and/or errors in sampling) ?
- Is there a statistical relationship between 2 variables?

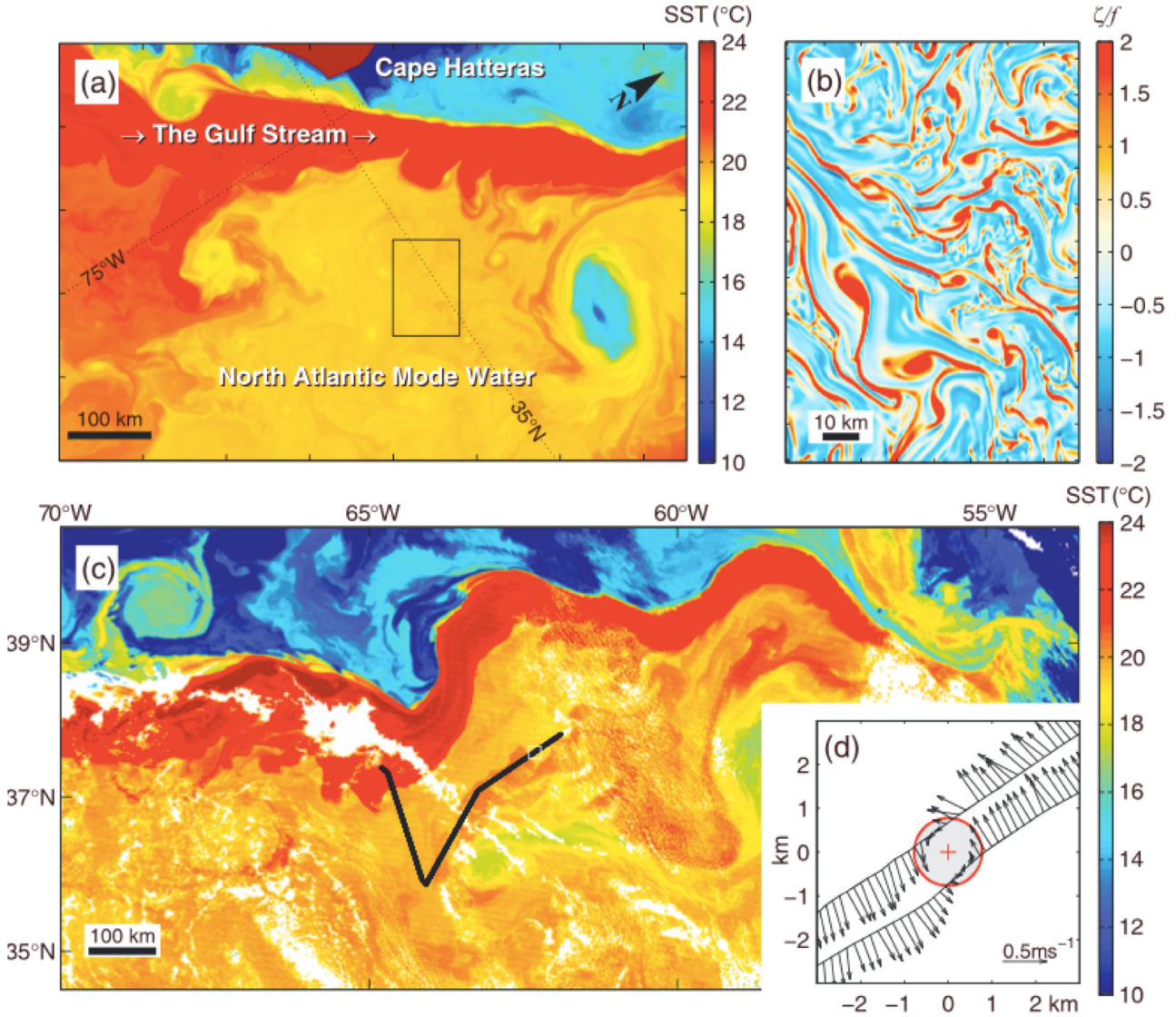
Sample distributions

Ex: How do we compare results from a model to observations:

Numerical Model

Observations

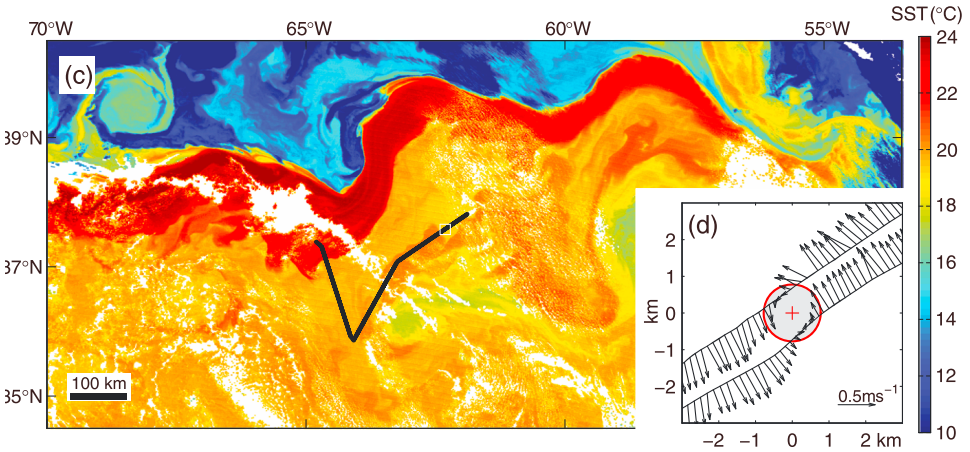
[satellite SST + velocity measurements from 2 parallel ships]



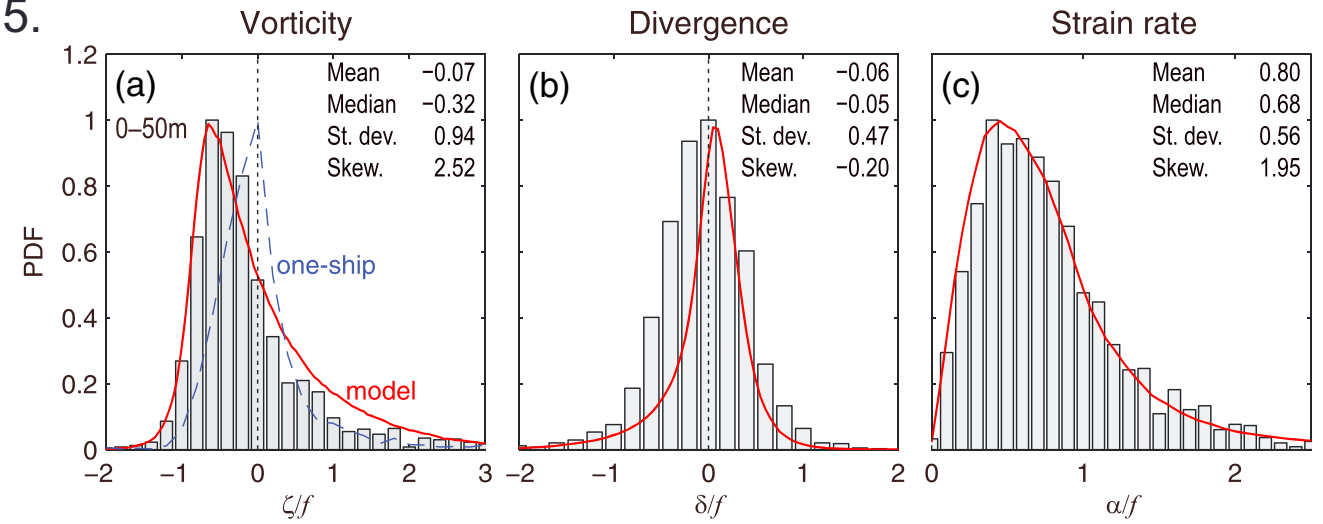
Shcherbina et al, 2015.

Sample distributions

The most basic estimate of a population distribution can be made by using the histogram of the measured data points.



Ex: Shcherbina et al, 2015.



Sample distributions

The most basic estimate of a population distribution can be made by using the histogram of the measured data points.

The most basic descriptive parameter is the sample mean: $\hat{\mu} = \frac{1}{N}\sum_k x_k$

To determine how the data are spread about the mean, we can compute the variance:

$$s^2 = \frac{1}{N-1}\sum_k (x_k - \hat{\mu})^2$$