Assimilating sparse historical data using large scale EOF error covariances

Daniel Lea and Matt Martin

GOV Meeting, Santa Cruz July 2016

Outline

- Motivation
- New NEMOVAR
- Diffusion modelled covariance
- Ensemble covariance
- EOF covariance

Motivation

Met Office

ERA Clim 2 coupled reanalysis planned from 1900 to approx present day

Problem the ocean observation network has changed markedly in the last 100 years or so.

Observations were sparse and sampling was inhomogeneous. Observations now much less sparse and more globally homogeneous, but we still have sparse sampling at depth

How can assimilation make best use of the sparse historical data while still doing a good job with today's data?

The key thing which gives data assimilation its power is the background error covariance which allows us to spread information from the observation locations.

Can we improve the error covariance structures to allow us to correctly spread sparse observation information over greater distances in order to fill in the gaps?

NEMOVAR what is it

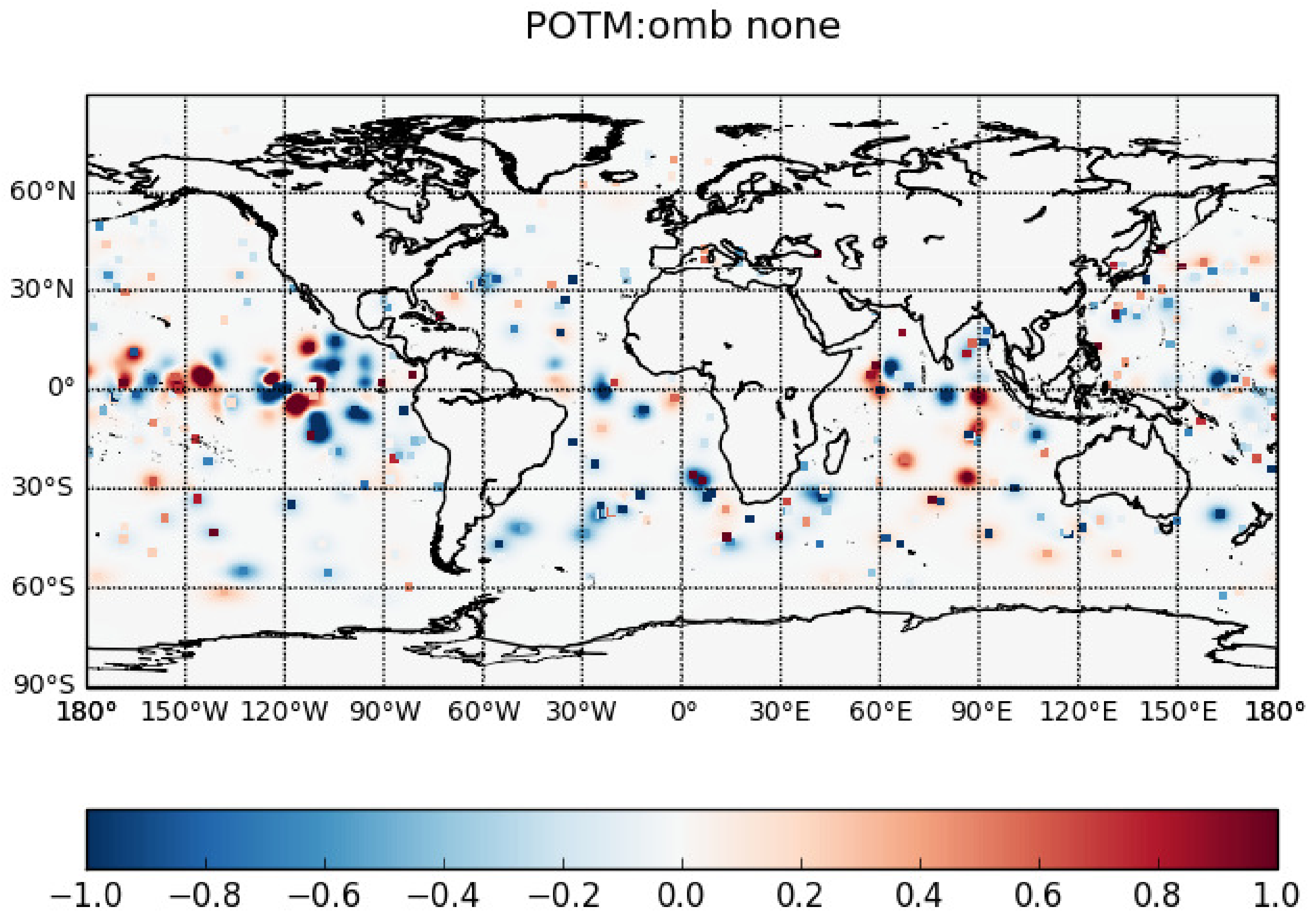
- Ocean data assimilation system for NEMO (3D/4D variational)
- It's the result of a collaboration between CERFACS, ECMWF, INRIA, Met Office ...
- At the Met Office this is used in FOAM (deep and shelf ocean) and OSTIA
- Main benefits: Compatible with the NEMO. It works with the ORCA grids natively. Efficient. There's an effective balance operator.

New hybrid B NEMOVAR – what’s new?

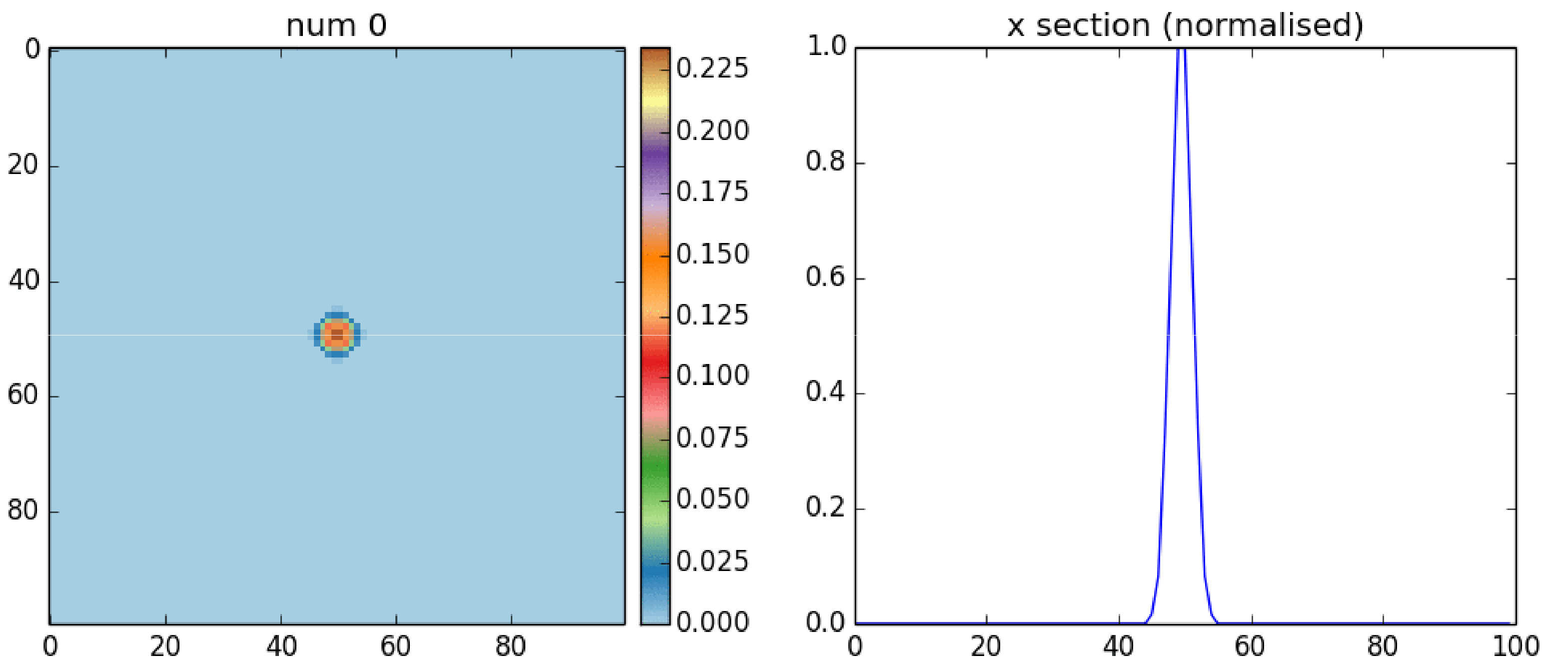
- Recently there has been a major improvement to NEMOVAR (mainly developed at CERFACS thanks to Anthony Weaver) (v4)
- We are still using an earlier version (v3) operationally
- The major changes from v3 to v4 are to improve the spatial spreading of observation information:
    - 2D implicit solver for correlation modelled by diffusion
    - Allows the use of ensemble information
    - Also the use of EOFs (by me!)

Some example Increments

using just the diffusion modelled covariances



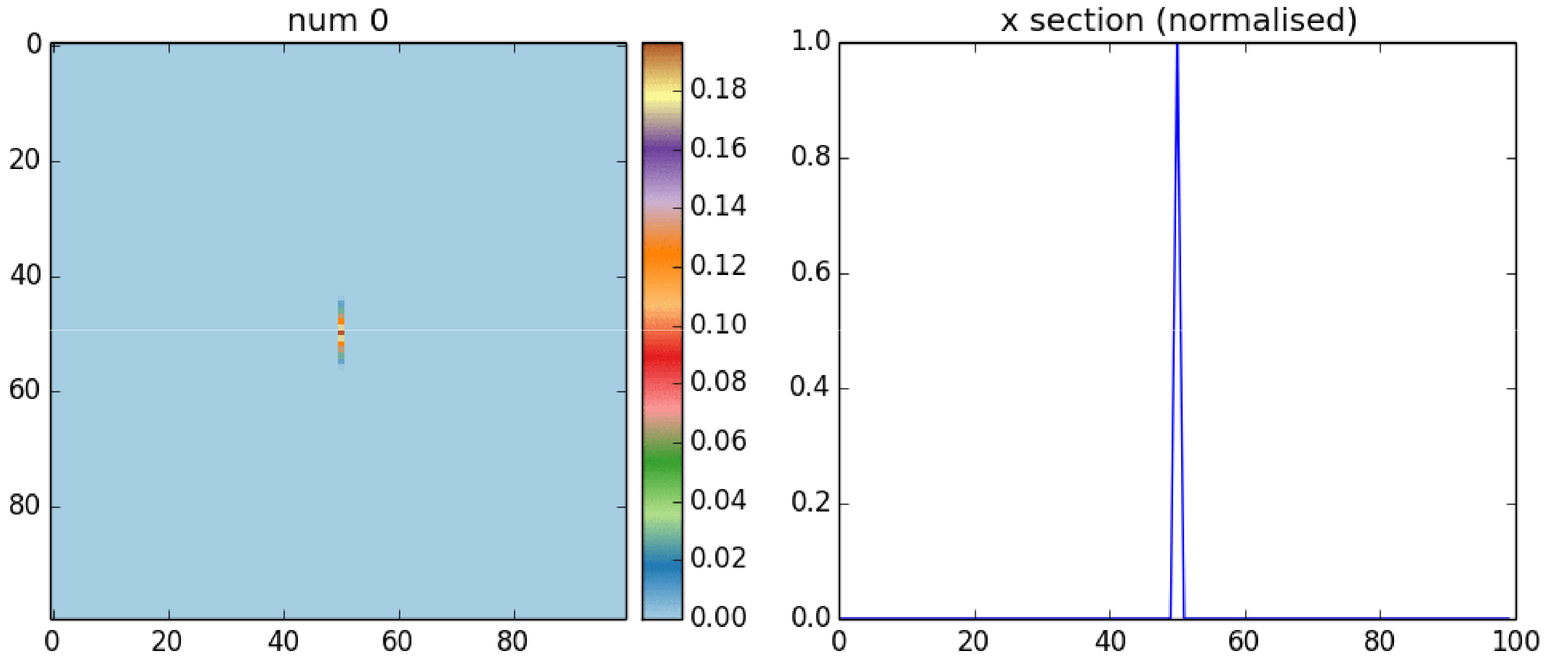
2D diffusion (explicit – Dirichlet BCs)



Toy example using python

Implicit solver for 2D diffusion used in New (v4) NEMOVAR

2x1D diffusion model (explicit)



Implicit version of this used in v3 (old) NEMOVAR

Ensemble covariance

The diffusion model gives a rather limited representation of the real covariance. Another idea is to use information from an ensemble of model runs to spread observation information in a flow dependent way.

This is akin to working in ensemble space rather than real space

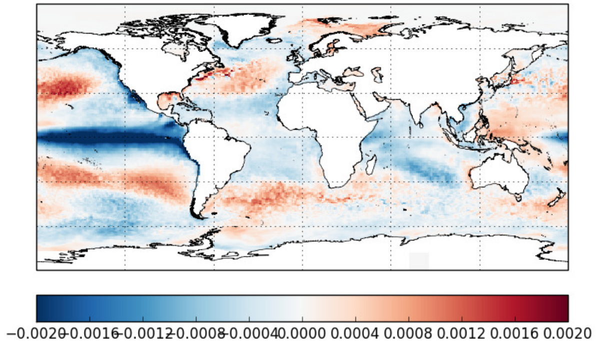
Met Office

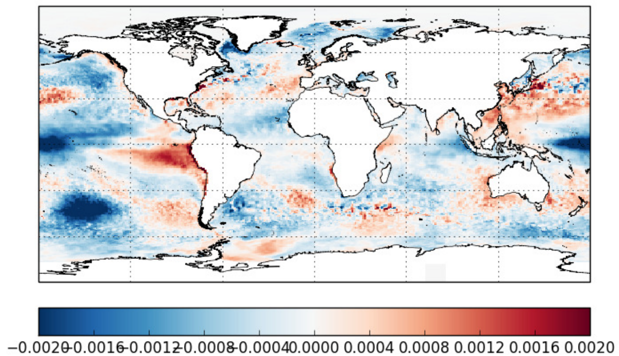
EOF DA

• I've adapted the ensemble covariance code for EOF DA. EOFs are just another "space" to spread the observation information in.

• This is work is motivated by the sparseness of historical data. This is for ERA-CLIM2 – ECMWF coupled reanalysis R&D - but may also be applicable to the decadal prediction system.

•Top two SST EOFs from Glosea 5 reanalysis >>>





EOF DA

Working in model space

$$J(\delta\mathbf{x}) = \tfrac{1}{2}\,\delta\mathbf{x}^{\mathrm{T}}\,\mathbf{B}^{-1}\,\delta\mathbf{x} + \tfrac{1}{2}\,(\mathbf{y}-\mathbf{H}(\mathbf{x}_b+\delta\mathbf{x}))^{\mathrm{T}}\,\mathbf{R}^{-1}\,(\mathbf{y}-\mathbf{H}(\mathbf{x}_b+\delta\mathbf{x}))$$

$$\delta\mathbf{x} = \mathbf{E}\,\mathbf{a}$$

Working in EOF space

$$J(\mathbf{a}) = \tfrac{1}{2}\,\mathbf{a}^{\mathrm{T}}\,\Lambda^{-1}\,\mathbf{a} + \tfrac{1}{2}\,(\mathbf{y}-\mathbf{H}(\mathbf{x}_b+\mathbf{Ea}))^{\mathrm{T}}\,\mathbf{R}^{-1}\,(\mathbf{y}-\mathbf{H}(\mathbf{x}_b+\mathbf{Ea}))$$

Hybrid

$$\delta\,\mathbf{x} = w_1\,\mathbf{E}\,\mathbf{a} + w_2\,\delta\,\mathbf{x}_{\mathrm{residual}}$$

$$J = w_1\,\tfrac{1}{2}\,\mathbf{a}^{\mathrm{T}}\,\Lambda^{-1}\,\mathbf{a} + w_2\,\tfrac{1}{2}\,\delta\,\mathbf{x}_{\mathrm{res}}{}^{\mathrm{T}}\,\mathbf{B}^{-1}\,\delta\,\mathbf{x}_{\mathrm{res}} + \text{[obs cost]}$$

a= vector of coefficients/weights for each EOF (Temperature anomaly = **Ea**)

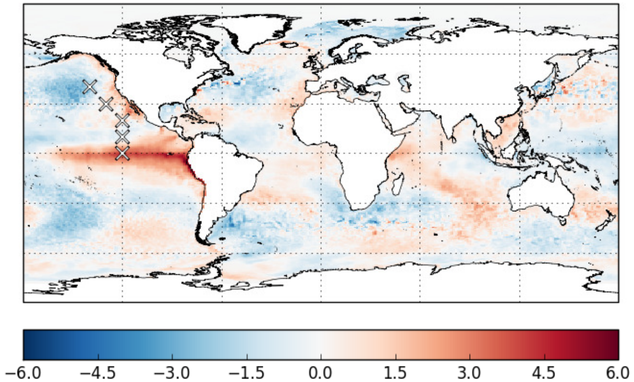
E= EOFs

Λ= diagonal matrix of eigenvalues (calculated with the EOFs) (squared)

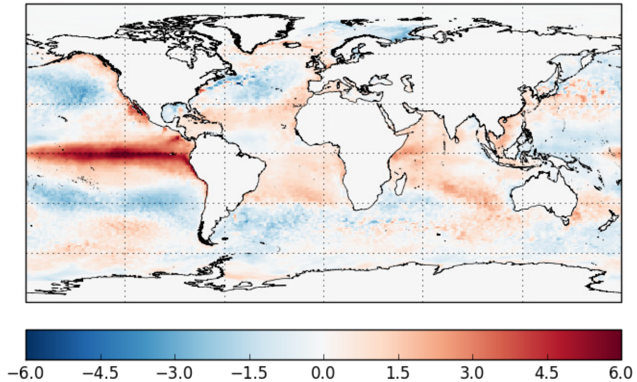
Increments with different hybrid EOF and diffusion modelled covariance weights

Met Office

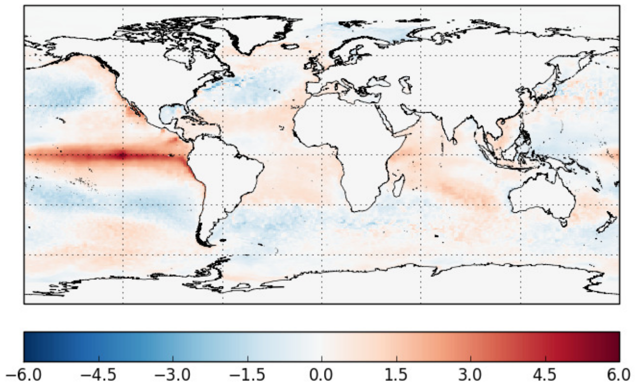
Truth and sim. obs locations



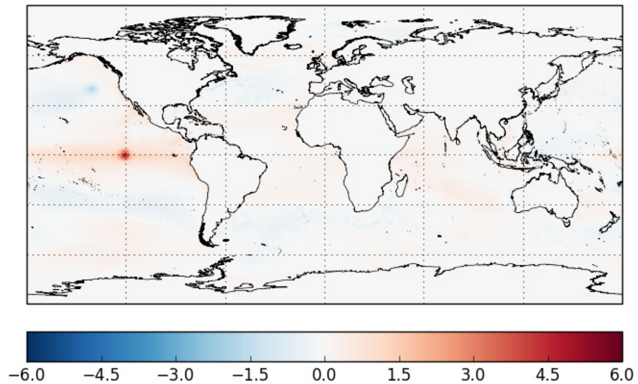
W_1=1 (W_1+W_2=1)



W_1=0.01



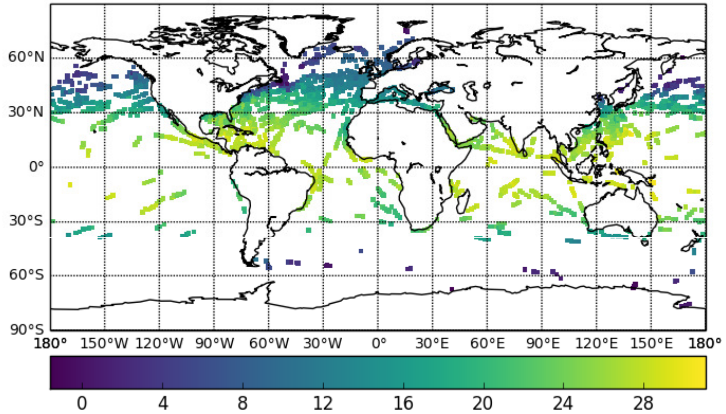
W_1=0.001



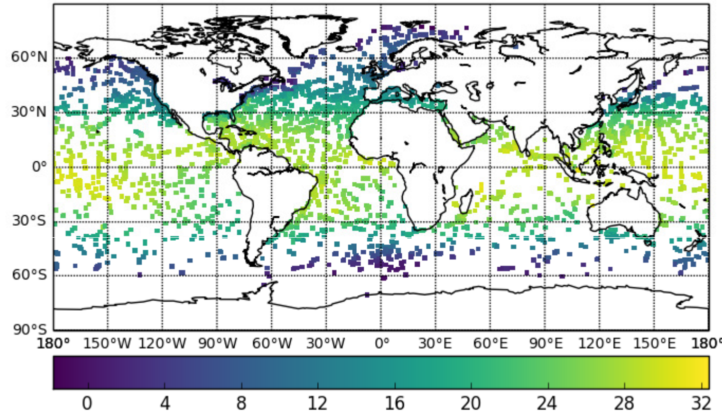
Observing system experiment

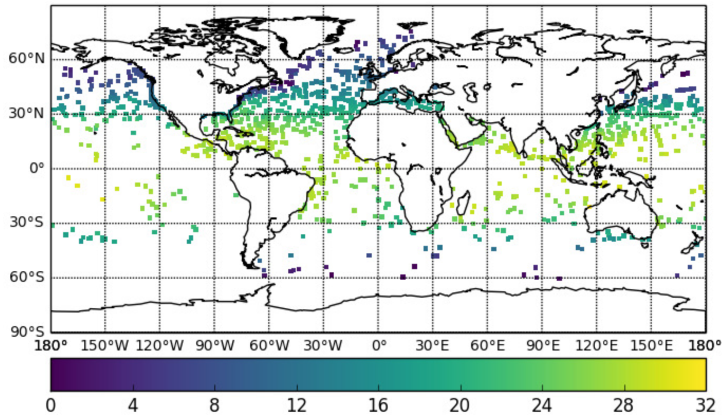
Subsample modern day observations to look like historical data (example of in-situ SST observations in ºC)

1 Jan 1960



1 Jan 2010





2010 data subsampled to1960 distribution

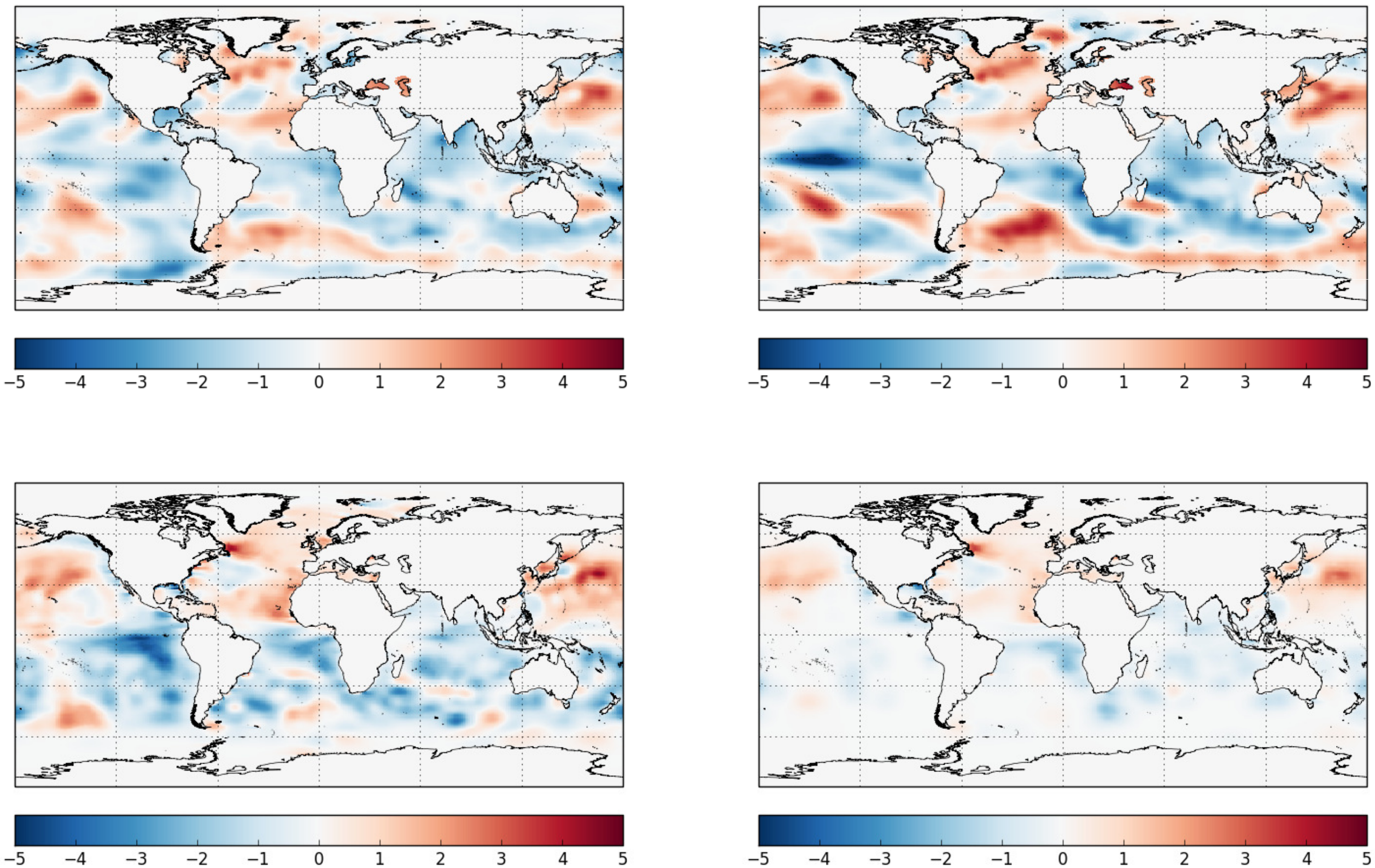
Met Office

Test results (very preliminary)
ocean surface temperature increments /ºC

Full 2010 data | Subsampled 2010 data

EOF DA

Stand
ard
DA



Met Office

Plans

• Run many observing system experiments to assess the robustness of the EOF estimates by comparing to the unassimilated data (and comparing the analysis fields)

• Compare to the standard FOAM type data assimilation (with shorter lengthscale guassian background errors)

• Look at the impact of the source of the EOFs (model/observation based...)

• Top EOFs will likely contain useful and robust correlation structures. The lower EOFs will be less useful. Will assess how many EOFs to use to minimise the spurious increments where there are few observations (and maximise the useful increments in those locations)

• Assess multivariate & 3D aspects

•Test in a reanalysis system

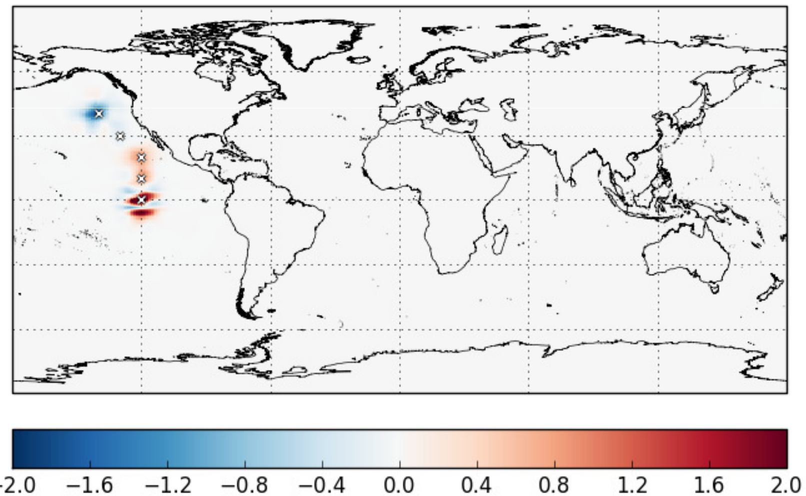
Met Office

End

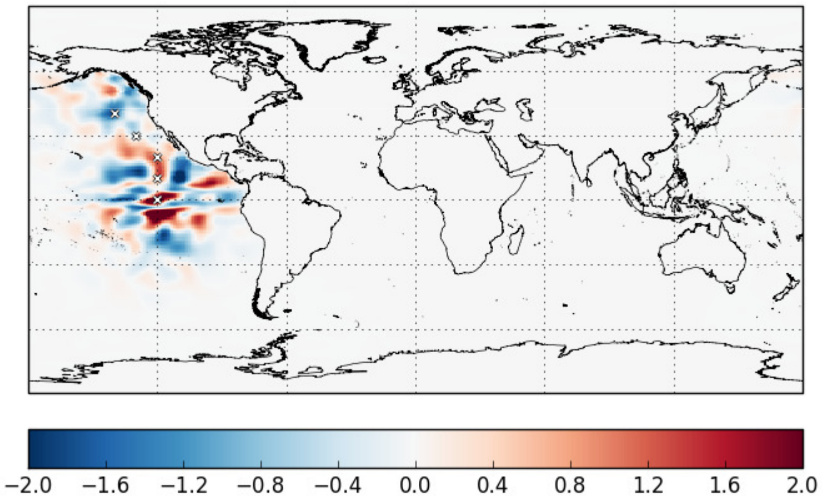
www.metoffice.gov.uk

Ensemble data assimilation test

Test: (random) ensemble (temperature) increments assimilating 5 simulated observations
Localisation uses diffusion equation (as described earlier)



8 degrees localisation lengthscale



60 degrees localisation lengthscale

Diffusion model

To spread the observation information we need to know about the covariance... – this is how much one point varies with respect to another (usually nearby point) (with a strong covariance if you make a change at one point you need to make the same change at another point).

Usually we expect close points co-vary more than distant points... (could do more complex things ... But wait until later!)

To implement this we use the diffusion equation - good for the ocean problem as it can deal nicely with coastlines:

$$du/dt = \nabla . (K \nabla u)$$

For uniform K(x,y,z)=k for a 3D field u=u(x,y,z):

$$du/dt = k \{ d^2u/dx^2 + d^2u/dy^2 + d^2u/dz^2 \}$$

Explicit discretisation/timestepping (on a grid):

$$u(x,t+1)-u(x,t) = k \{u(x+1,t)-2u(x,t)+u(x-1,t) \} + \dots$$

1D implicit diffusion

Explicit timestepping (forwards Euler)

$u(x,t+1)-u(x,t) = k\,\{u(x+1,t)-2u(x,t)+u(x-1,t)\}$

Implicit timestepping (backwards Euler)

$u(x,t+1)-u(x,t) = k\,\{u(x+1,t+1)-2u(x,t+1)+u(x-1,t+1)\}$

Rearrange and make u a vector (in space)

$\mathbf{u}(t+1) = (\mathbf{I} - \mathbf{M})^{-1}\,\mathbf{u}(t)$

This makes it stable even with long timesteps. But this is harder to solve this matrix problem (NEMOVAR v3 uses Cholesky decomposition) but it is still much cheaper than the explicit version for longer length scales relative to the grid spacing (like in orca ¼ degree)

But coastal artifacts if not careful