Managing Large Volumes of Environmental Data: the Work of the Reading e-Science Centre

Alastair Gemmell, Jon Blower, Keith Haines – Reading e-Science Centre

In the workshop presentation, the title of the talk was actually adapted to read as "Managing and Visualizing Large Volumes of Environmental Data: the Work of the Reading e-Science Centre". This was to reflect the fact that the presentation would have been equally at home in the data management, or the visualization sessions, as well as the fact that at the Reading e-Science Centre (ReSC) the management and visualization of data are very much connected. As the presentation was formally in the data management session, and as this session was concerned mainly with observed data, the work presented focuses largely on the observed data handled by the Reading e-Science Centre.

1. The Reading e-Science Centre

The ReSC is hosted by the Environmental Systems Science Centre (ESSC) at the University of Reading. The director of ReSC is Keith Haines who works in oceanography, and hence the ReSC deals primarily, although not exclusively, with marine datasets. We also work with a number of atmospheric products including satellite data. At the ReSC we enjoy the benefits of collaborating with other academic and non-academic institutes, and the work that was presented was done with the help of a number of people external to ESSC - Andy Saulter and Martin Price at the Met Office, and Keiran Millard and Quillon Harpham at HR Wallingford chief among them.

2. ReSC datasets

The ReSC handles both gridded datasets, and point-based in-situ datasets. The former composed of model output, climatology, and satellite products. The model output is largely from the NEMO model (ref. 1) which is run locally by the marine group at ESSC. We also hold some output from models run at other centres. The satellite data held at ReSC may well increase in volume with the formation of the National Centre for Earth Observation (NCEO) of which ESSC is a part.

The in-situ data which we handle includes the ENACT/ENSEMBLES EN3 dataset from the Met Office Hadley Centre (ref. 2). This comprises around 10 million temperature and salinity profiles throughout the global ocean from 1950 onwards. Other data include that from the RAPID array at 26°N, Argo float data, and data from the SEPRISE project (from moorings around Europe).

3. Problems in managing these datasets

There are various problems inherent in handling these datasets. Firstly the large volumes of data. This is particularly acute for gridded model data, but is increasingly becoming an issue with other datasets. There is also a problem with data being dispersed around various institutes - we handle data from other centres via OPeNDAP for example. One of the most challenging problems we face is one of heterogeneity in the data products. This is particularly problematic in observed data where many formats are in use.

This heterogeneity inhibits interoperability between centres and projects. Each data provider may choose to serve data using different protocols, and formatted in different ways. Downstream users such as ReSC wishing to integrate different data into one portal are then faced with writing custom code for each different dataset. This in turn reduced reusability for future projects.

4. Tools and technologies to solve these problems

There are a number of useful tools and technologies which can ease some of the difficulties mentioned above. These include Open Geospatial Consortium (OGC) standards, Climate Science Modelling Language (CSML), OPeNDAP, THREDDS and relational databases such as PostgreSQL/PostGIS.

4.1 OGC standards

The OGC is a body in charge of standards for geospatial and location based services. Recent ReSC projects have found their WMS and WFS standards particularly helpful in handling diverse datasets. The WMS (Web Map Service) standard specifies how to serve geo-referenced images of data over the web. This is ideal for gridded model output. The WFS (Web Feature Service) standard specifies how to serve geo-referenced points, lines and polygons over the web. This is ideal for in-situ observations such as from moorings and Argo floats. In each case the important point is that the standards allow users to know precisely how to request data, and what format they are going to get back in return. ReSC has developed a very successful WMS for NetCDF data know as ncWMS (ref. 3). This connects to a web interface client known as Godiva2 also developed at ReSC, and which uses OpenLayers to display the data.

4.2 *THREDDS Data Server (TDS)*

The TDS enables data providers to serve NetCDF and similar data easily online via the OPeNDAP protocol. Importantly, subsetting of the data is possible in the request, enabling users to avoid having to download a huge file in order to read in perhaps only a few data points of interest. There is a new version of THREDDS released which contains a copy of ncWMS bundled into it. This then enables data to be served either via OPeNDAP (perhaps for observations) or via WMS (perhaps for gridded data).

4.3 *Climate Science Modelling Language (CSML)*

CSML is a standards-based way of representing features of interest to the climate scientist. There are 13 main features encoded in the CSML data model including profiles, trajectories, timeseries and swaths. CSML provides a common view onto datasets regardless of their storage format or physical location, and is ideal for integrating diverse data products. The ReSC and collaborators are currently developing a set of reusable Java libraries for CSML which will further enhance its usefulness.

4.4 *PostgreSQL/PostGIS*

PostgreSQL is the most advanced open source database, and PostGIS is an open source extension to PostgreSQL giving it geospatial capabilities, and following the OGC 'Simple Features for SQL' specification. It should be noted that at ReSC we require the ability to be able to search for data within an arbitrary polygon on the Earth's surface, and this functionality can be added to PostgreSQL/PostGIS by installing the GEOS library. PostgreSQL has good support for Java in the form of the JDBC driver, which is important in enabling integration with a suite of Java web applications developed at ReSC.

5. Some examples of ReSC products using these tools and technologies

5.1 *OceanDIVA: Ocean Data Intercomparison and Visualization Application*

OceanDIVA (ref. 4) is a Java web application which allows the comparison of model and observed ocean temperature and salinity data. It makes use of NetCDF, OPeNDAP, KML and Google Earth. There are two possible types of output — KML for the viewing of geospatial data in Google Earth or similar, and probability distribution functions of the model/observation differences over various regions. Observations are from the EN3 dataset held locally in NetCDF files at ESSC, and models include locally run NEMO code, as well as various other models run elsewhere and accessed via OPeNDAP.

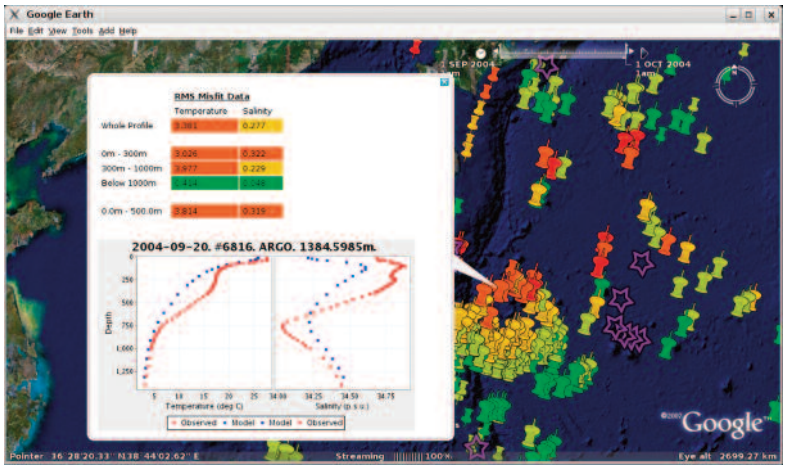

Fig. 1 KML output from OceanDIVA shown in Google Earth

Figure 1 shows an example of KML output in Google Earth. The coloured pins represent the location of all the available observations for the requested time period. They are coloured according to the RMS misfit between the temperature profile from the instrument and the nearest model data. Observations with a close match to the model towards the green end of the scale, and those with a poor match are towards the red end of the scale. If the user clicks a profile pin, the application queries the particular data in question and generates a profile plot on the fly. This can show the user at what depths the model and observations are in agreement or disagreement.

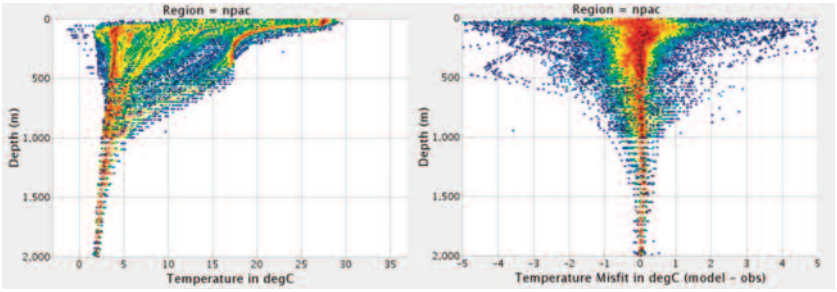

Fig. 2 Example PDF output from OceanDIVA for the North Pacific

Figure 2 shows an example of probability density function output from OceanDIVA. The panel on the left shows all the observed temperature profiles in the North Pacific for a given month. The more red areas of the plot indicate greater data density. The right panel shows the range of misfits between these observations and the model. In this case data on the right of the centre show that the model is too warm, and those on the left show the model is too cold. If the model and observations were in 100% perfect agreement the right hand plot would be a straight vertical line down the middle of the panel.

5.2 The ECOOP Ecosystem Portal

The ECOOP project is a large European (FP6) Coastal Operational Oceanography Project. As part of ECOOP ReSC and collaborators at the Met Office were tasked with developing a web portal which integrates ecosystem-relevant model and observed data from the North Sea and allows users to compare between model and observations. This was intended as a technology demonstration for a decision support system for predicting harmful algal blooms (HABs). The idea being that if one can show agreement between model and observations for the past few days, one is more likely to believe the model forecast of a HAB.

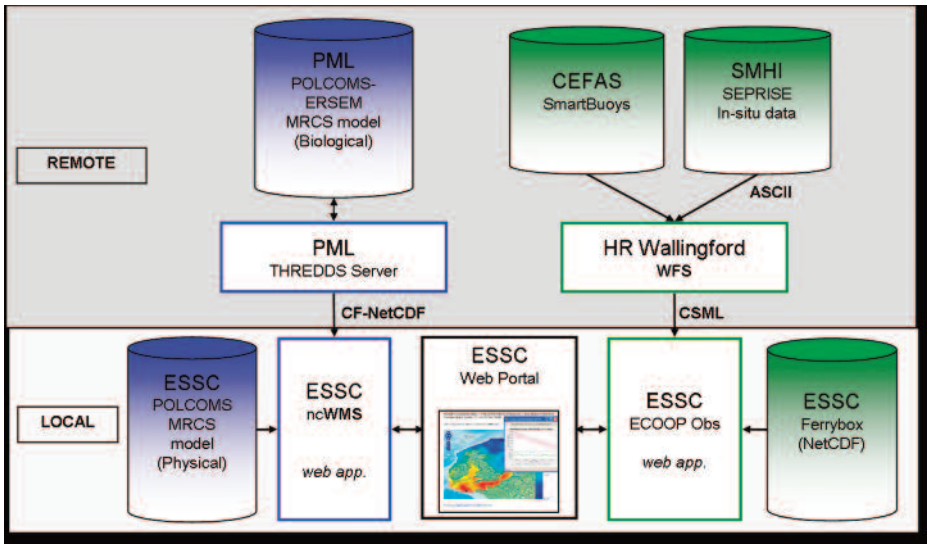

Fig. 3 The data flows involved in the ECOOP ecosystem project

Figure 3 shows how the data arrives at the web portal. Model data is in blue, and observed data is in green. The model data is all formatted as NetCDF, and is ingested into the portal via THREDDS and ncWMS. The observed data is formatted in a variety of ways, but is represented as CSML in the HR Wallingford WFS (with the exception of the NetCDF Ferrybox data), and this aids the feeding of the data into the portal via the ECOOP Obs java web application.

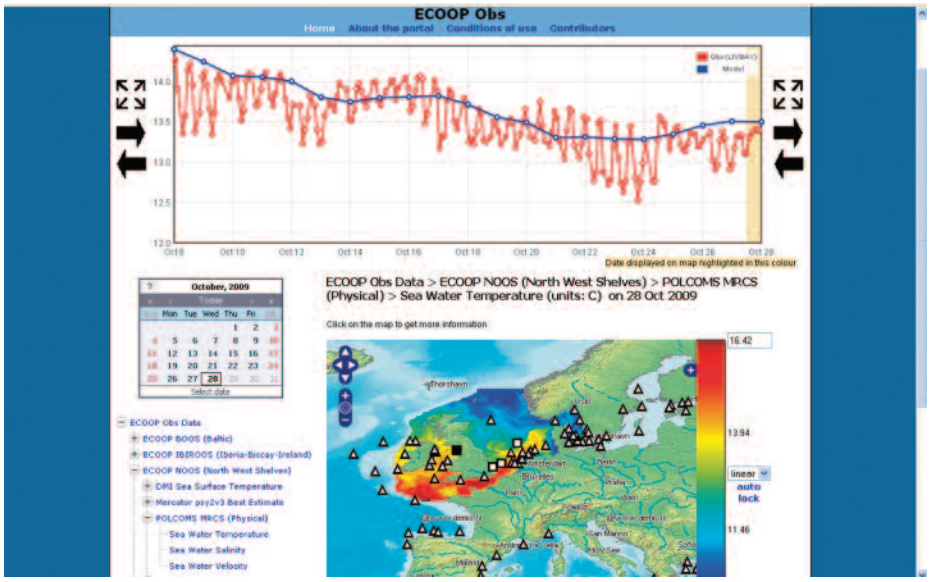

Fig. 4 Screenshot from the ECOOP Ecosystem portal

Figure 4 shows a screenshot of the portal. The users has clicked on a menu on the left hand side to request ocean temperature from the POLCOMS MRCS model. The portal (via the web application) has then requested the WFS to return the locations of all observations containing temperature data within the previous 10 days. The user has then clicked on an observation in Liverpool Bay. This triggers the portal (via the web application) to request two data streams. One is the previous 10 days of temperature data from the Liverpool Bay buoy. The other is the previous 10 days of temperature data from the model output at this point. The latter is returned via the OGC WMS getFeatureInfo specification. The portal then uses a Javascript plotting library called Flot (ref. 5) to create the timeseries plot at the top of the map. This can be extended or contracted in either direction, and the user can zoom in, and query any point for its value.

6. 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. One can either store metadata alone, or metadata plus data.

Data format is also an important consideration. There are a number of projects looking to standardise this in marine community - 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, and OPeNDAP can enable dynamic access to data from a range of holdings over the web. We have paired this with THREDDS and ncWMS in the work presented here.

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. However, there is scope for harnessing the power of tools such as Google Earth. This gives a powerful visualization platform and standard KML at cost of being closed source.

References

1. http://www.nemo-ocean.eu/
2. **Ingleby, B.**, and **M. Huddleston**, 2007: Quality control of ocean temperature and salinity profiles – historical and real-time data. *Journal of Marine Systems*, **65**, 158-175 10.1016/j.jmarsys.2005.11.019
3. **Jon Blower, Keith Haines, Adit Santokhee, Chunlei Liu**, Godiva2: Interactive visualization of environmental data on the web, *Phil. Trans. Roy. Soc. A*, **367**, 1035-9, 2009
4. **A.L. Gemmell, G.C. Smith, K. Haines, J.D. Blower**, Validation of ocean model syntheses against hydrography using a new web application, *Journal of Operational Oceanography* **2(2)** August 2009, pp. 29-41
5. http://code.google.com/p/flot/