and visualising

Managing large volumes of environmental data: the work of the Reading e-Science Centre

Alastair Gemmell

Environmental Systems Science Centre (ESSC) &

Reading e-Science Centre (ReSC),

University of Reading

ECMWF workshop, Reading 3rd Nov 2009

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NATURAL ENVIRONMENT RESEARCH COUNCIL
ENVIRONMENTAL SYSTEMS SCIENCE CENTRE

Reading e-Science Centre

Many Thanks to:

Jon Blower, Keith Haines (ESSC)

Martin Price, Andy Saulter (Met Office)

Quillon Harpham, Keiran Millard (HR Wallingford)

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

NATURAL ENVIRONMENT RESEARCH COUNCIL
ENVIRONMENTAL SYSTEMS SCIENCE CENTRE

Reading e-Science Centre

Talk Outline

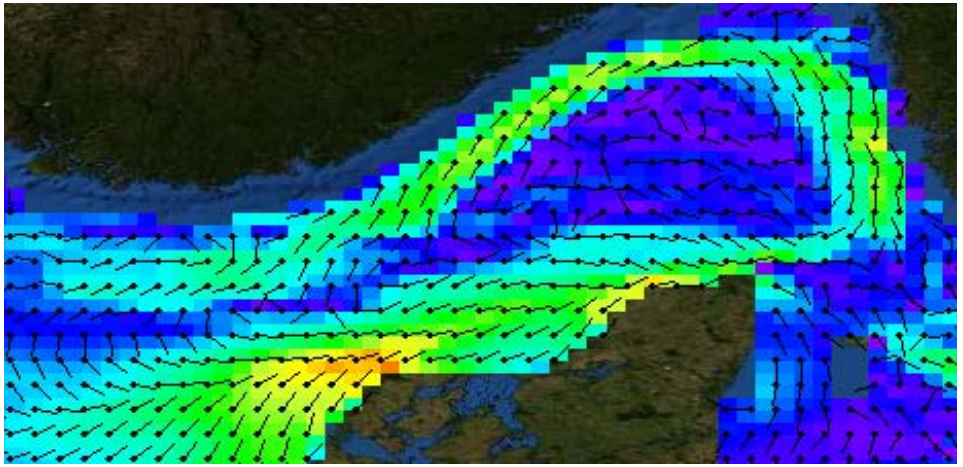
- Introduction to ESSC and ReSC
- Datasets that we handle
- Problems
- Tools and technologies to help
- Some solutions
- Conclusions

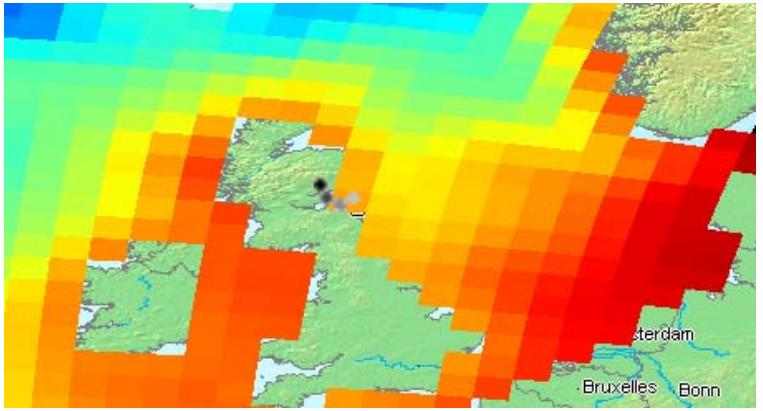
Introduction to ESSC & ReSC

- ESSC is the **Environmental Systems Science Centre** at the University of Reading
- ReSC is the **Reading e-Science Centre** which is hosted by ESSC.
- ReSC generally works with **marine** datasets. Director is Keith Haines who is an oceanographer.
- We also work with some atmospheric products including satellite data.
- We have experience of **managing** and **visualising** large datasets.

Datasets that ReSC handles: gridded data

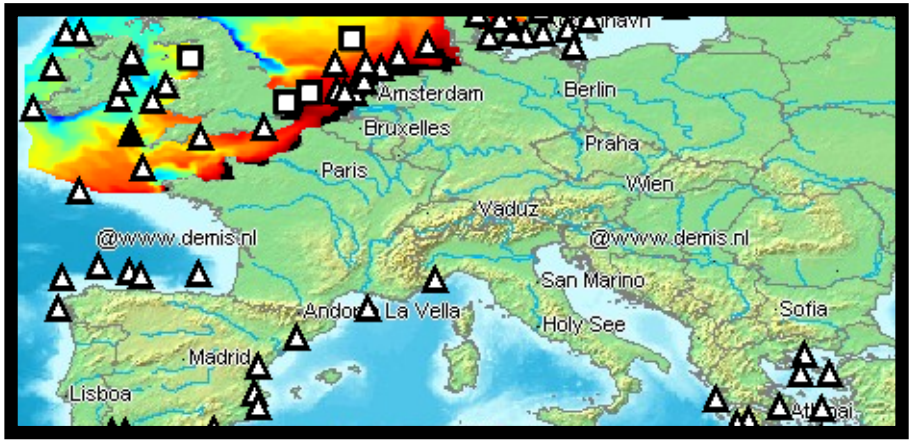
- A large amount of ocean **model** data, mostly NEMO output
- Also other model output and reanalyses from remote locations. Usually accessed via OPeNDAP
- **Satellite** data e.g. from Sciamachy. This may increase as ESSC is now part of National Centre for Earth Observation (NCEO)
- **Climatology**





Datasets that ReSC handles: in situ data

- Profile data from ENACT/ENSEMBLES EN3 dataset
  - ~10 million Temperature and Salinity profiles from 1950 to present
- Data from the Rapid array at 26N
- Argo float data.
- Data from the SEPRISE project – moorings around Europe

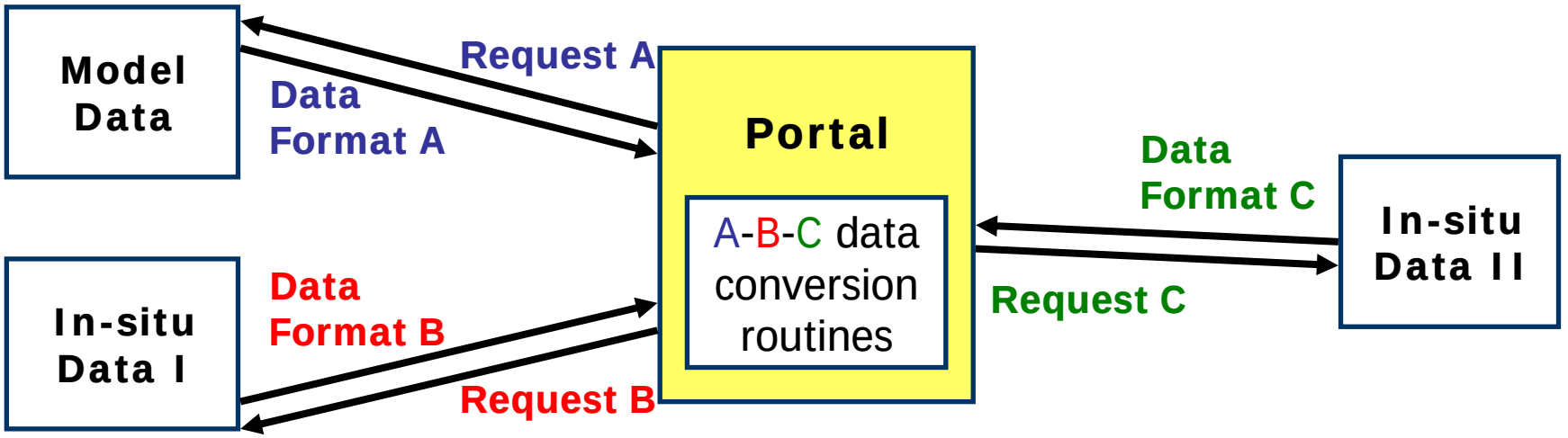


Problems in managing these datasets

- **Large volumes**
  - particularly acute for gridded model data, but becoming problematic for obs too
- **Dispersed data**
  - how best to combine and compare datasets resident in different locations?
- **Heterogeneity**
  - particularly problematic for observed data. Many formats in use. Inhibits interoperability (see next slide).

Heterogeneity can lead to lack of interoperability

- Each data provider chooses a way of serving their own data. May be formatted in one of a number of ways.
- Accessing the data may require many different forms of requests
- Future projects cannot easily employ these data access, analysis, visualisation routines, unless they happen to use same datasets



Tools and technologies to tackle the problems

- OGC Standards
  - Web Map Service (WMS)
  - Web Feature Service (WFS)
  - KML
  - [Styled Layer Descriptor (SLD) & Sensor Observation Service (SOS)]
- Climate Science Modelling Language (CSML)
- OPeNDAP / THREDDS
- Spatial Databases – e.g. PostgreSQL with PostGIS

OGC Web Services

Open Geospatial Consortium, Inc.

http://www.opengeospatial.org

- Open Geospatial Consortium
  - Body in charge of standards for ‘geospatial and location based services’
- WMS/WFS
  - WMS = Web Map Service. Serve geo-referenced images. Ideal for model output (also satellite data).
  - WFS = Web Feature Service. Serve geo-referenced points, lines, polygons. Ideal for in-situ observations, trajectories etc.
  - OGC has specifications for these services, allowing data to be served in a consistent manner.
  - Applications know what data format to expect and how to ask for it.
- ncWMS
  - OGC-compliant WMS for NetCDF developed at Reading e-Science Centre
  - Connects to ‘Godiva2’ web portal which uses OpenLayers to display data

THREDDS Data Server

http://www.unidata.ucar.edu/projects/THREDDS/

- THematic Realtime Environmental Distributed Data Services
- THREDDS enables data providers to serve NetCDF and similar data easily online via OPeNDAP protocol. Subsetting of data can be built into request.
- New version of THREDDS in collaboration with Reading e-Science Centre contains a version of ncWMS bundled in.
- Data can then be served by THREDDS either via OPeNDAP, or WMS.
- Diverse datasets held in different places can be read in via THREDDS servers at each institute.

CSML

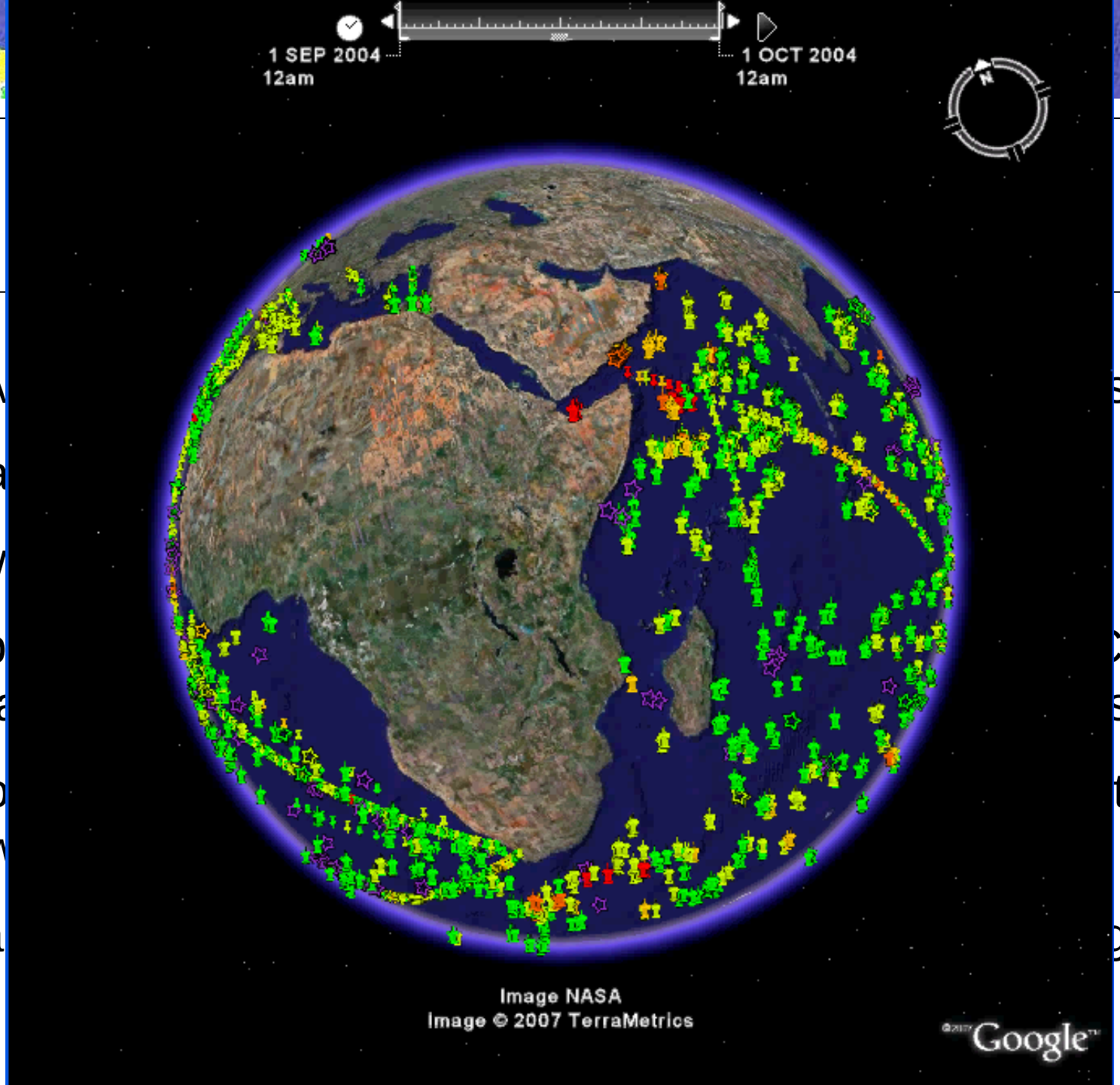
http://csml.badc.rl.ac.uk/

- Climate Science Modelling Language
- A standards-based way of representing data features pertinent to the Climate Sciences.
- 13 main feature types including profiles, trajectories, swaths, timeseries.
- Provides a common view onto datasets, independent of their storage format or physical location. Ideal for integrating diverse data products.
- We are developing a set of reusable Java libraries that embody the CSML concepts - can then apply these techniques to a number of other projects.

PostgreSQL with PostGIS

- PostgreSQL is most advanced open source database
- PostGIS is an open source extension to PostgreSQL giving geospatial capabilities
- PostGIS follows the OGC ‘Simple Features for SQL’ specification
- Addition of GEOS library enables searching within any polygon
- Good support for Java aids integration with web services

Some examples of ReSC solutions



- Jav
- Ma
- Tw
- Ob
- Mo
- Ca

files a

elsew

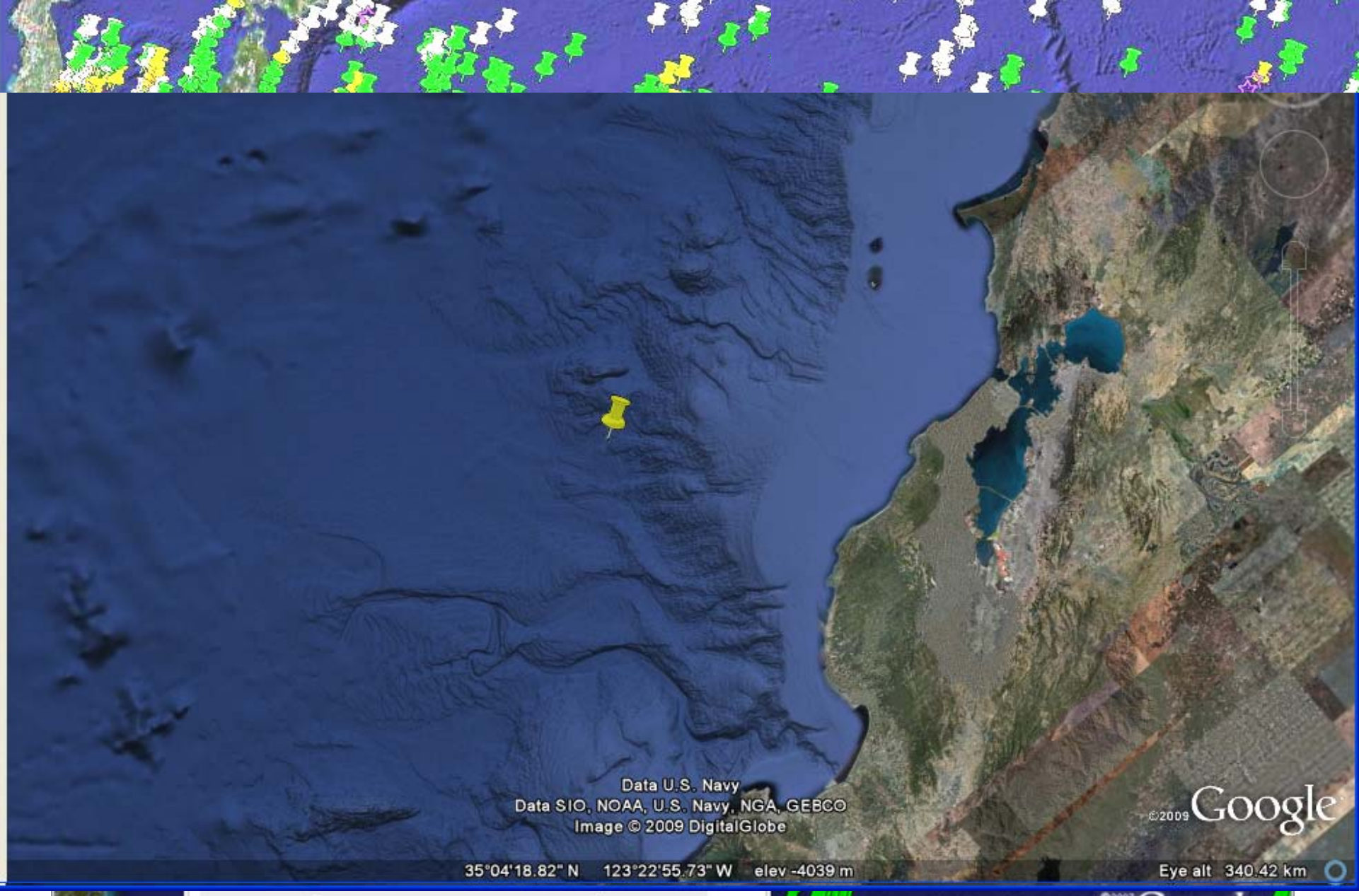
s

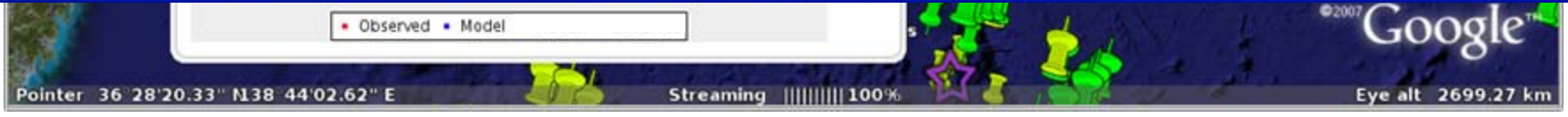
CDF

s.

thers

gy

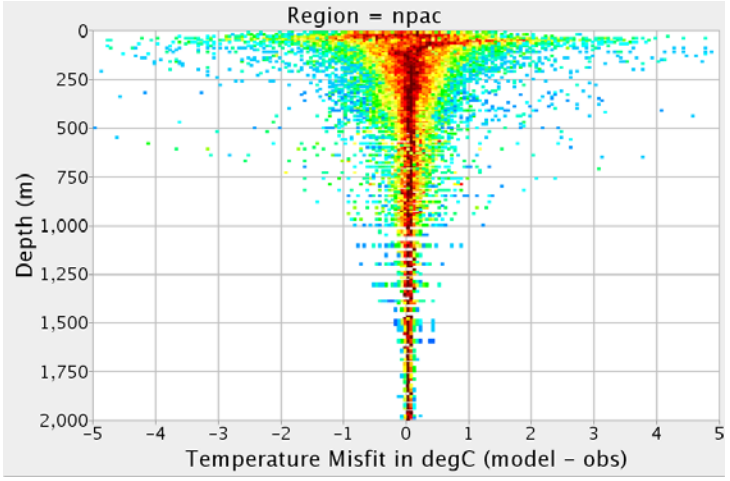

Data U.S. Navy
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image © 2009 DigitalGlobe
©2009 Google
35°04'18.82" N 123°22'55.73" W elev -4039 m
Eye alt 340.42 km


Observed
Model
©2007 Google
Pointer 36 28'20.33" N138 44'02.62" E
Streaming 100%
Eye alt 2699.27 km

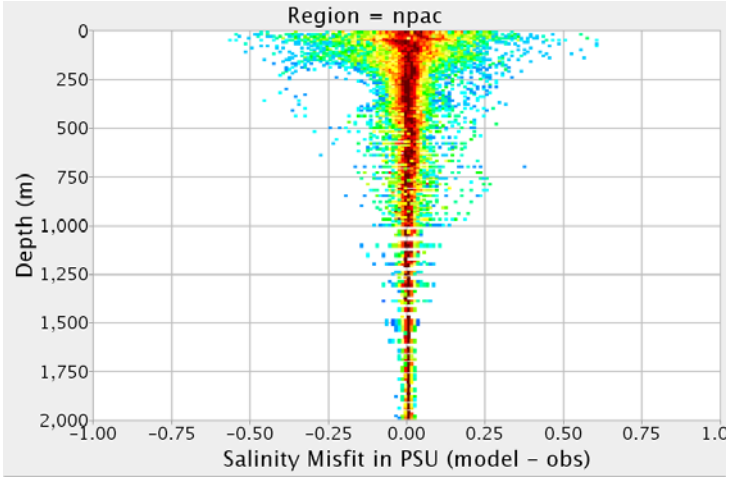
Statistical representation of the data

PDFs covering the north Pacific. Model is NEMO ¼ degree

Depth

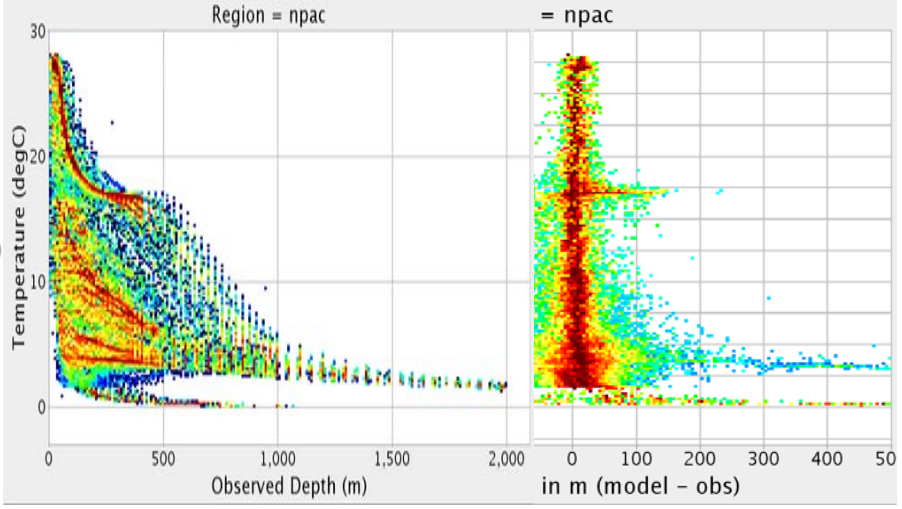


Temperature Misfit

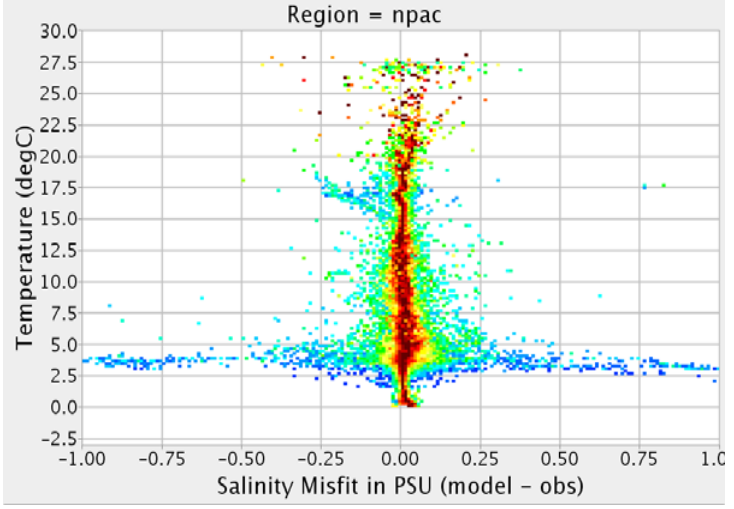


Salinity Misfit

Temp



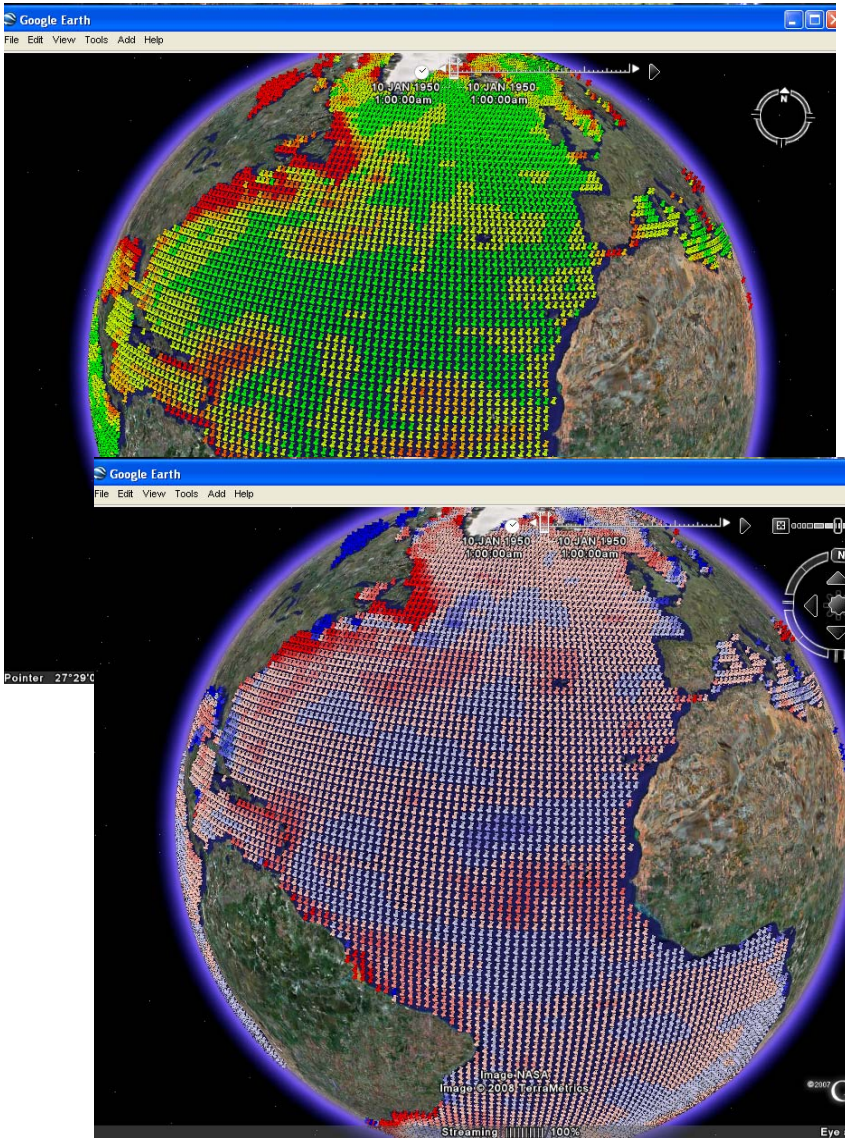
Observed Depth Depth Misfit

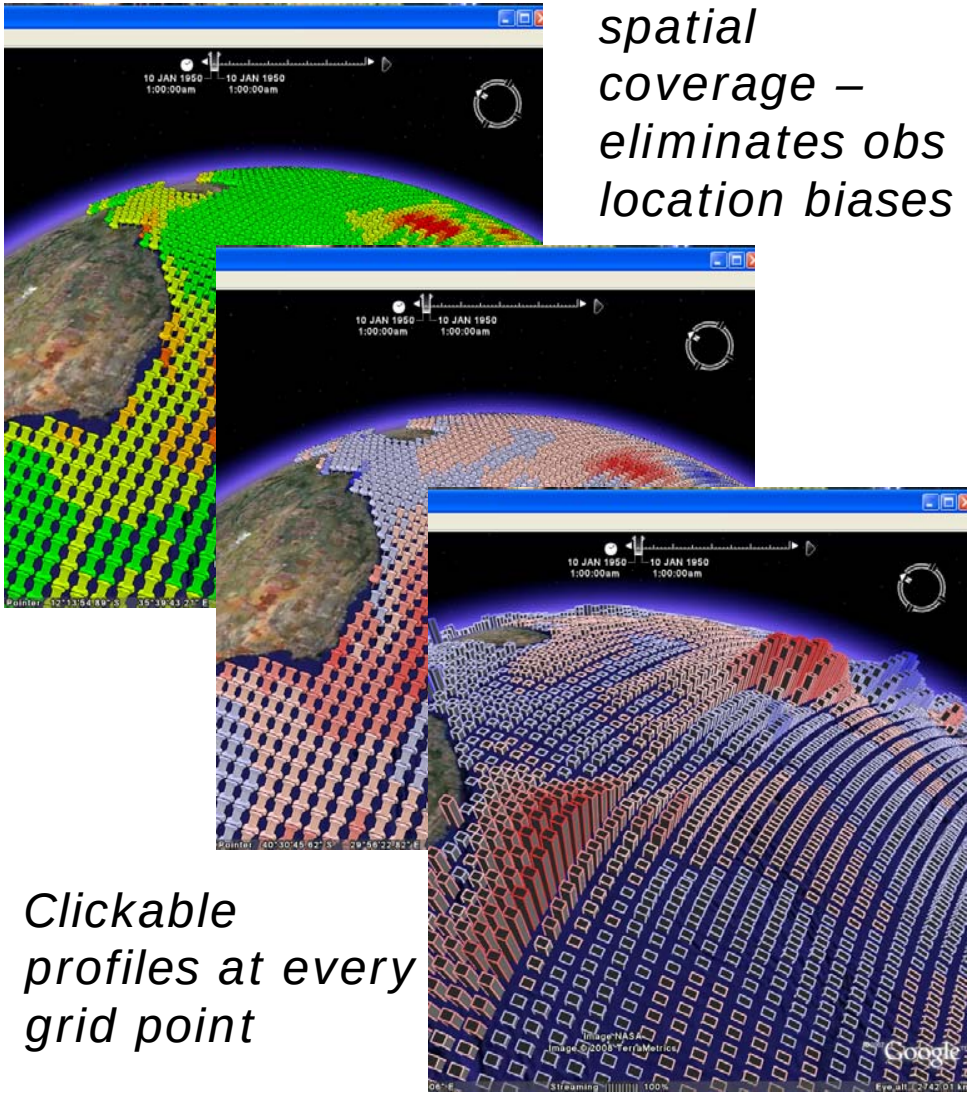


Salinity Misfit

Comparing Climatology with Models

World Ocean Atlas 2005

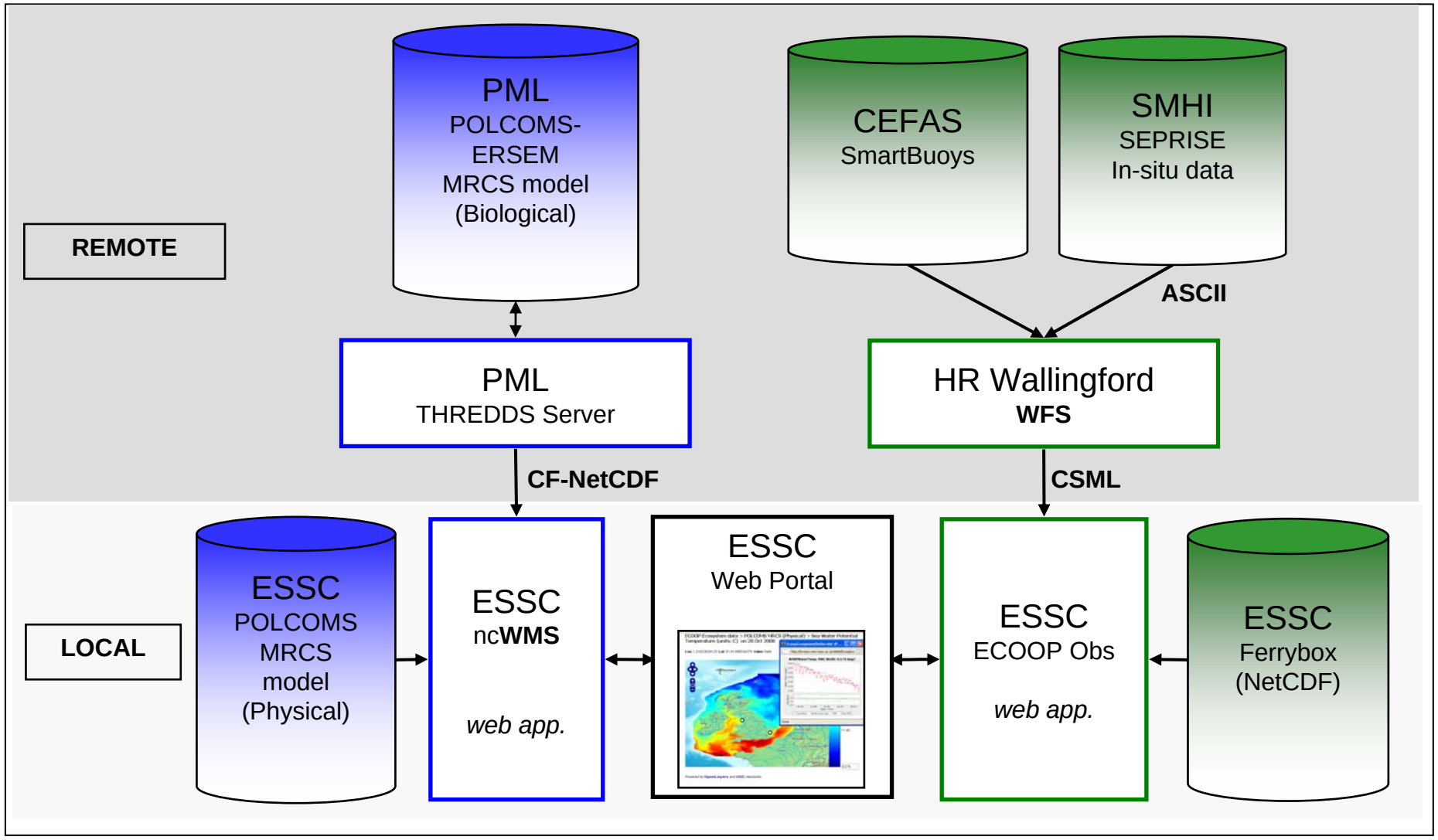




More uniform spatial coverage – eliminates obs location biases

Clickable profiles at every grid point

Examples – ECOOP Ecosystem Portal



ECOOP Obs

Home | About the portal | Conditions of use | Contributors

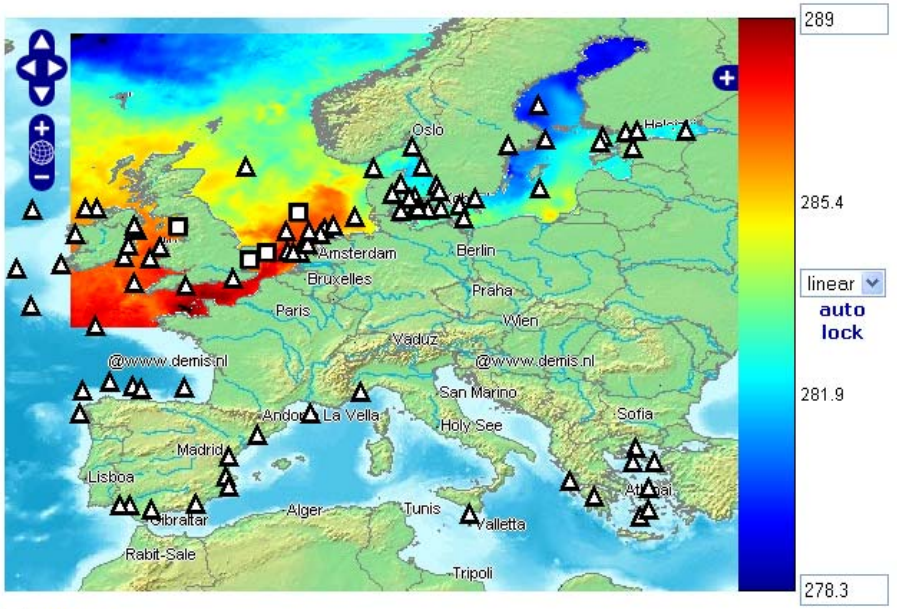
?	October, 2009					
«	‹	Today			›	»
Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Select date

- ECOOP Obs Data
  - ECOOP BOOS (Baltic)
    - DMI BALTIC BEST ESTIMATES
    - DMI Sea Surface Temperature
      - Sea Surface Temperature
    - BSH Best Estimates
  - ECOOP IBIROOS (Iberia-Biscay-Ireland)
  - ECOOP NOOS (North West Shelves)
  - ECOOP MOON (Mediterranean)
  - ECOOP BSGOOS (Black Sea)

ECOOP Obs Data > ECOOP BOOS (Baltic) > DMI Sea Surface Temperature > Sea Surface Temperature (units: kelvin) on 28 Oct 2009

Click on the map to get more information



- **CEFAS** SmartBuoy
- Pride of Bilbao Ferrybox (2008 data available, 2009 coming soon after a refit)
- **SEPRISE** Data

Powered by **OpenLayers** and **OGC** standards

ECOOP is funded by the European Commision's Sixth Framework Programme, under the priority Sustainable Development, Global Change and Ecosystems. Contract No. 36355.

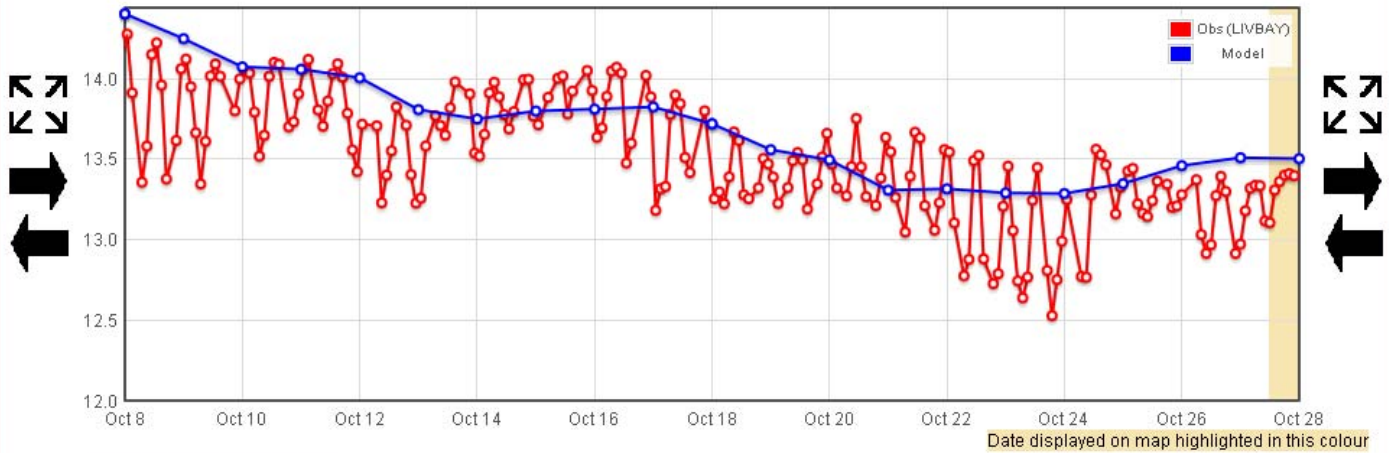
To give feedback on the portal, please use this **enquiry form** on the National Centre for Ocean Forecasting website.

Reading e-Science Centre | Met Office | Natural Environment Research Council | Cefas | Institute of Marine Research | HR Wallingford | TechWorks Marine | NIVA | PML Plymouth Marine Laboratory | National Oceanography Centre, Southampton

http://www.resc.reading.ac.uk/ecoop_ecosystem_portal

ECOOP Obs

Home | About the portal | Conditions of use | Contributors



Date displayed on map highlighted in this colour

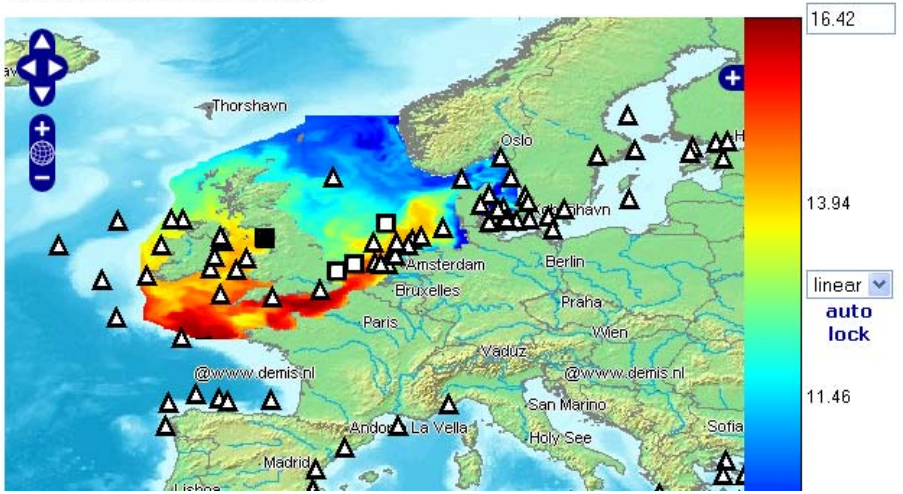
?	October, 2009					
«	‹	Today			›	»
Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Select date

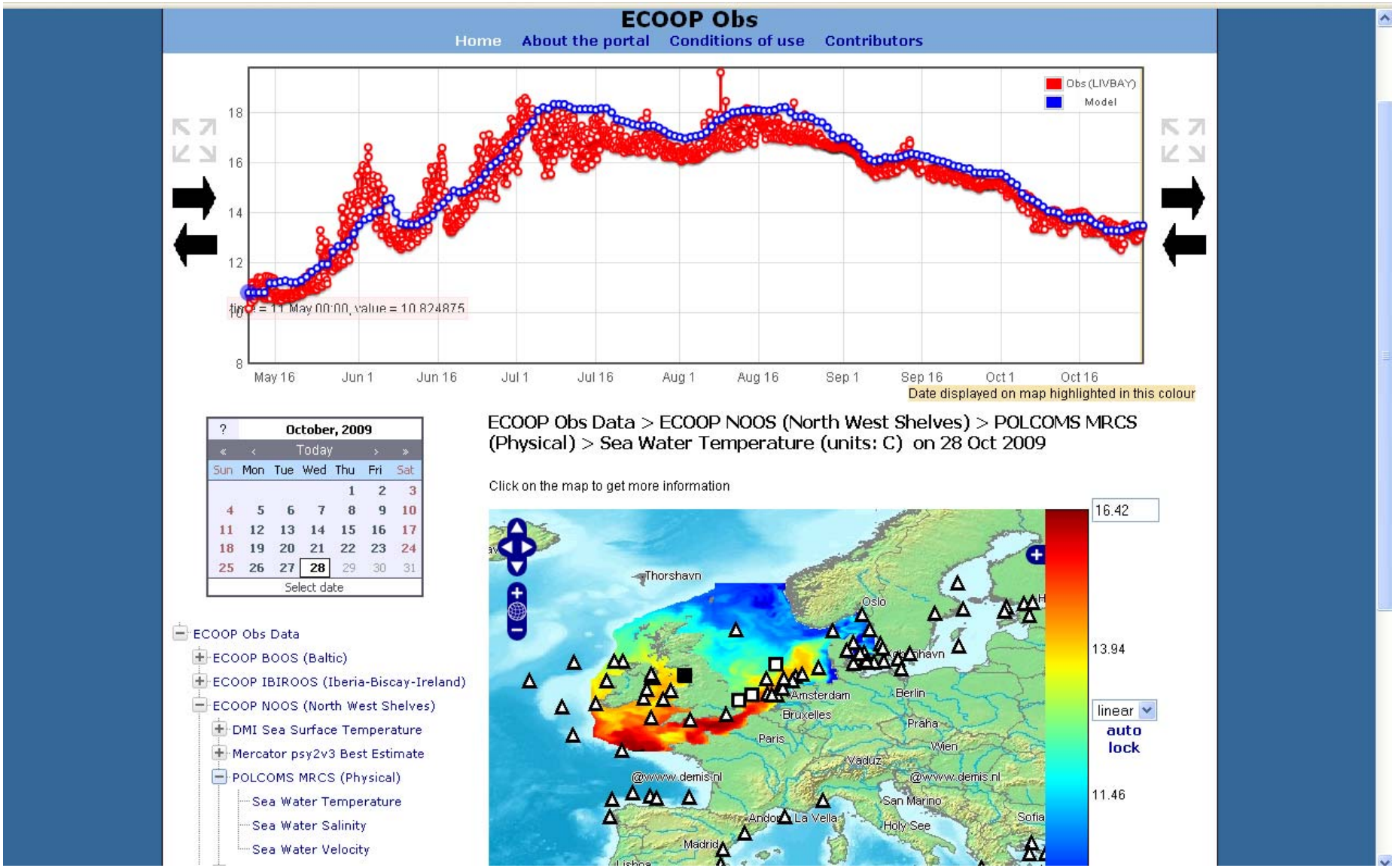
- ECOOP Obs Data
  - ECOOP BOOS (Baltic)
  - ECOOP IBIROOS (Iberia-Biscay-Ireland)
  - ECOOP NOOS (North West Shelves)
    - DMI Sea Surface Temperature
    - Mercator psy2v3 Best Estimate
    - POLCOMS MRCS (Physical)
      - Sea Water Temperature
      - Sea Water Salinity
      - Sea Water Velocity

ECOOP Obs Data > ECOOP NOOS (North West Shelves) > POLCOMS MRCS (Physical) > Sea Water Temperature (units: C) on 28 Oct 2009

Click on the map to get more information

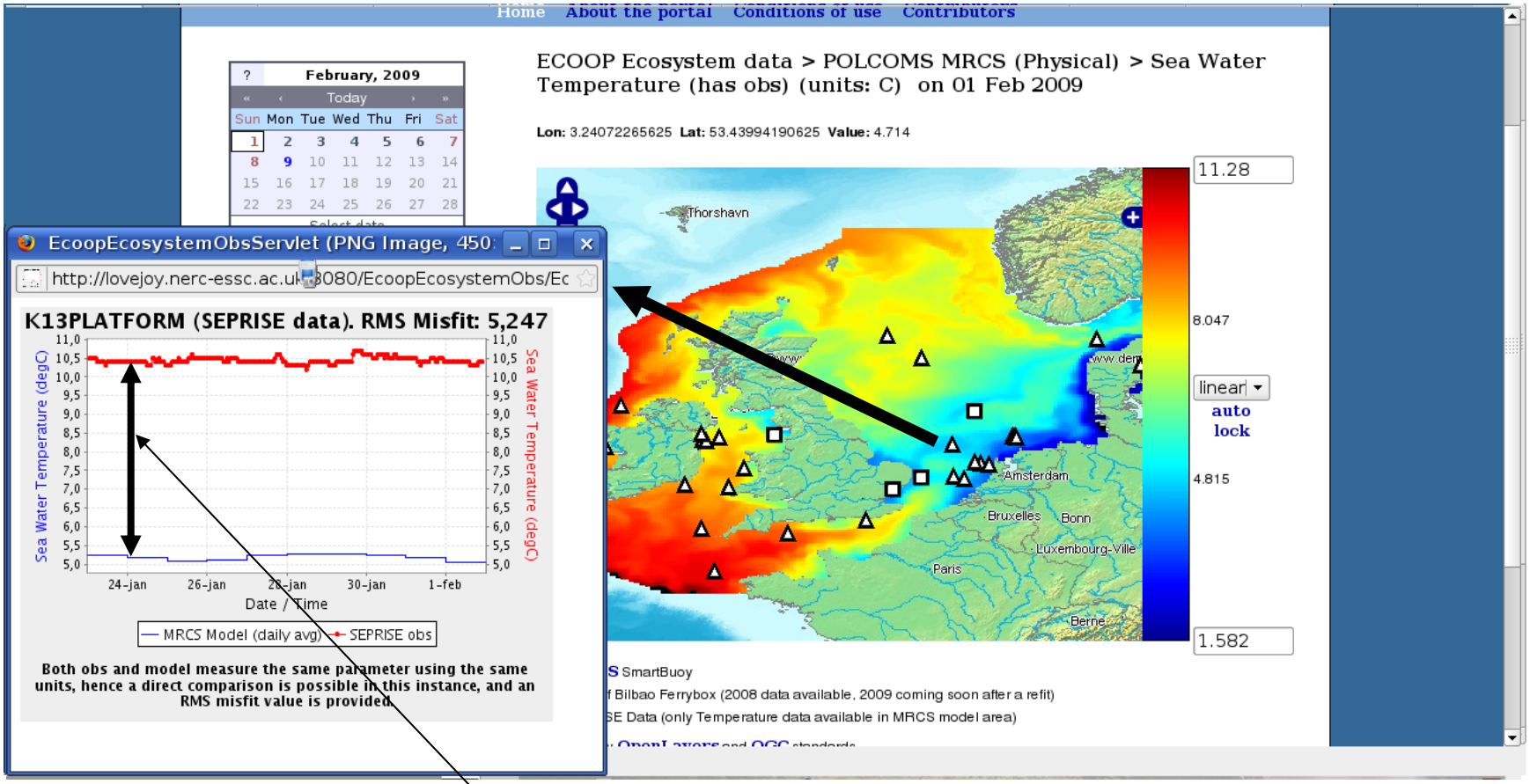


http://www.resc.reading.ac.uk/ecoop_ecosystem_portal



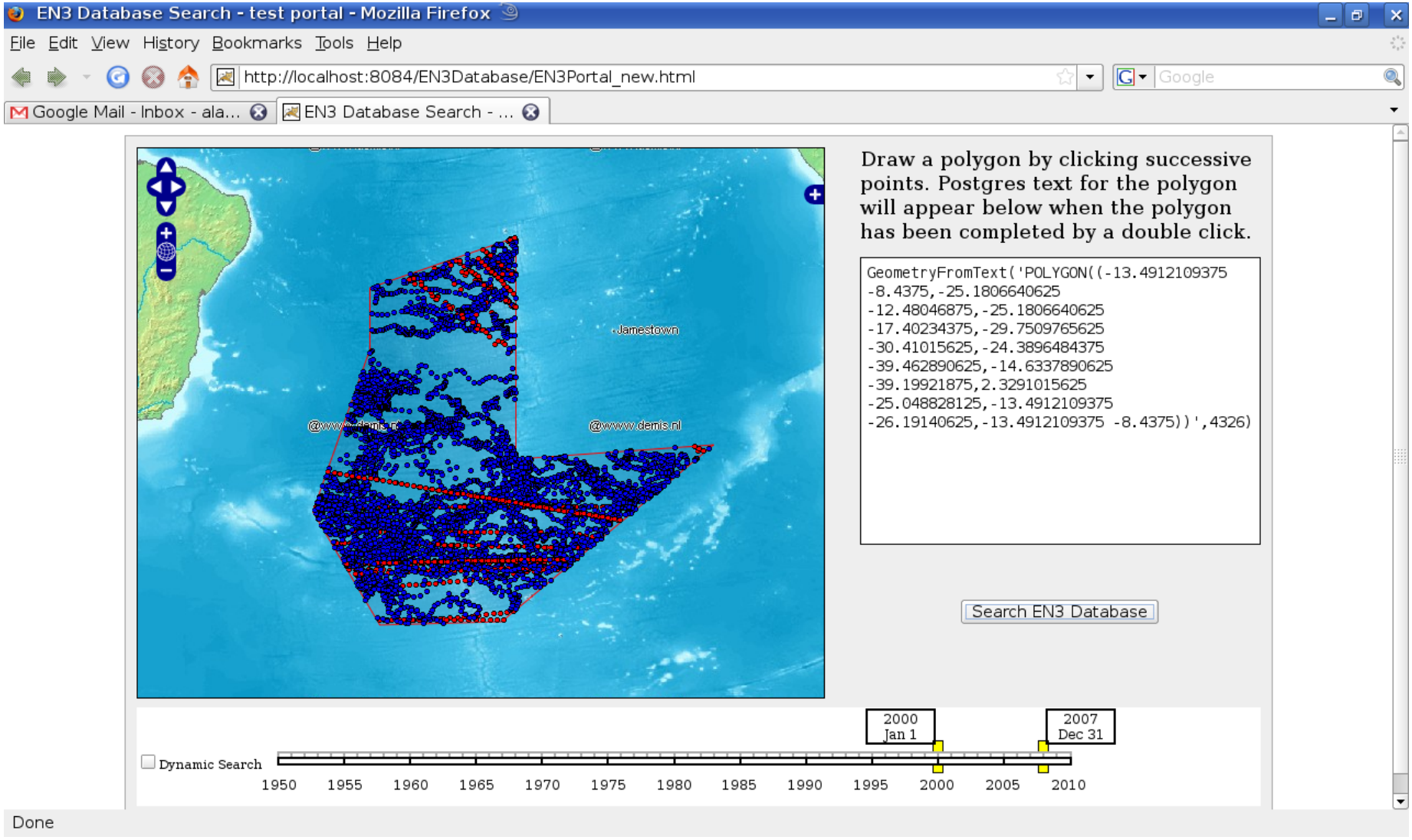
http://www.resc.reading.ac.uk/ecoop_ecosystem_portal

Comparing / co-plotting datasets can catch errors!



This looks like a suspiciously large and constant difference between obs and model

Some work in progress: searching and displaying 50 years of global in-situ data



Some work in progress: searching and displaying 50 years of global in-situ data

- We store profile metadata (date, lat, lot, id number) in a PostgreSQL/PostGIS database
- We keep actual data in original EN3 netCDF files.
- Use Java web application to communicate to both the database and the netCDF files.
- Some example times to retrieve profile metadata:
  - All profiles in North Atlantic in Sep 2008: **9 sec**
  - All profiles in the example polygon 2000 through 2007: **13 sec**
  - All profiles in Indian Ocean 1950-2000: **10 sec**

Concluding thoughts

Observations are heterogeneous. Particularly as volumes increase we need standards-based solutions for data management and visualization

Spatial databases can aid fast retrieval of observations when volumes increase. Can either store metadata alone, or metadata plus data.

Data format is an important consideration. Number of projects looking to standardise this in marine oommunity – e.g. SeaDataNet, OceanSITES.

A model such as CSML can help to abstract away some of the differences between formats if you end up working with multiple ones.

OPeNDAP can enable dynamic access to data from range of holdings over the web. We have paired this with THREDDS and ncWMS

(final!) Concluding thoughts

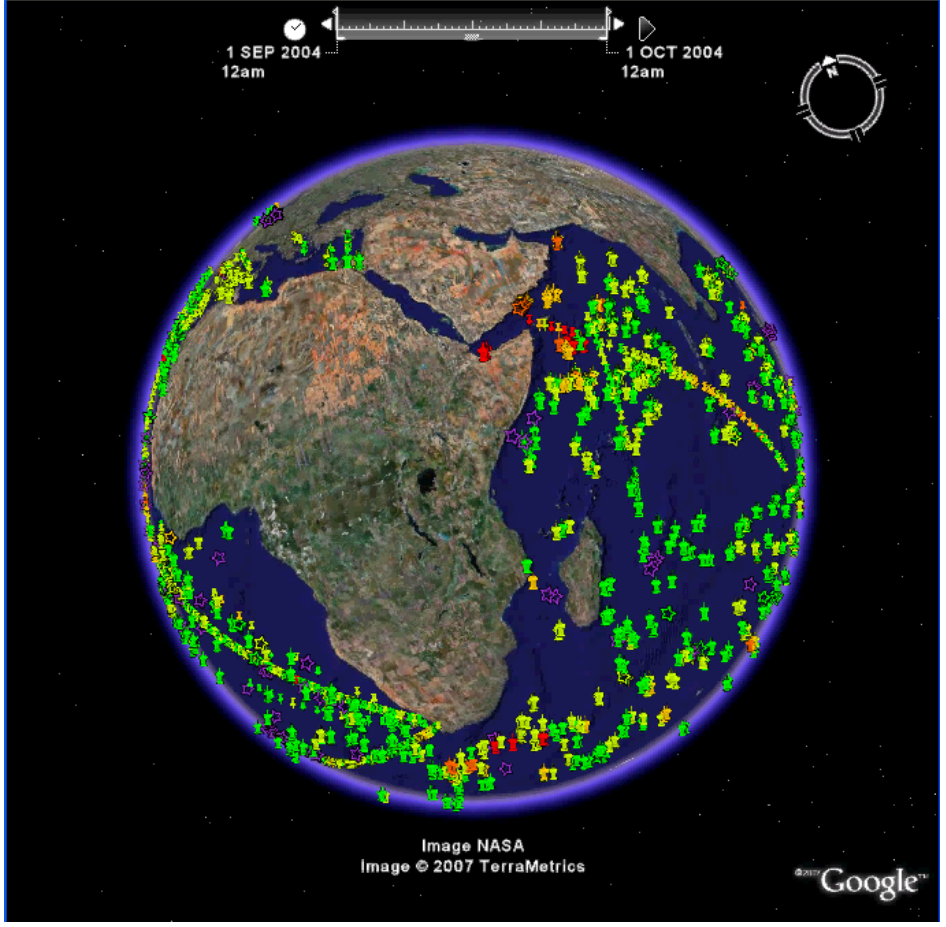
Visualization needs to be built on solid standards-compliant data management foundations.

In this case, it should then be 'simple' to develop the visualization system itself.

Sticking to open source and open standards gives greater flexibility, interoperability, and potential for reuse.

There is scope for harnessing power of tools such as Google Earth. Gives powerful viz platform and standard KML at cost of closed source.

I have spoken from an academic R&D type of perspective. Obviously one must get things right before employing in operational setting.



Thanks for your attention

a.l.gemmell@reading.ac.uk