British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UNIVERSITY OF CAMBRIDGE

Oceanographic data and how to use it

Dr Dan Jones | Oceanographer
British Antarctic Survey (BAS)
BAS Artificial Intelligence Lab

DanJonesOcean.com
@DanJonesOcean
github.com/DanJonesOcean

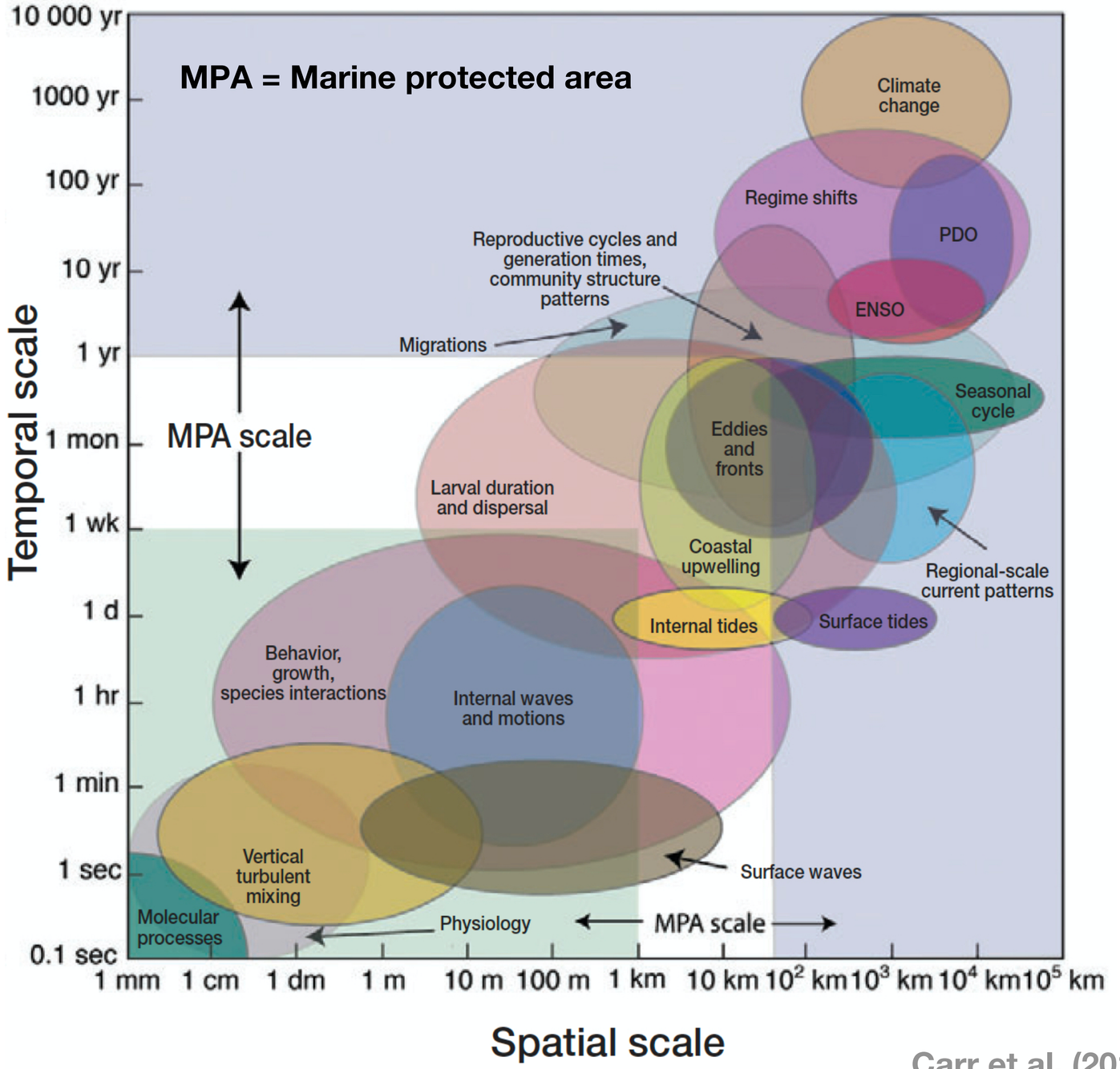
AI4ER CDT

Oceanographic data:

Spans a wide range of spatial and temporal scales

Is highly correlated in space and time

Often requires multi-disciplinary knowledge to interpret



Carr et al. (2011)

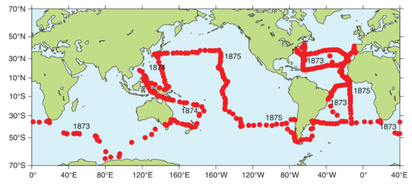
Types of oceanographic information

Observational

Data from sensors on ships, robots, satellites, etc.

“In situ”

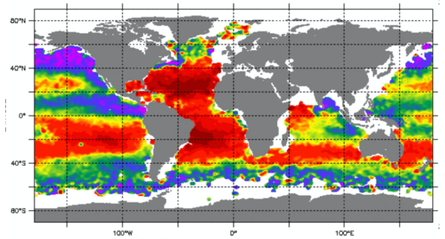
Data from the time and place where the sensor made its measurements



Sometimes called “station data”

Gridded

Interpolated / extrapolated for better spatial and temporal coverage



Often on a latitude/longitude grid

Numerical/simulated

Solutions to systems of equations, typically derived from physical/biogeochemical principles (e.g. conservation of energy)

Complexity

Idealised models

More abstract

Realistic models

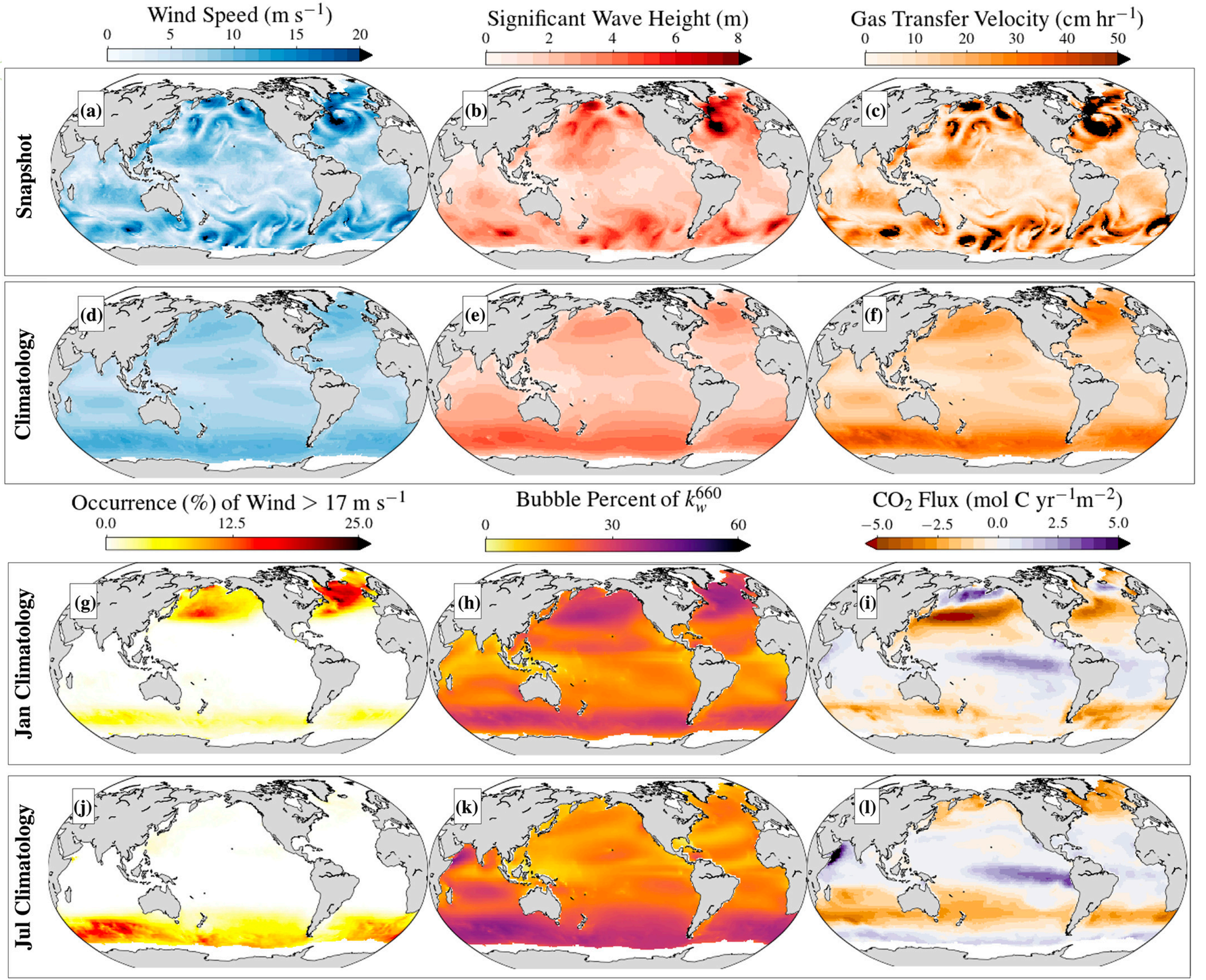
More detailed

Data assimilation products

Observations and simulations inform each other

Snapshots: instantaneous values (possibly averaged over a few days)

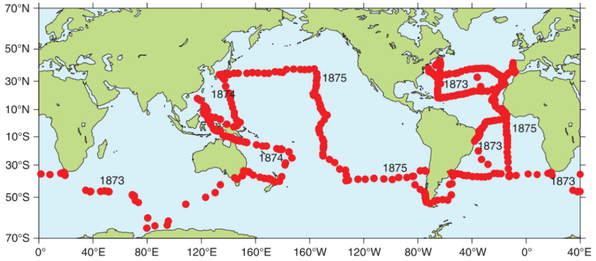
Climatology: A long-term average. Could be over a particular month or season.

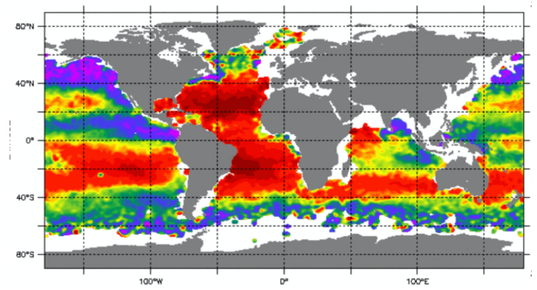


Reichl and Deike (2020)

What are observations used for?

- **Mapping** fields (i.e. creating interpolated data products)
- **Initiating** and/or constraining numerical models
- **Validating** numerical models
- **Improving** numerical model physics and parameterisations
- **Improving** process understanding



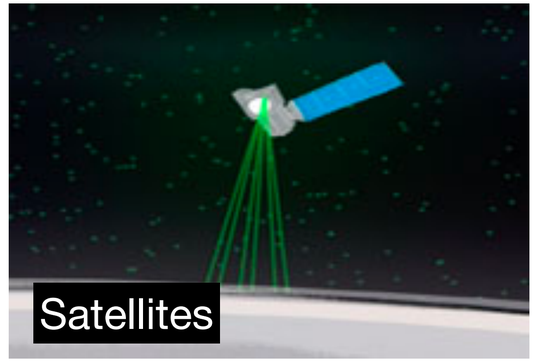


Various platforms used to collect observational data

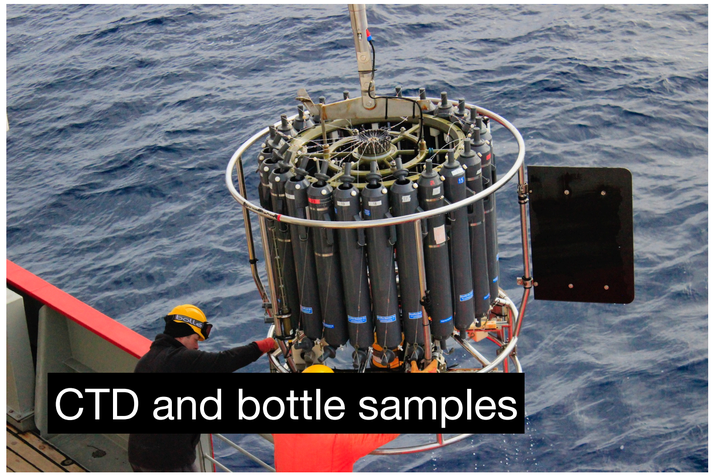

Gliders

Autosubs

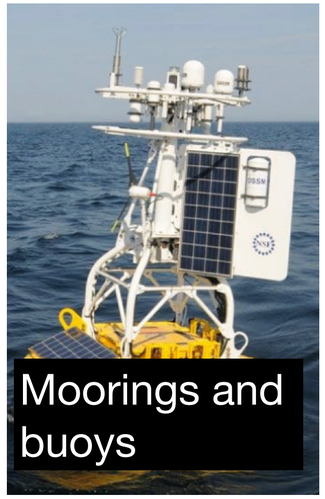

Ships


Satellites


Floats

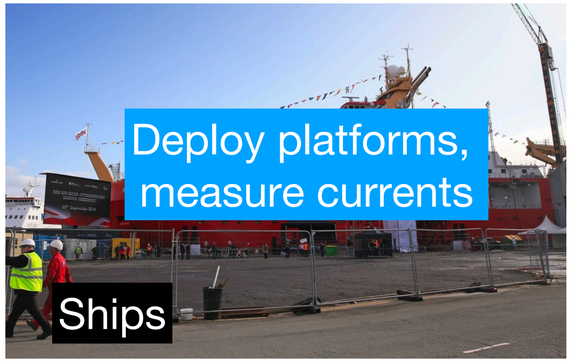

CTD and bottle samples

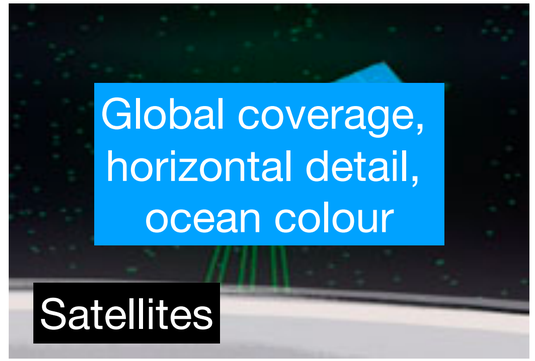

Seal tags

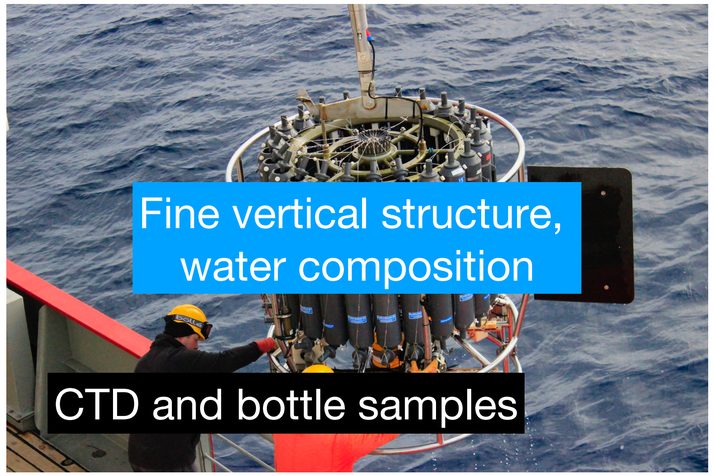

Moorings and buoys

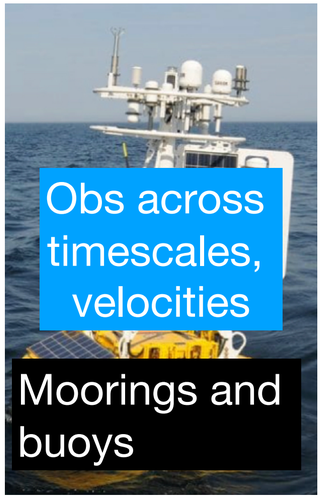
CTD: conductivity-temperature-depth sensor
Measures salinity using the conductivity of the seawater. Measures pressure as a proxy for depth.

Various platforms used to collect observational data



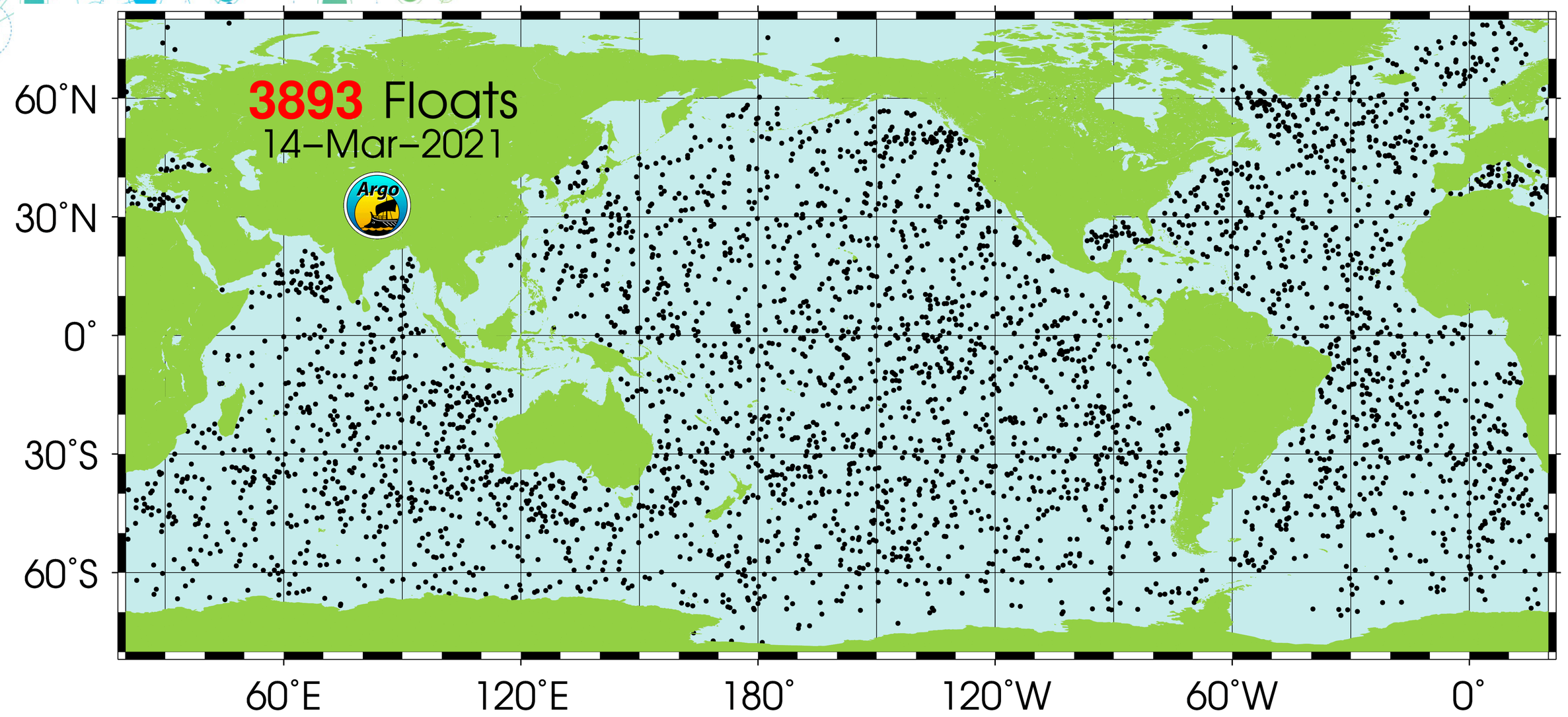




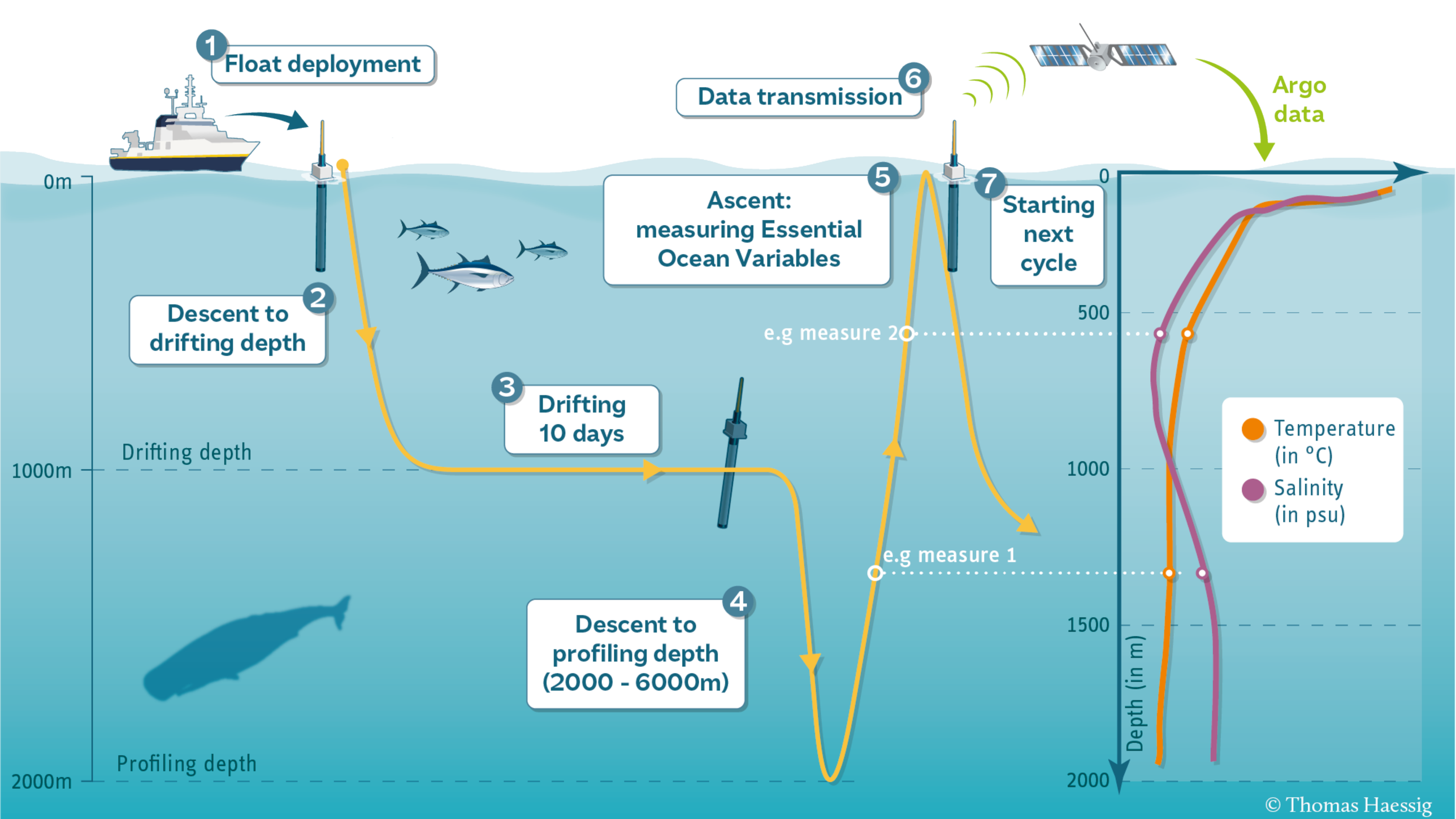


CTD: conductivity-temperature-depth sensor
Measures salinity using the conductivity of the seawater. Measures pressure as a proxy for depth.

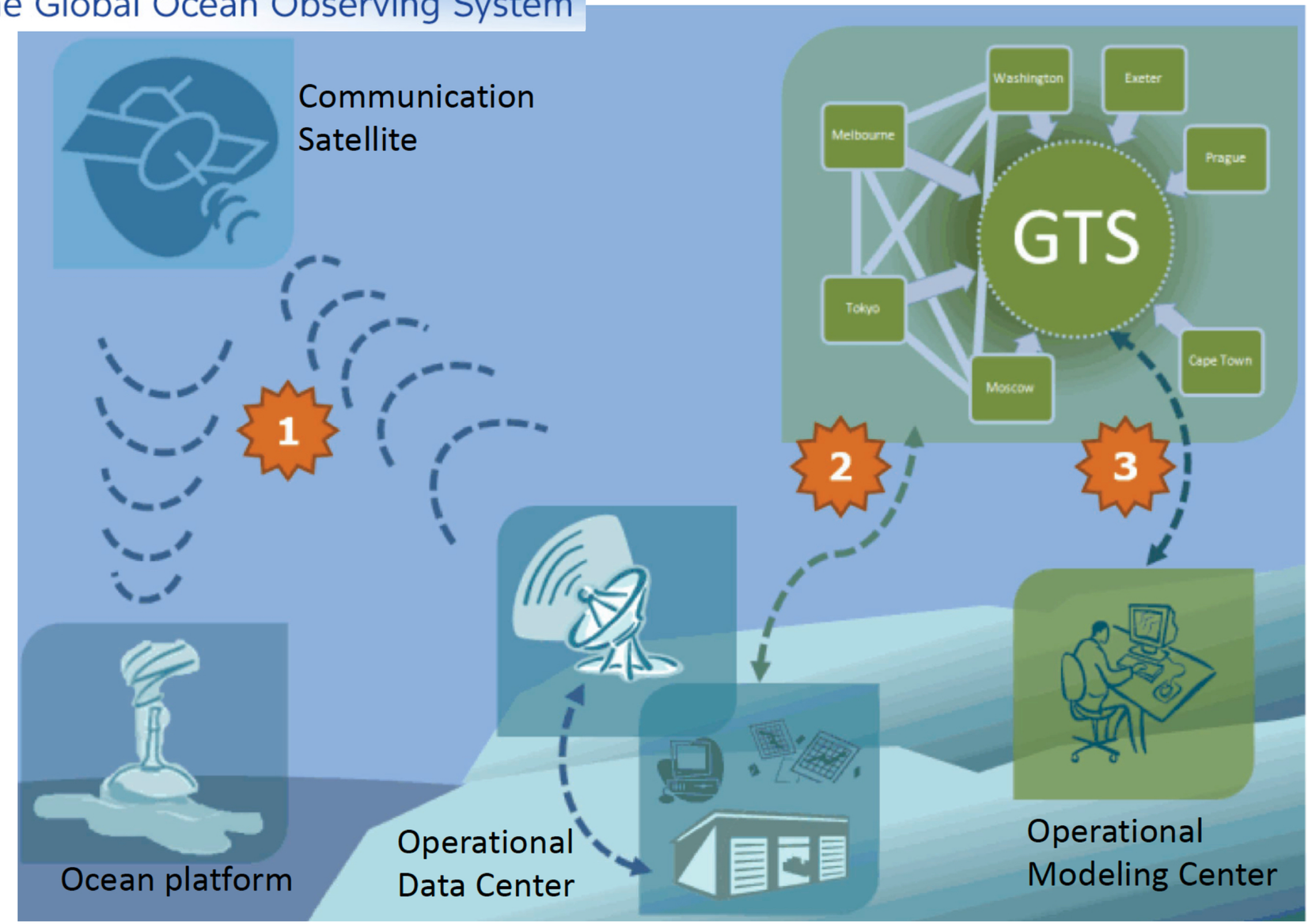
The global Argo float network



https://argo.ucsd.edu/


1 Float deployment
2 Descent to drifting depth
3 Drifting 10 days
4 Descent to profiling depth (2000 - 6000m)
5 Ascent: measuring Essential Ocean Variables
6 Data transmission
7 Starting next cycle
Argo data
0m
Drifting depth
1000m
Profiling depth
2000m
e.g measure 2
e.g measure 1
0
500
1000
1500
2000
Depth (in m)
Temperature (in °C)
Salinity (in psu)
© Thomas Haessig

The Global Ocean Observing System



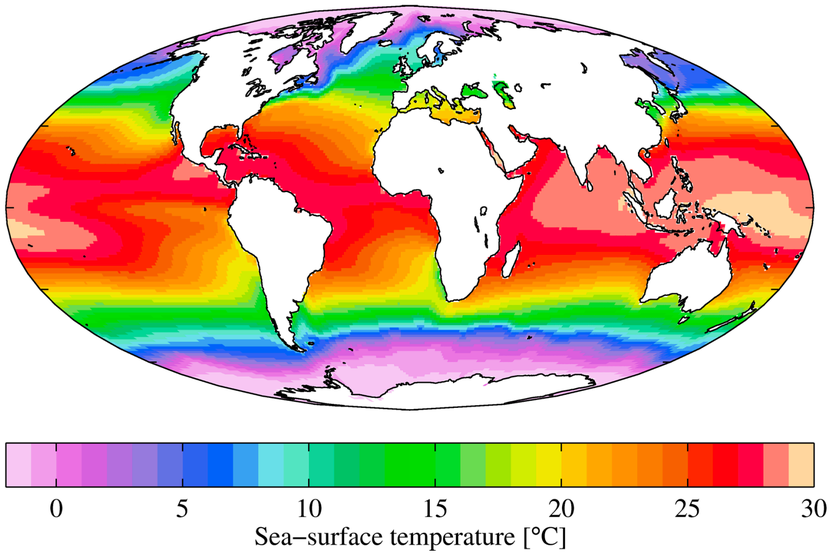
1) From platforms to satellites to data centres

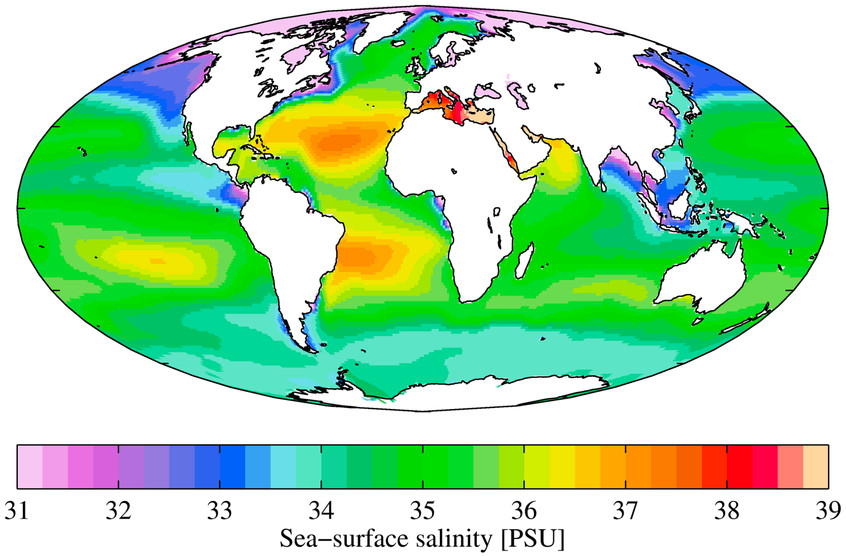
2) From data centres to the Global Telecommunications System (GTS)

3) From the GTS to a user (e.g. a modelling centre)

To see what's on GTS: https://www.ocean-ops.org/

What is a climatology?
A gridded estimate of the long-term mean ocean state





World Ocean Atlas (WOA)

Produced using *objective analysis* at 1° horizontal resolution

Vertical data on intervals from the surface to the seafloor (5500 m)

Available on annual, seasonal, and monthly timescales

How are climatologies constructed?

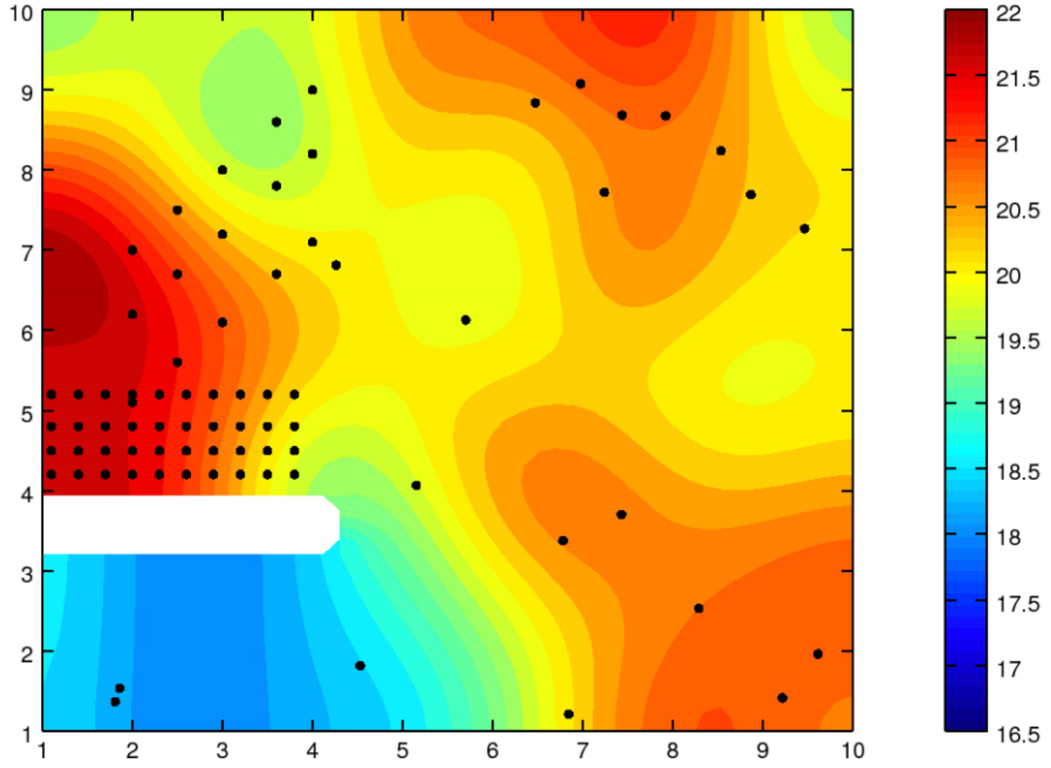
Traditional objective analysis in oceanography: estimating a scalar or vector field by applying the ordinary least squares method. This minimises the overall difference between the dataset and the statistical model.

The result is a statistical estimator with the following properties:

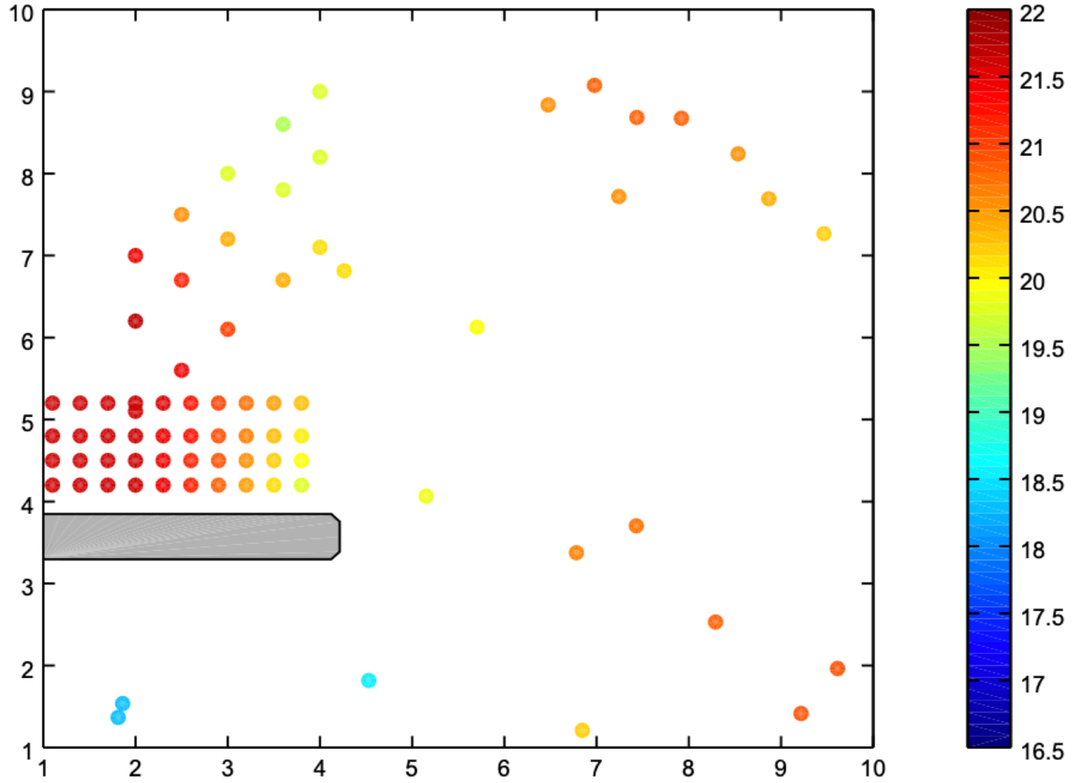
- Unbiased (i.e. expected value of distribution matches the true value)
- Linear in the data
- Minimises the variance, given the expected value and covariance of both the data and the estimated field

McIntosh (1990)

How are climatologies constructed? One method: optimal interpolation



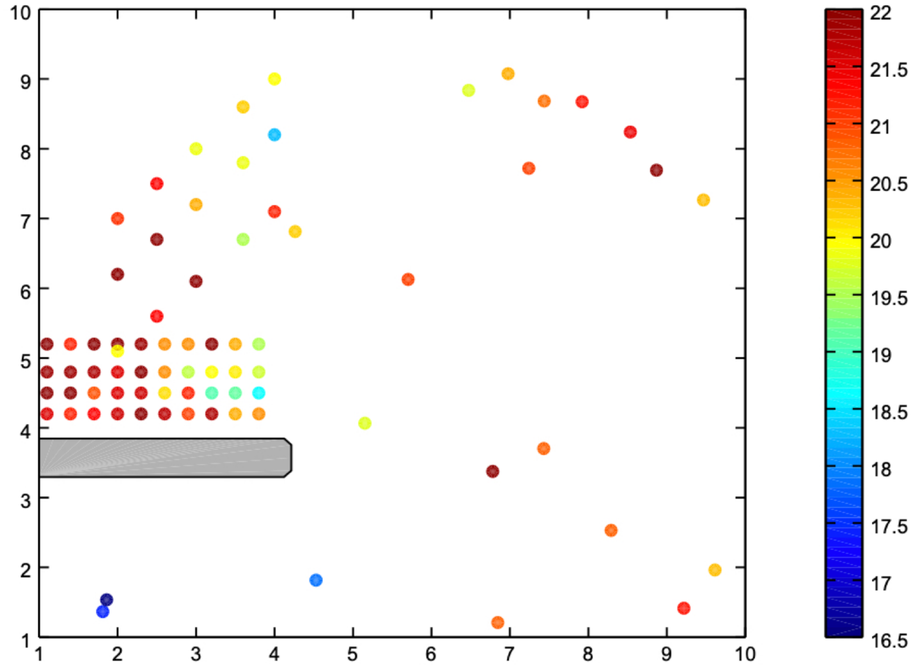
"True" field with sampling locations



Sampling-based view of field

Barth et al. (2008)

How are climatologies constructed? One method: optimal interpolation



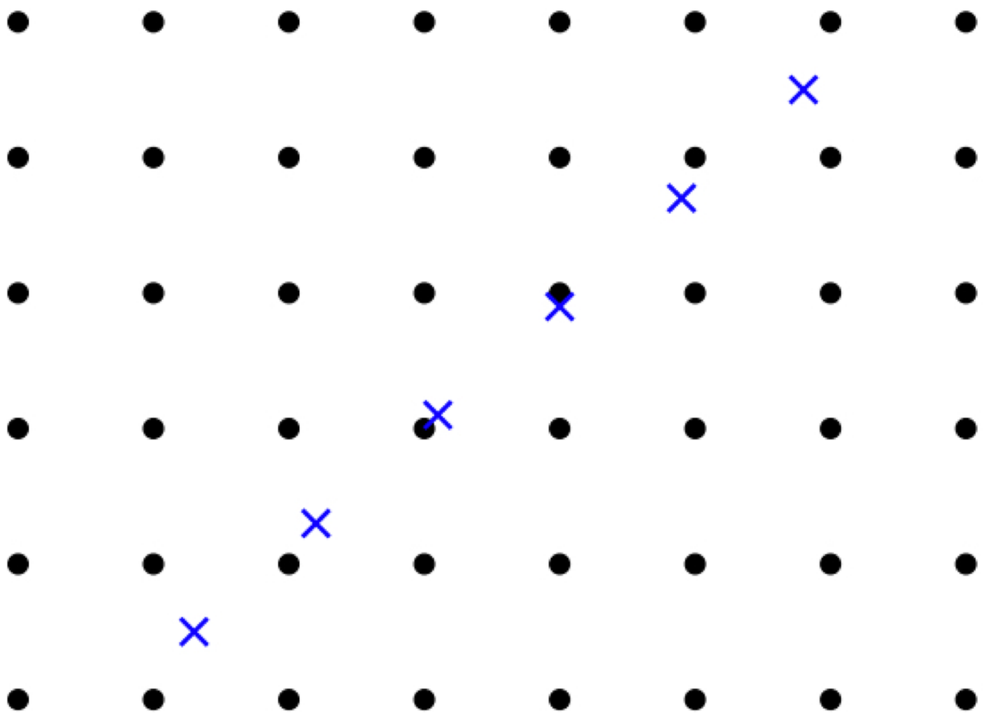
Observations with errors

Potential observation errors:

- **Instrumental errors** (limited precision or sensor bias)
- **Representative errors** (e.g. we want a monthly mean value when all we have is an instantaneous value)
- **Time sampling errors** (all observations are not taken at the same time)
- **Other errors** (e.g. human errors, transmission errors, malfunctioning of the instrument)

Barth et al. (2008)

How are climatologies constructed?
One method: optimal interpolation



First guess state vector (background)

True field

Error of the first guess

$$\mathbf{x}^b = \mathbf{x}^t + \eta^b$$

$$\mathbf{y}^o = H\mathbf{x}^t + \epsilon$$

Observations

Matrix *H* returns interpolated field

Error of the observations

The operator H(·) interpolates gridded field to the locations of the observations. This figure shows schematically the position of measurements (crosses) and the gridded field (dots). In its simplest form, H(·) performs a bilinear interpolation.

Barth et al. (2008)

The optimal interpolation (OI) scheme can be derived as the Best Linear Unbiased Estimator (BLUE) of the true state $\mathbf{x}^t$ which has the following properties:

- The estimator is linear in $\mathbf{x}^b$ and $\mathbf{y}^o$
- The estimator is not biased:

$$E\left[\mathbf{x}^a\right] = \mathbf{x}^t \tag{1.18}$$

- This estimate has a minimal total variance *i.e.* no other estimator would have an error variance lower that the BLUE estimator.

The only unbiased linear combination between $\mathbf{x}^b$ and $\mathbf{y}^o$ is the following:

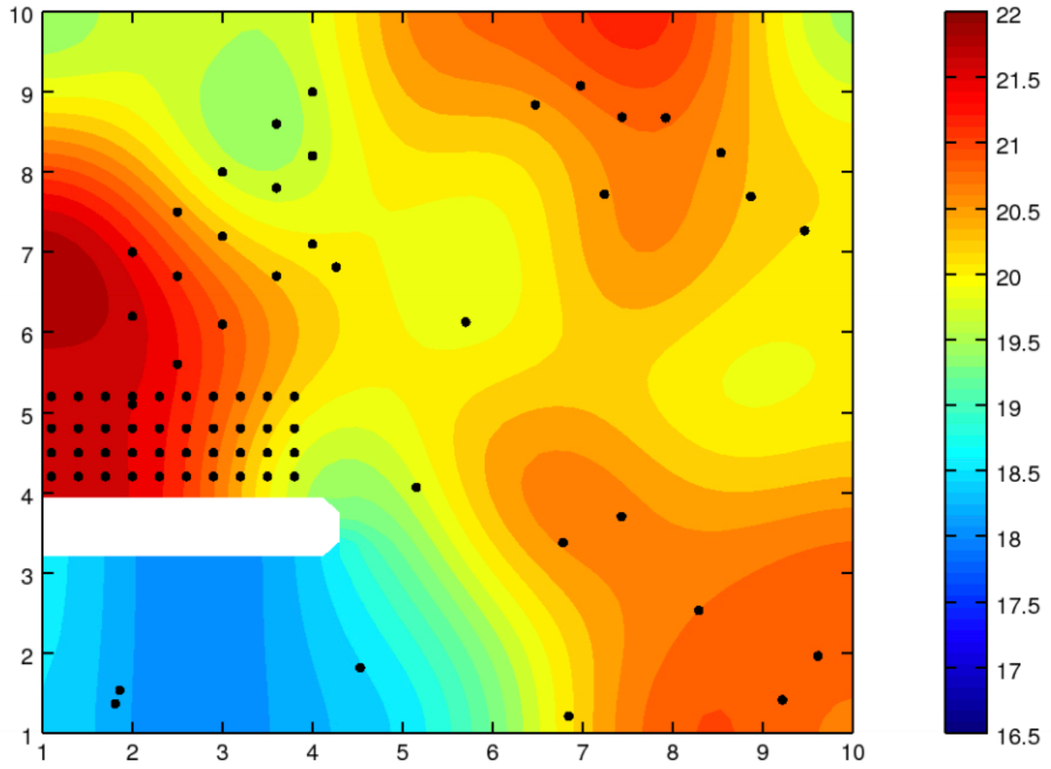
$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}\left(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b\right) \tag{1.19}$$

where the matrix $\mathbf{K}$ is the called the "Kalman gain". This matrix represents the gridding operation or "analysis" since it returns a gridded field when it is applied to a vector of observations minus the first guess. By subtracting the true state $\mathbf{x}^t$ from this equation, one can derive how an error on the first guess or on the observations affects the analysis:

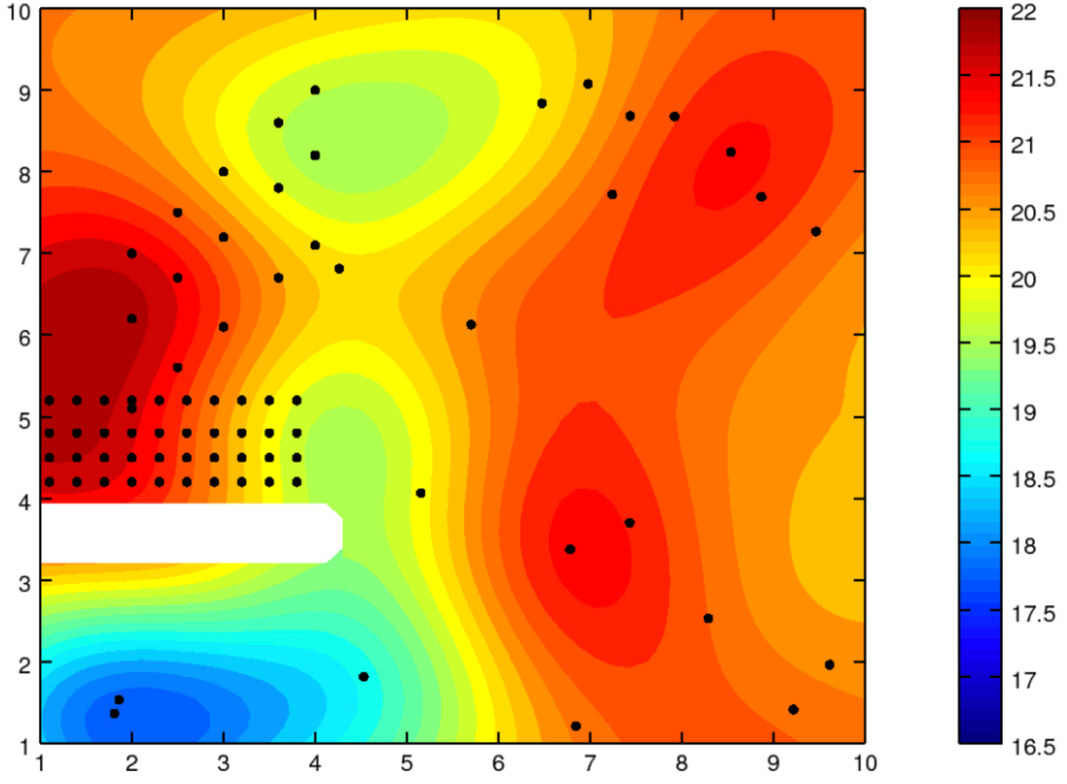
$$\boldsymbol{\eta}^a = \boldsymbol{\eta}^b + \mathbf{K}\left(\boldsymbol{\varepsilon} - \mathbf{H}\boldsymbol{\eta}^b\right) = \left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)\boldsymbol{\eta}^b + \mathbf{K}\,\boldsymbol{\varepsilon} \tag{1.20}$$

Barth et al. (2008)

How are climatologies constructed?
One method: optimal interpolation



"True" field with sampling locations



Optimally-interpolated field

Barth et al. (2008)

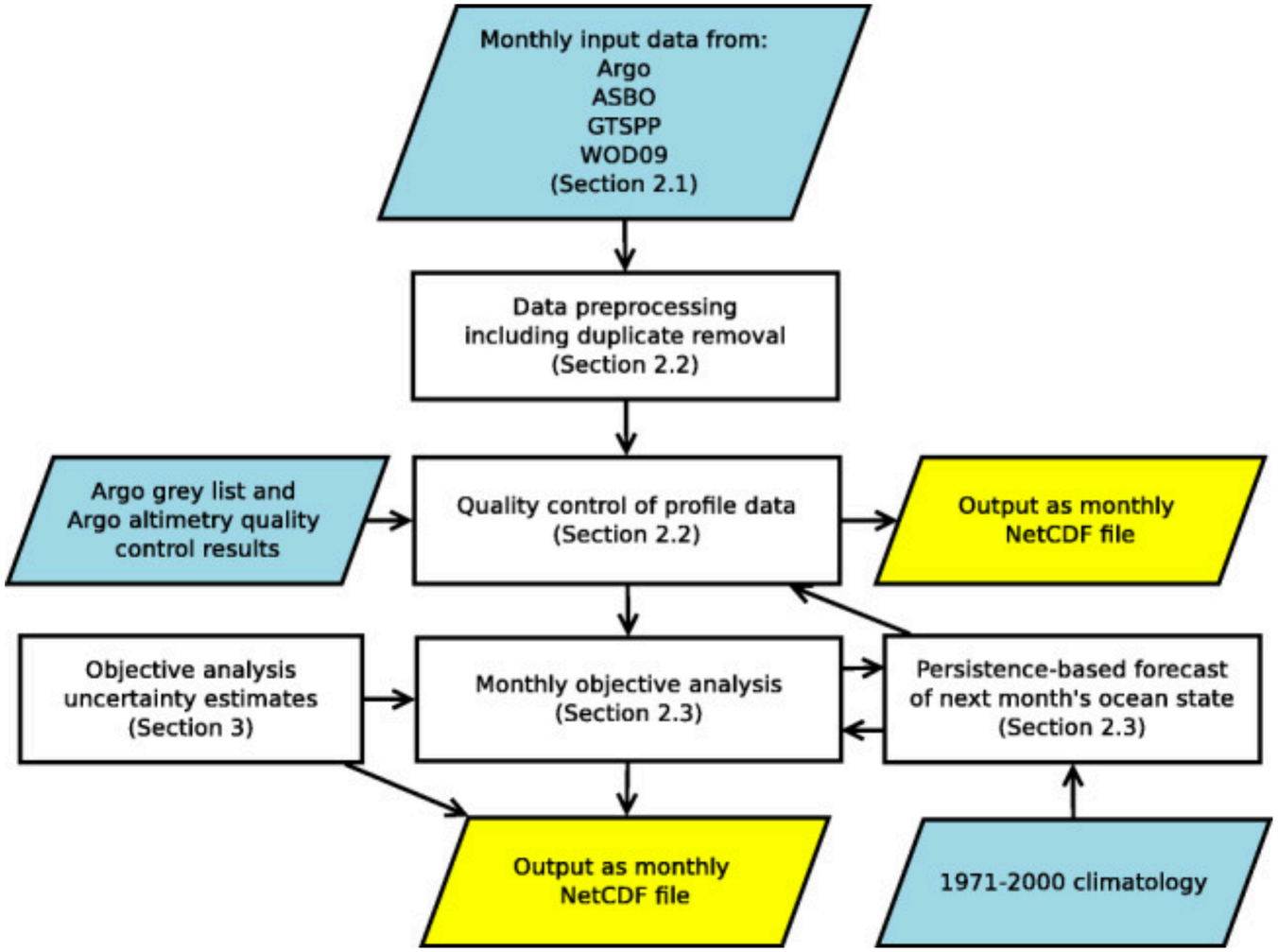
EN4: a climatology by the UK Met Office

Uses smoothed objective analysis to make a climatology

Uses “persistence” as a starting point (i.e. no change)

Assimilate data using optimal interpolation (OI)

Estimates will relax to climatology in the absence of observations



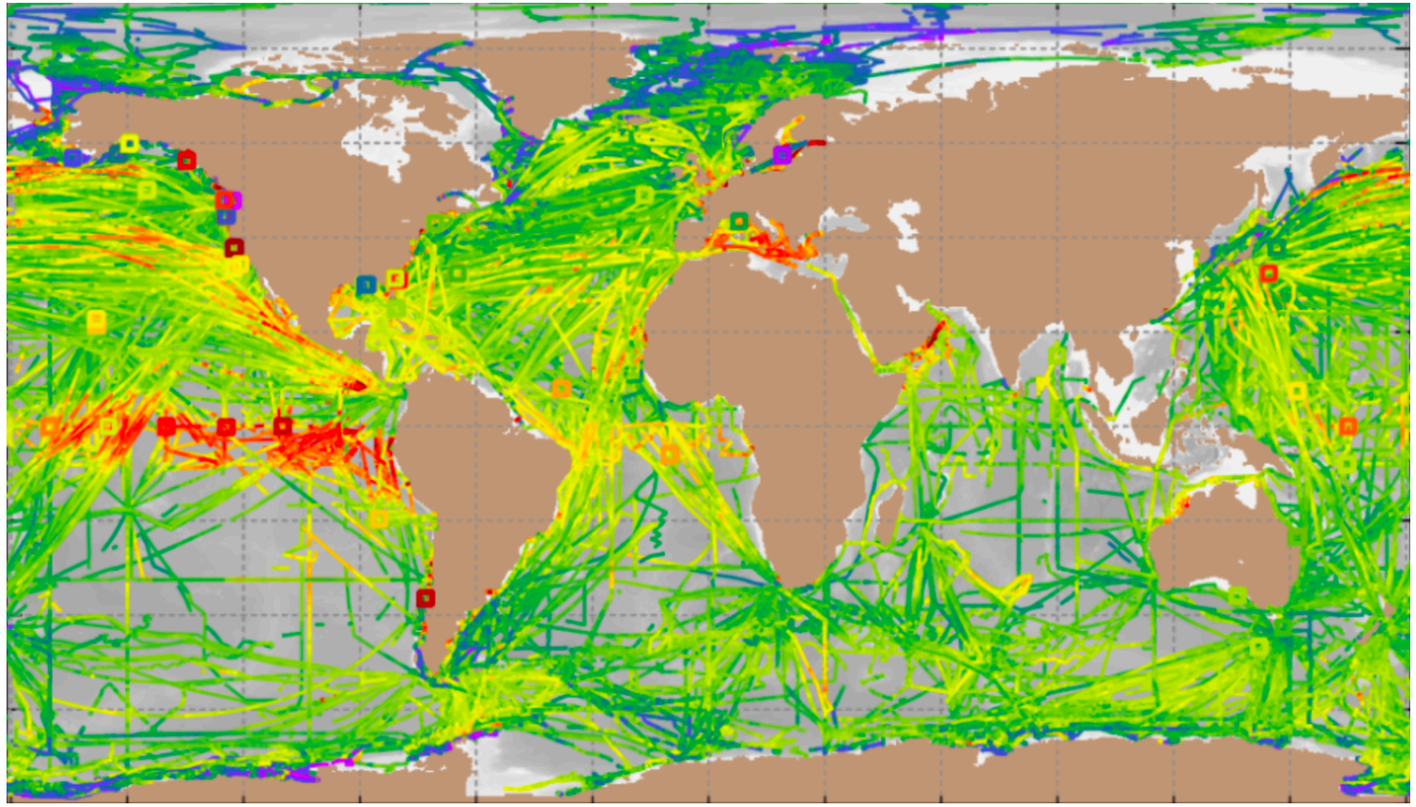
Good et al. (2013)

The Surface Ocean CO_2 Atlas (SOCAT)
One of the datasets we'll use for the challenge

SOCAT is a synthesis activity for quality-controlled, **surface ocean pCO_2** (partial pressure of carbon dioxide) observations by the ocean carbon research community (>100 contributors).

SOCAT enables quantification of the **ocean carbon sink** and ocean acidification and evaluation of ocean biogeochemical models.

The partial pressure of carbon dioxide in seawater (pCO_2) is the atmospheric concentration of CO_2 that would be **in equilibrium** with that seawater (no net transfer of molecules)



Locations of moorings and tracks of ships and drifters for all data in SOCAT version 2019 (1970-2019)

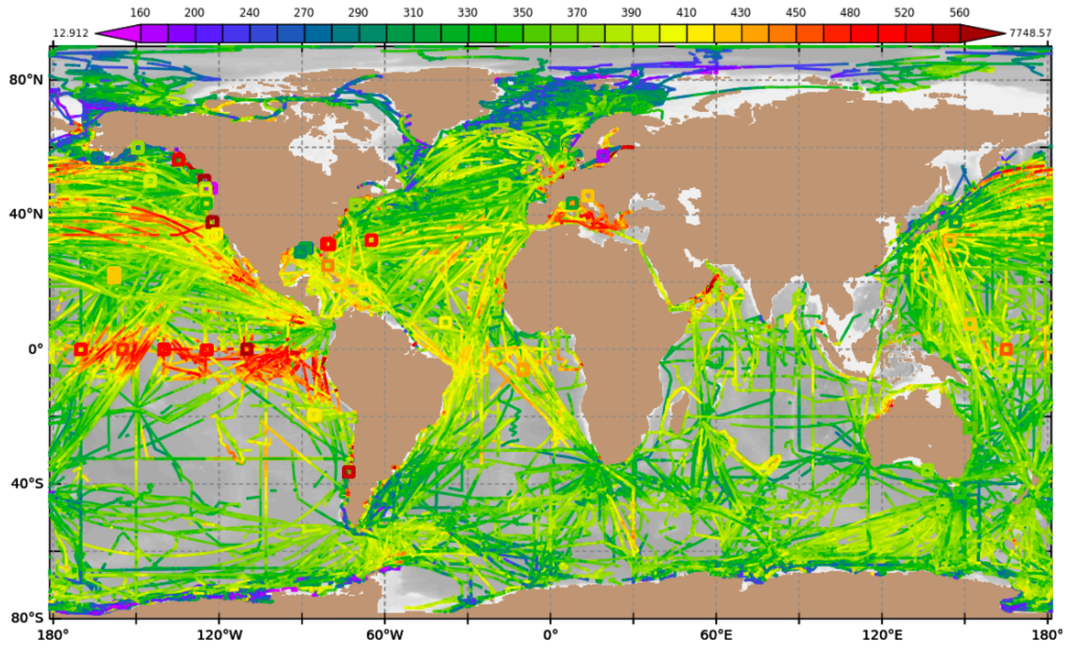
SOCAT
· SURFACE OCEAN CO2 ATLAS ·

What is a gridded product? SOCAT as an example

- Combine all data within each 1°x1° box for each month. Only include data with good quality control flags.

- Grid cells that have no measurements are not assigned a value

- Report an unweighted mean and a cruise-weighted mean, to account for different sampling frequencies on different ships. Also report minima, maxima, and standard deviations for each grid cell.

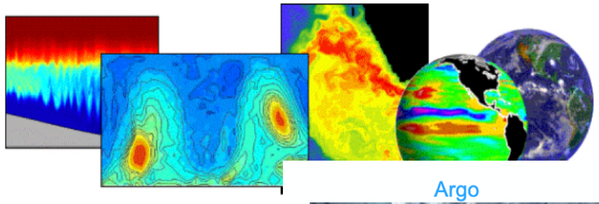


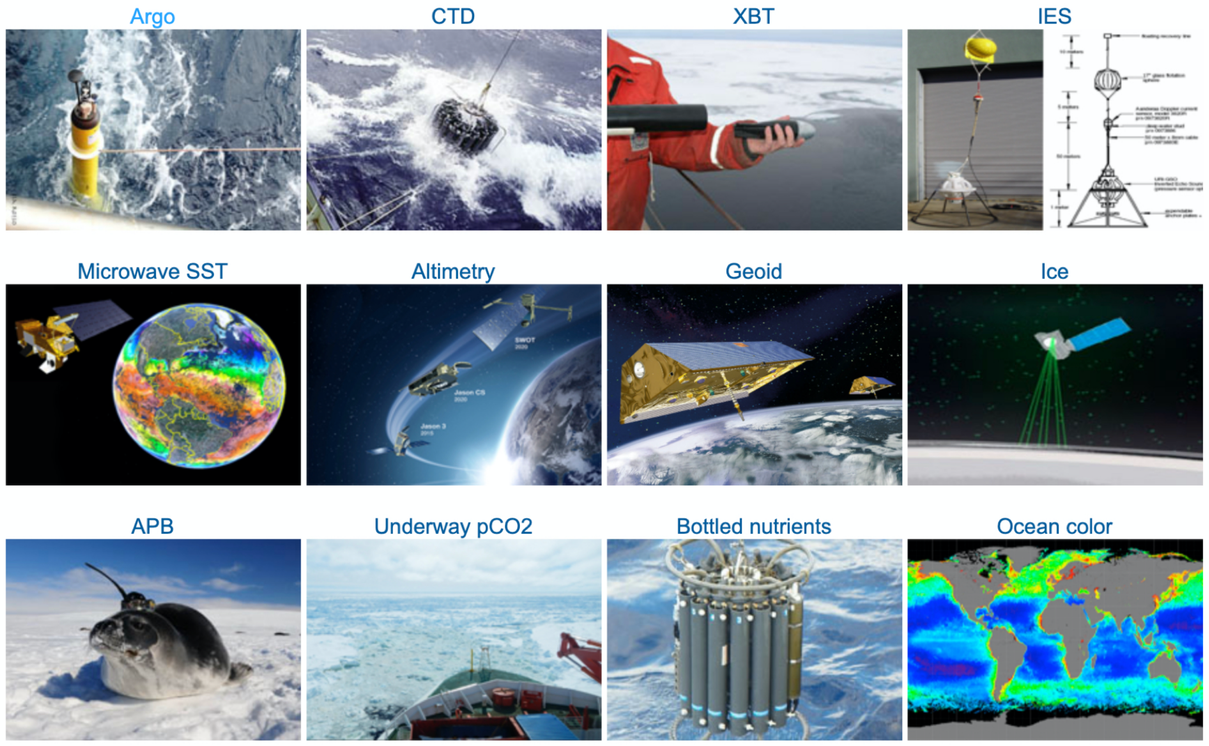
Sabine et al. (2012)
https://doi.org/10.5194/essd-5-145-2013

SOCAT
· SURFACE OCEAN CO2 ATLAS ·

What is a state estimate?

M.I.T. general circulation model (MITgcm)





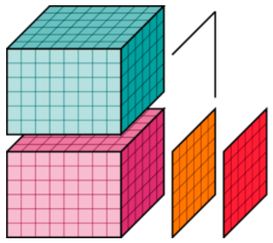
- A state estimate (like ECCO2 used in the data challenge) is an observationally-constrained numerical simulation that has been brought into consistency with a large suite of observations.
- By construction, state estimates feature **closed budgets of heat and salt**. This isn't the case for all reanalysis products in general!
- **Most reanalysis products use artificial sources and sinks of heat/salt in the interior.** Be aware!

NetCDF (FYI, not used in challenge)

NetCDF stands for Network Common Data Form and it is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. It is also a community standard for sharing scientific data.

NetCDF is maintained by Unidata, one of the University Corporation for Atmospheric Research (UCAR)'s Community Programs (UCP).

Unidata also supports and maintains netCDF programming interfaces for C, C++, Java, and Fortran. Programming interfaces are also available for Python, IDL, MATLAB, R, Ruby, and Perl.



xarray

xarray is a very useful python package for dealing with NetCDF data

https://euroargodev.github.io/

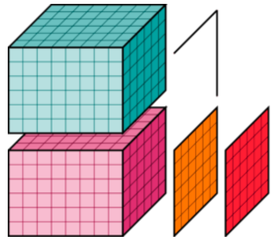
NetCDF (FYI, not used in challenge)

- **Self-Describing:** A netCDF file includes information about the data it contains.
- **Portable:** A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- **Sharable:** One writer and multiple readers may simultaneously access the same netCDF file.
- **Scalable:** Small subsets of large datasets in various formats may be accessed efficiently through netCDF interfaces, even from remote servers.
- **Appendable:** Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
- **Archivable:** Access to all earlier forms of netCDF data will be supported by current and future versions of the software

https://euroargodev.github.io/

NetCDF (FYI, not used in challenge)


```
<xarray.Dataset>
Dimensions:  (lat: 720, lon: 1440, time: 365)
Coordinates:
  * time     (time) datetime64[ns] 2019-01-01 2019-01-02 ... 2019-12-31
  * lat      (lat) float32 -89.875 -89.625 -89.375 ... 89.375 89.625 89.875
  * lon      (lon) float32 0.125 0.375 0.625 0.875 ... 359.375 359.625 359.875
Data variables:
    sst      (time, lat, lon) float32 ...
Attributes:
    Conventions:    CF-1.5
    title:          NOAA/NCEI 1/4 Degree Daily Optimum Interpolation Sea Surf...
    institution:    NOAA/National Centers for Environmental Information
    source:         NOAA/NCEI https://www.ncei.noaa.gov/data/sea-surface-temp...
    References:     https://www.psl.noaa.gov/data/gridded/data.noaa.oisst.v2....
    dataset_title:  NOAA Daily Optimum Interpolation Sea Surface Temperature
    version:        Version 2.1
    comment:        Reynolds, et al.(2007) Daily High-Resolution-Blended Anal...
```



xarray

https://euroargodev.github.io/

References

Carr, M.H., Woodson, C.B., Cheriton, O.M., Malone, D., McManus, M.A. and Raimondi, P.T. (2011), Knowledge through partnerships: integrating marine protected area monitoring and ocean observing systems. Frontiers in Ecology and the Environment, 9: 342-350. https://doi.org/10.1890/090096

Good, S. A., Martin, M. J., and Rayner, N. A. (2013), EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, J. Geophys. Res. Oceans, 118, 6704– 6716, https://doi.org/10.1002/2013JC009067

Jones, D. C., Ito, T., Takano, Y., and Hsu, W.-C. (2014), Spatial and seasonal variability of the air-sea equilibration timescale of carbon dioxide, Global Biogeochem. Cycles, 28, 1163– 1178, https://doi.org/10.1002/2014GB004813

McIntosh, P. C. (1990), Oceanographic data interpolation: Objective analysis and splines, J. Geophys. Res., 95(C8), 13529–13541, https://doi.org/10.1029/JC095iC08p13529

Reichl, B.G, & Deike, L. (2020). Contribution of sea-state dependent bubbles to air-sea carbon dioxide fluxes. Geophysical Research Letters, 47, e2020GL087267. https://doi.org/10.1029/2020GL087267

What drives the surface ocean carbon dioxide concentration?

Sea surface salinity (SSS)	Salt concentration at the surface
Sea surface temperature (SST)	Water temperature at the surface (relevant to solubility)
Mixed layer depth (MLD)	Depth beneath the surface over which the density is roughly uniform (this layer is well mixed, not stratified)
Chlorophyll-a concentration	Affects ocean colour, relevant for the inorganic and organic carbon concentrations in the surface layer
Sea surface height	Height of the sea surface relative to the geoid (i.e. the ocean surface if gravity and rotation were the only effects)
Atmospheric CO_2 mole fraction	A measure of atmospheric carbon dioxide concentration, relevant for air-sea carbon fluxes