LET’S TALK ABOUT DATA ASSIMILATION

ARYA PAUL
INCOIS

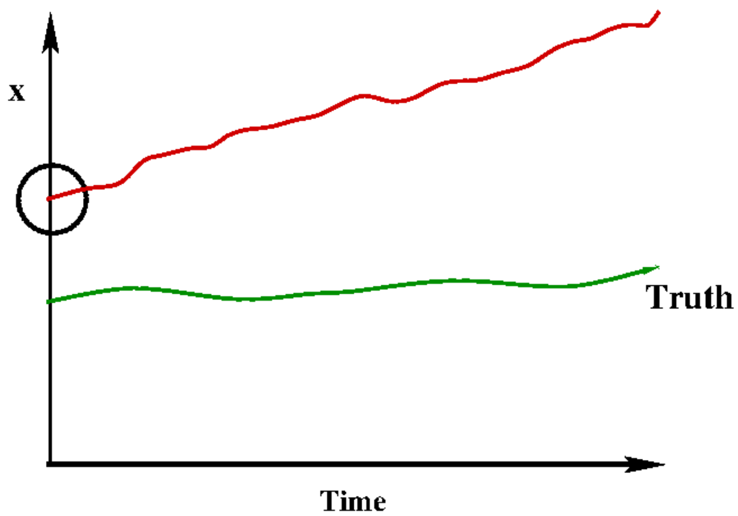
Gauss, 1809: Theoria Motus Corporum Coelestium
-1823: Theoria combinationis Observationum erroribus minimis obnoxiae

If the astronomical observations and other quantities on which the computations of orbits is based, were absolutely correct, the elements also, whether deduced from three or four observations, would be strictly accurate (so far indeed as the motion is supposed to actually take place exactly according to the laws of Kepler), and therefore, if other observations were used, they might be confirmed, but not corrected. But since our measurements and observations are nothing more than approximations to truth, the same must be true of all calculations resting upon them, and the highest aim of all computation made concerning concrete phenomena must be to approximate, as nearly as practicable to the truth. But this can be accomplished in no other way than by a suitable combination of more observations than the number absolutely requisite for the determination of the unknown quantities. This problem can only be properly be undertaken when an approximate knowledge of the orbit has been already taken into account, which is afterwards to be corrected, so as to satisfy all the observations in the most accurate manner possible.

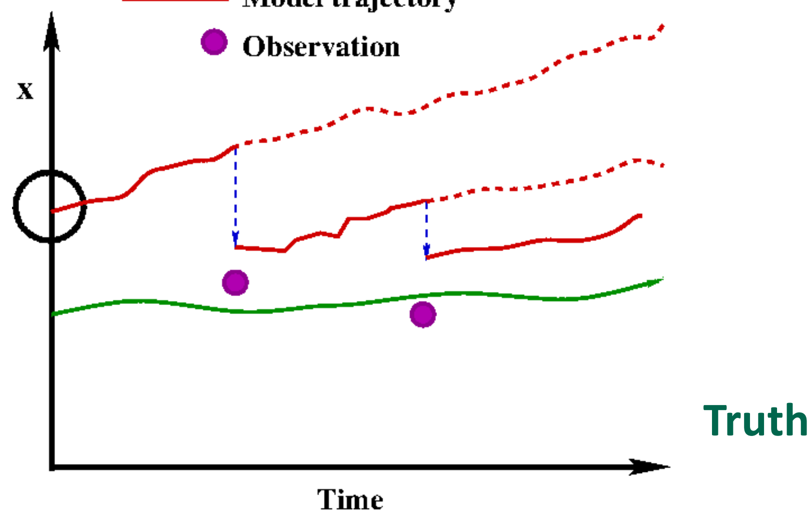
Summary of Gauss’s idea

- Both model and observations are approximate.
- Truth is not known.
- The resulting analysis will also be approximate.
- It’s better to have enough observations to over-determine the problem.
- The model is used to provide a preliminary estimate.

Flowchart of Data Assimilation



NO CORRECTION



CORRECTION

DATA ASSIMILATION

Finding maximum likelihood (using Bayes' Theorem)

Minimize the cost function (Least square approach)

WHAT IS BAYES' THEOREM ?

Bayes' Theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

P(A|B) = Probability of finding A given B
P(B|A) = Probability of finding B given A
P(A) = Probability of A with no knowledge of B
P(B) = Probability of B with no knowledge of A.

Did you ever bet on horses ?

Total Number of Races	12
Fleetfloot winning	7
Bolt winning	5

Probability of Bolt winning = 5/12 = 41.7%

Probability of Fleetfoot winning = 7/12 = 58.3%

Now let's add a new factor into the calculation. It turns out that on 3 of Bolt's previous 5 wins, it had rained heavily before the race. However, it had rained only once on any of the days that he lost. It appears, therefore, that Bolt is a horse who likes 'soft going', as the bookies say. On the day of the race in question, it is raining.

Given this new information (raining), what is the probability of Bolt winning ?

Ref : http://www.kevinboone.net/bayes.html

	It's raining	Not raining
Bolt winning	3	2
Bolt losing	1	6

What we need to know is the probability of Bolt winning, ***given that it is raining*** **?**

Like any other probability, we calculate it by dividing the number of times something happened, by the number of times if could have happened.
We know that Bolt won on 3 occasions on which it rained, and there were 4 rainy days in total.
So Bolt's probability of winning, given that it is now raining, is 3 / 4, or 0.75, or 75%.

The probability shifts from 41.7% to 75%.

This is important information if you plan to bet — **if it is raining you should back Bolt; if it is not, you should back Fleetfoot**.

Revisiting Bayes' Theorem

p(A|B) = p(B|A) p(A) / p(B)

P(A|B) = Probability of finding A given B
P(B|A) = Probability of finding B given A
P(A) = Probability of A with no knowledge of B
P(B) = Probability of B with no knowledge of A.

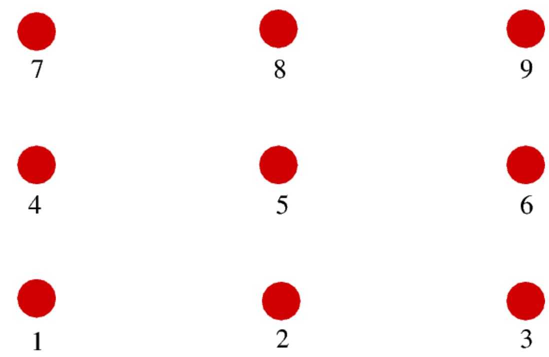
P(A|B) = Probability of Bolt winning when it rains
P(B|A) = Probability of raining when Bolt wins = 3/5
P(A) = Probability of Bolt winning = 5/12
P(B) = Probability of raining = 4/12

$$p(A|B) = \left(\frac{3}{5} \times \frac{5}{12}\right) \div \frac{4}{12} = \frac{3}{4}$$

BASICS

$$x_{k+1}^{b} = Mx_{k}^{b} + \varepsilon^{b}$$

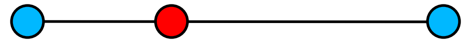
$$y_{k+1} = Hx_{k+1}^{b} + \varepsilon^{o}$$



What is x ?

What is y ?

What is H ?



What are the error characteristics ?

- Unbiased model and observation error i.e., $\langle \varepsilon^{b} \rangle = \langle \varepsilon^{o} \rangle = 0$
- Model and observation error are uncorrelated i.e., $\langle \varepsilon^{b} \varepsilon^{o^{T}} \rangle = 0$
- Non-trivial error covariances i.e., $\langle \varepsilon^{b} \varepsilon^{b^{T}} \rangle = B, \quad \langle \varepsilon^{o} \varepsilon^{o^{T}} \rangle = R$

$$p(x \mid y) \propto p(y \mid x)p(x)$$

Given observations, what is the best estimate of the state x.

$$p(x) \propto \exp\left(-\frac{1}{2}(x-x^b)B^{-1}(x-x^b)^T\right)$$ ⟸ B is Gaussian

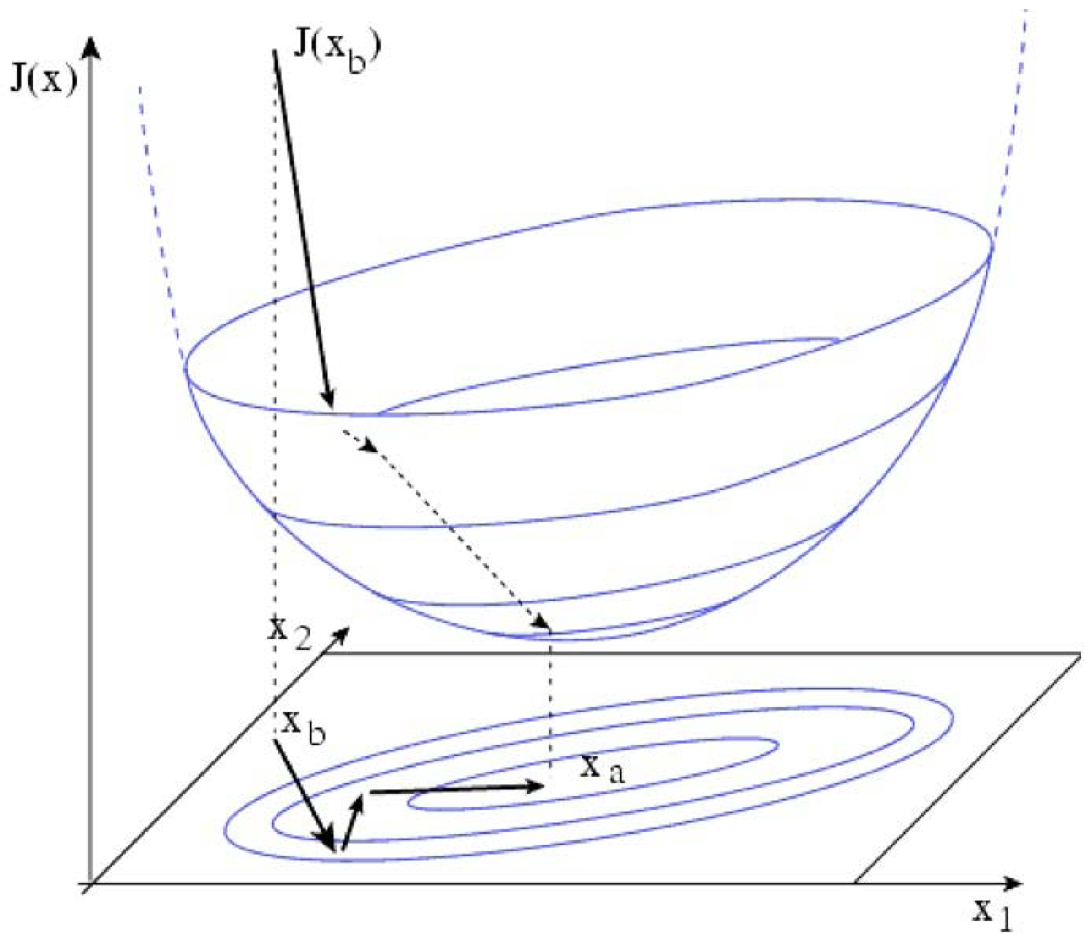
$$p(y \mid x) \propto \exp\left(-\frac{1}{2}(Hx-y)R^{-1}(Hx-y)^T\right)$$ ⟸ R is Gaussian

$$p(x \mid y) \propto \exp\left(-\frac{1}{2}J(x)\right) \quad \text{where}$$

$$J(x) = (x-x^b)B^{-1}(x-x^b)^T + (Hx-y)R^{-1}(Hx-y)^T$$

Maximizing $p(x \mid y)$ is same as **Minimizing** $J(x)$

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

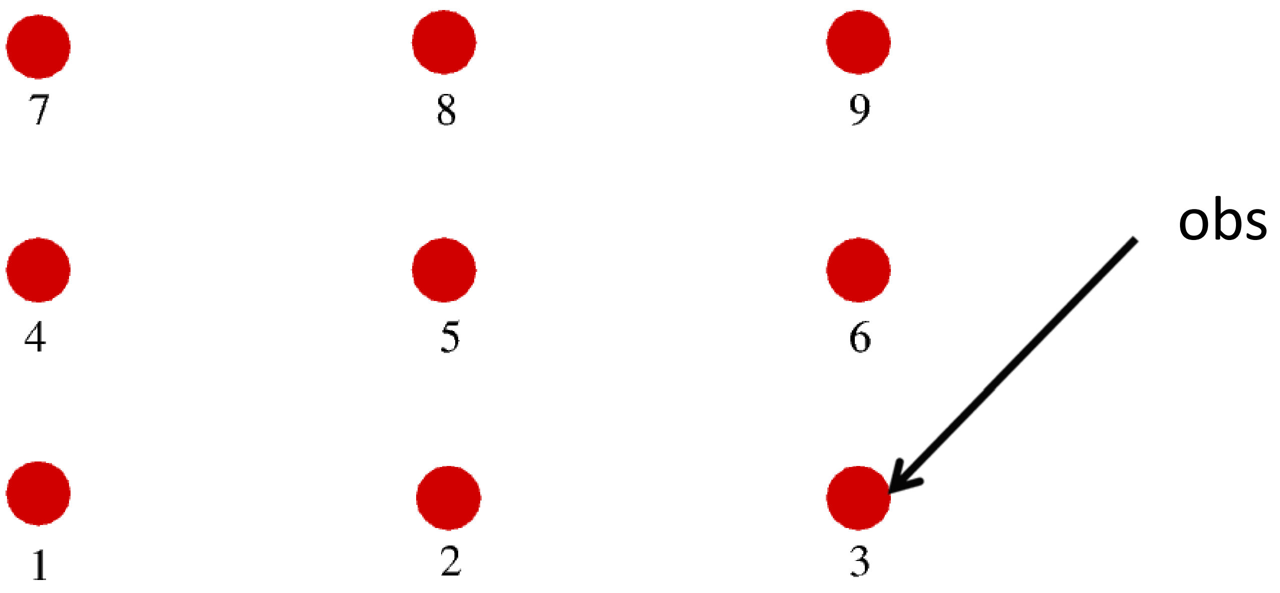


The cost function is parabolic and the minimization is done using steepest descent.

POPULAR DATA ASSIMILATION METHODS

- KALMAN FILTER -- B evolves according to model dynamics.
- 3D VAR – B is stationary.
- 4D VAR – B evolves within the time window of cost function minimization.
- ENSEMBLE BASED KALMAN FILTER -- B is estimated from the ensembles.

What is the role of B ??



Under what condition will the information at grid location 3 propagate to other grid points ?

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

$$x^b = \begin{bmatrix} x_1^b \\ x_2^b \\ . \\ . \\ x_9^b \end{bmatrix}, \quad y = y_0, \quad H = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$R = \sigma_0^2 \qquad B = \begin{bmatrix} B_{11} & B_{12} & . & . & . & B_{19} \\ B_{21} & B_{22} & . & . & . & B_{29} \\ . & . & B_{33} & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ B_{31} & B_{32} & . & . & . & B_{39} \end{bmatrix}$$

$$y - Hx^b = y_0 - \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^b \\ x_2^b \\ . \\ . \\ x_9^b \end{bmatrix} = y_0 - x_3^b$$

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

$$HB = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & . & . & . & B_{19} \\ B_{21} & B_{22} & . & . & . & B_{29} \\ . & . & B_{33} & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ B_{31} & B_{32} & . & . & . & B_{39} \end{bmatrix} = \begin{bmatrix} B_{31} & B_{32} & . & . & B_{39} \end{bmatrix}$$

$$HBH^T = \begin{bmatrix} B_{31} & B_{32} & . & . & B_{39} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ . \\ 0 \end{bmatrix} = B_{33}$$

$$\left(HBH^T + R\right)^{-1} = \frac{1}{B_{33} + \sigma_0^2} \qquad \Rightarrow \left(HBH^T + R\right)^{-1}\left(y - Hx^b\right) = \frac{y_0 - x_3^b}{B_{33} + \sigma_0^2}$$

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

$$\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right) = \frac{y_0 - x_3^b}{B_{33} + \sigma_0^2}$$

$$BH^T = \begin{bmatrix} B_{11} & B_{12} & . & . & . & B_{19} \\ B_{21} & B_{22} & . & . & . & B_{29} \\ . & . & B_{33} & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ B_{31} & B_{32} & . & . & . & B_{39} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ . \\ 0 \end{bmatrix} = \begin{bmatrix} B_{13} \\ . \\ . \\ . \\ B_{93} \end{bmatrix}$$

$$BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right) = \begin{bmatrix} B_{13} \\ . \\ . \\ . \\ B_{93} \end{bmatrix} \frac{\left(y_0 - x_3^b\right)}{\left(B_{33} + \sigma_0^2\right)}$$

$$
\begin{bmatrix} x_1^a \\ x_2^a \\ . \\ . \\ x_9^a \end{bmatrix} = \begin{bmatrix} x_1^b \\ x_2^b \\ . \\ . \\ x_9^b \end{bmatrix} + \begin{bmatrix} B_{13} \\ B_{23} \\ . \\ . \\ B_{93} \end{bmatrix} \frac{\left(y_0 - x_3^b\right)}{\left(B_{33} + \sigma_0^2\right)}
$$

$$
x_k^a = x_k^b + \left(\frac{y_0 - x_3^b}{B_{33} + \sigma_0^2} \right) B_{k3}
$$

B propagates information from one site to another !!!

What is the relative significance of B & R ?

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

Let's estimate the temperature of Hyderabad.

Given

$$x^b = 31.0, \sigma_b^2 = 2$$
$$y_0 = 30.0, \sigma_0^2 = 1$$

In this case, $H = 1, R = \sigma_o^2, B = \sigma_b^2$

$$x^a = x^b + \sigma_b^2\left(\sigma_b^2 + \sigma_o^2\right)^{-1}\left(y_0 - x^b\right)$$
$$\Rightarrow x^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

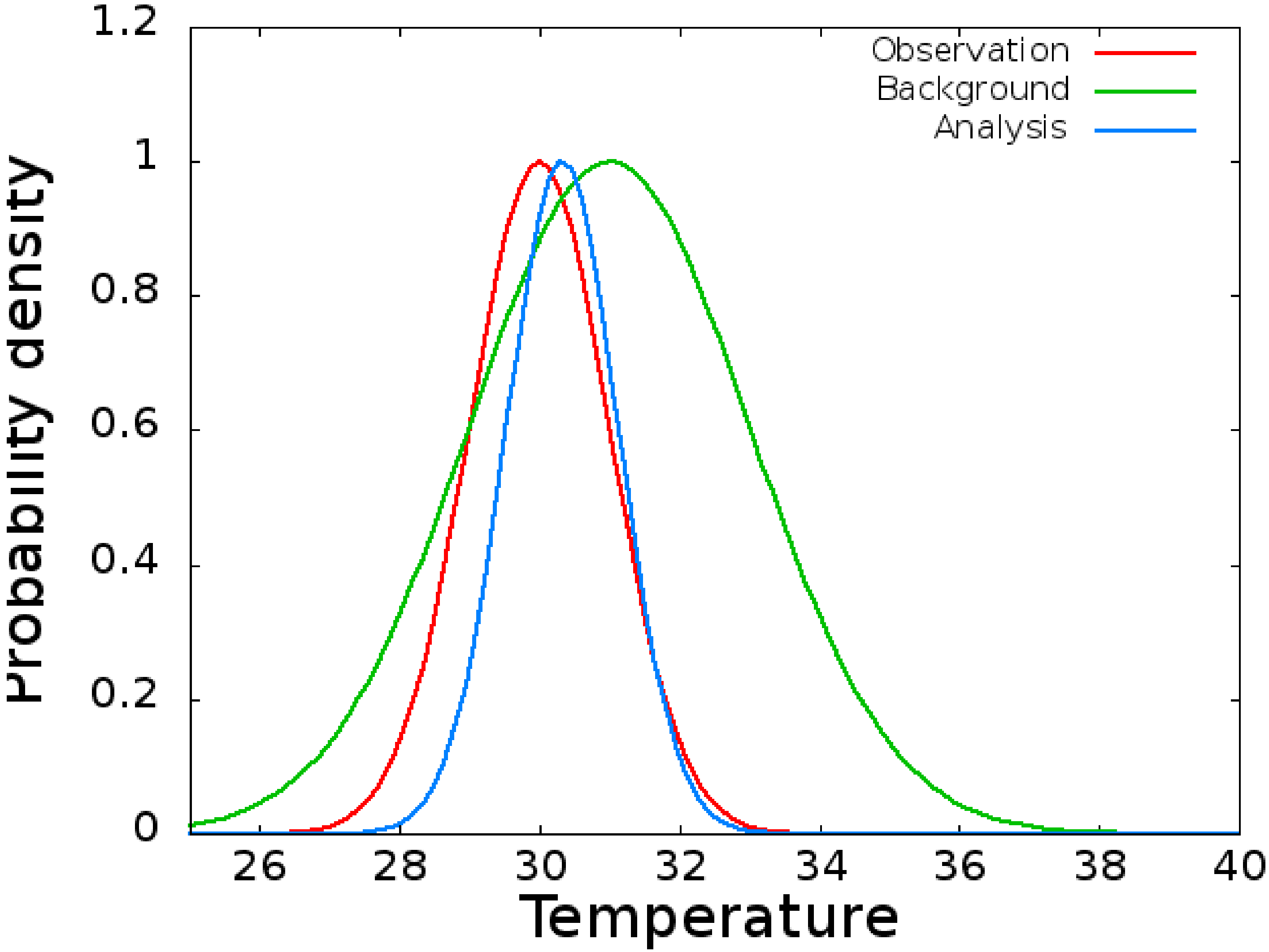
$$x^a = 30.33, \sigma_a^2 = 0.8$$

If $\sigma_b >> \sigma_0$

$x^a \approx y_0$

If $\sigma_0 >> \sigma_b$

$x^a \approx x_b$



$$x^b = 31.0, \sigma_b^2 = 2$$

$$y_0 = 30.0, \sigma_0^2 = 1$$

$$x^a = 30.33, \sigma_a^2 = 0.8$$

Some more exercises

$$x^a = x^b + BH^T\left(HBH^T + R\right)^{-1}\left(y - Hx^b\right)$$

Suppose we observe a point in between two grid points.

$$Hx^b = \alpha x_1^b + (1-\alpha)x_2^b; \quad 0 \le \alpha \le 1$$

Assume

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 & \mu\sigma_b^2 \\ \mu\sigma_b^2 & \sigma_b^2 \end{bmatrix}; \qquad R = \sigma_0^2$$

$$\begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} \alpha + \mu(1-\alpha) \\ \mu\alpha + (1-\alpha) \end{pmatrix} \frac{y_0 - \left[\alpha x_1^b + (1-\alpha)x_2^b\right]}{\left[\alpha^2 + 2\alpha(1-\alpha)\mu + (1-\alpha)^2\right]\sigma_b^2 + \sigma_0^2}$$

Case 1: No cross-correlation between two grid points, $\mu = 0$ and $\alpha = 1$

$$\begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} \alpha + \mu(1-\alpha) \\ \mu\alpha + (1-\alpha) \end{pmatrix} \frac{y_0 - \left[\alpha x_1^b + (1-\alpha) x_2^b\right]}{\left[\alpha^2 + 2\alpha(1-\alpha)\mu + (1-\alpha)^2\right]\sigma_b^2 + \sigma_0^2}$$

$$\Rightarrow \begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \frac{y_0 - x_1^b}{\sigma_b^2 + \sigma_0^2}$$

$$x_1^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x_1^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

$$x_2^a = x_2^b$$

The analysis at grid point 1 is same as the analysis of the previous example
The analysis at grid point 2 is equal to the background. Observation had no effect.

Case 2: $\alpha = 1, \quad \mu \neq 0$

$$\begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} \alpha + \mu(1-\alpha) \\ \mu\alpha + (1-\alpha) \end{pmatrix} \frac{y_0 - \left[\alpha x_1^b + (1-\alpha) x_2^b\right]}{\left[\alpha^2 + 2\alpha(1-\alpha)\mu + (1-\alpha)^2\right]\sigma_b^2 + \sigma_0^2}$$

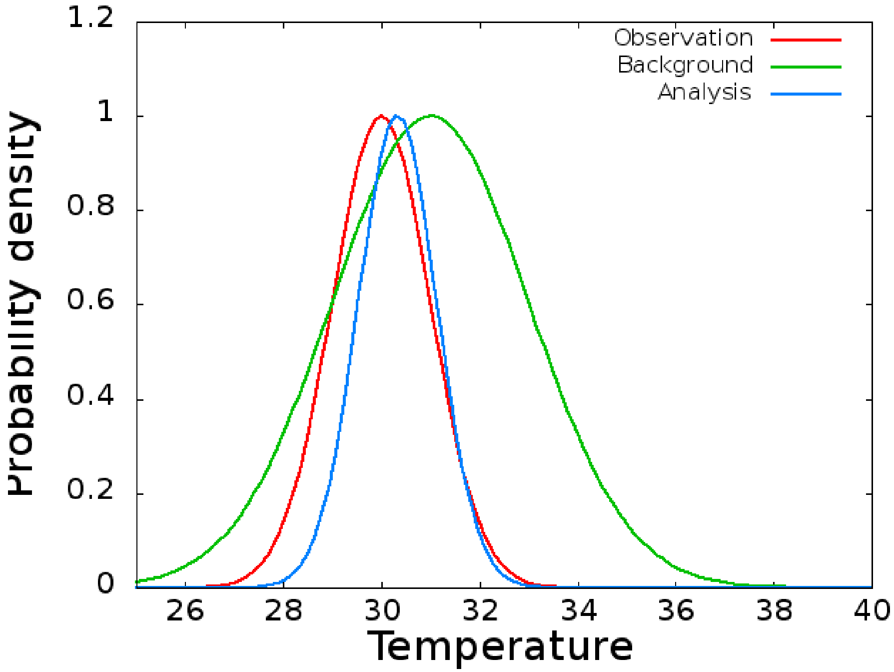
$$\Rightarrow \begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} = \begin{pmatrix} x_1^b \\ x_2^b \end{pmatrix} + \sigma_b^2 \begin{pmatrix} 1 \\ \mu \end{pmatrix} \frac{y_0 - x_1^b}{\sigma_b^2 + \sigma_0^2}$$

$$x_1^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x_1^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

$$x_2^a = x_2^b + \mu\sigma_b^2 \frac{y_0 - x_1^b}{\sigma_0^2 + \sigma_b^2}$$

Now the solution at grid point 2 is influenced by the observation. The role of Background error covariance is to spread information from one grid point to the other.

PRACTICAL ISSUES



$$x^a = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_b^2} x^b + \frac{\sigma_b^2}{\sigma_0^2 + \sigma_b^2} y_0$$

If $\sigma_b >> \sigma_0$

$x^a \approx y_0$

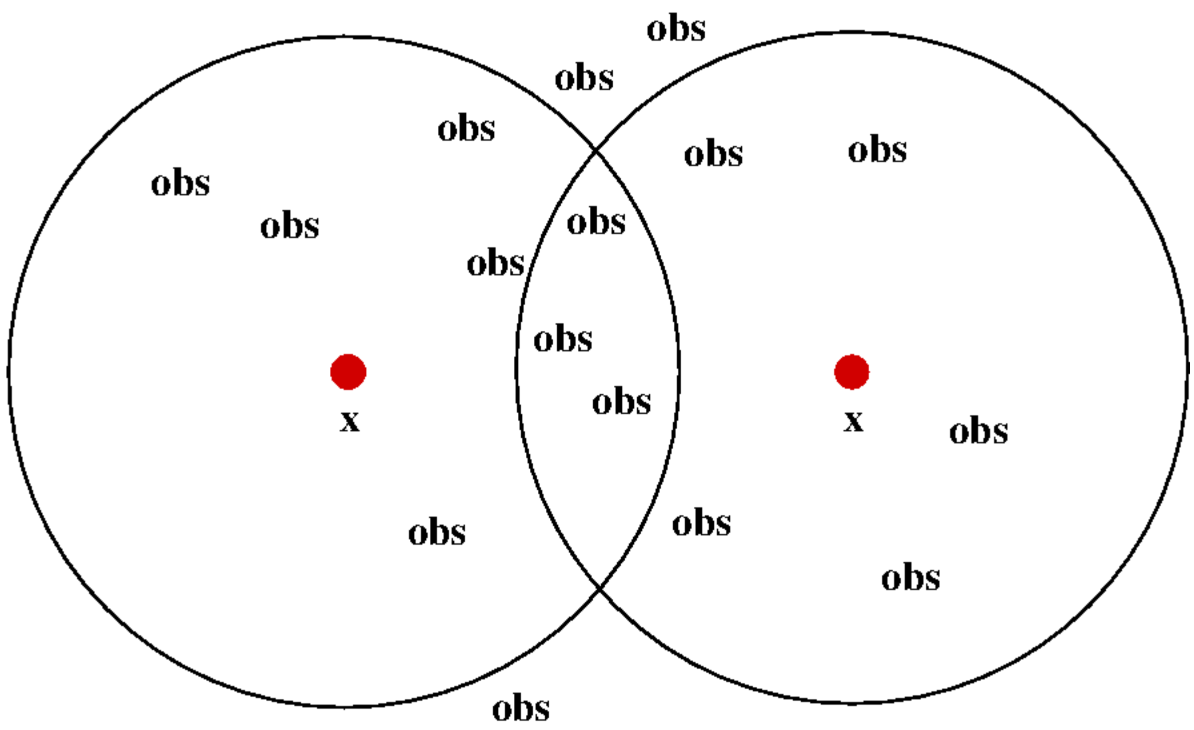
If $\sigma_0 >> \sigma_b$

$x^a \approx x_b$

COVARIANCE INFLATION IS NECESSARY !!!

Assimilating distant observations leads to spurious correlations

Idea of Localization



PRACTICAL APPLICATIONS IN INCOIS

System Comparison

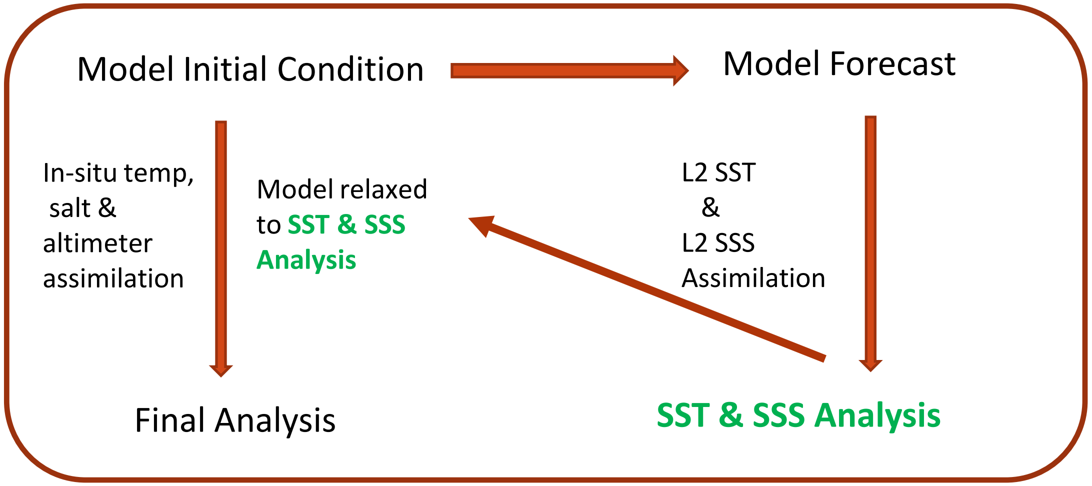
	INCOIS GODAS	LETKF-MOM
Model	MOM4.0	MOM4.1
Resolution	0.5 degree in zonal 0.25 degree in meridional between 10S & 10N.	0.5 degree in zonal 0.25 degree in meridional between 10S & 10N.
No. of levels	40	40
Features	• River RunOff ON • Relaxed to REYNOLDS SST • SSS not relaxed	• River RunOff OFF • Relaxed to LETKF derived SST • Relaxed to LETKF derived SSS
Assimilation Scheme	3D VAR	LETKF with 56 ensemble members
Observations	• In-situ temp & salinity	• In-situ temp & salinity • Altimeter • L2 SST & L2 SSS
Initial Condition	Well Trained	2003 IC pretended as 2010
Run Period	2002 onwards till date	Jan 03, 2010 – June 15, 2010

Why Local Ensemble Transform Kalman Filter (LETKF) ?

- LETKF is a more advanced assimilation technique and has been proved to work better and faster than Ensemble Kalman Filter for the same number of ensemble members.
- LETKF can account for more growing modes in the local subspace thereby making it a potential better technique.

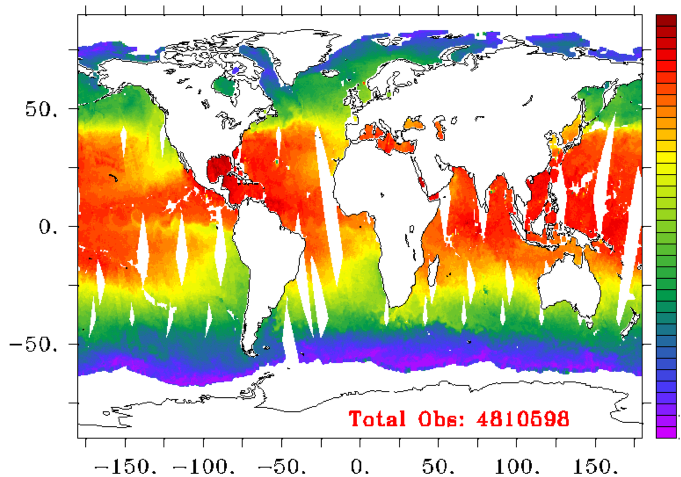
In what sense is this indigenous ?

- We are the first to assimilate L2 SST and L2 SSS in LETKF. (Disclaimer : As far as we are aware of)
- Direct SST assimilation is degrading the vertical temperature profile.
- The algorithm that goes into the assimilation scheme.

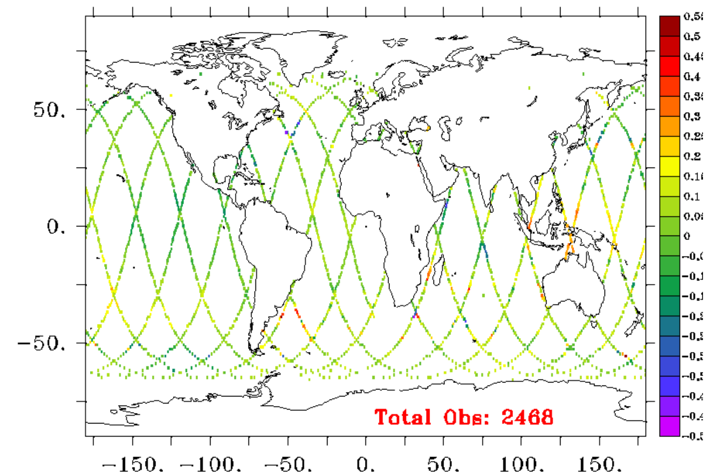


Observations

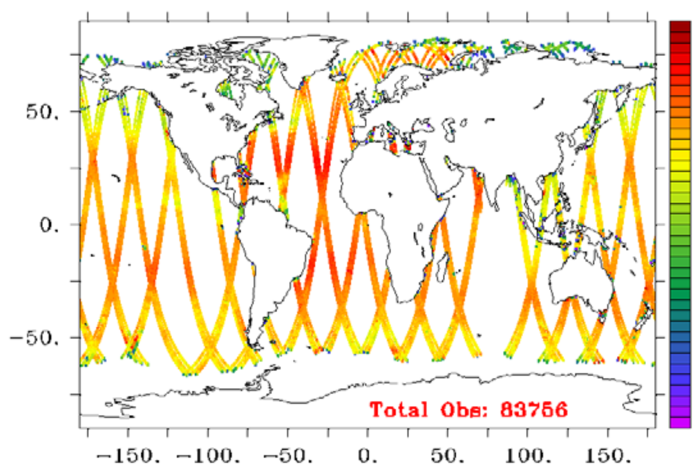
L2 SST



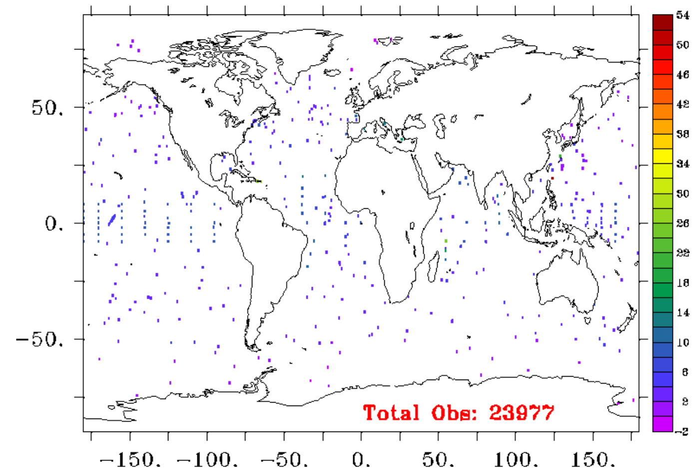
L2 SLA



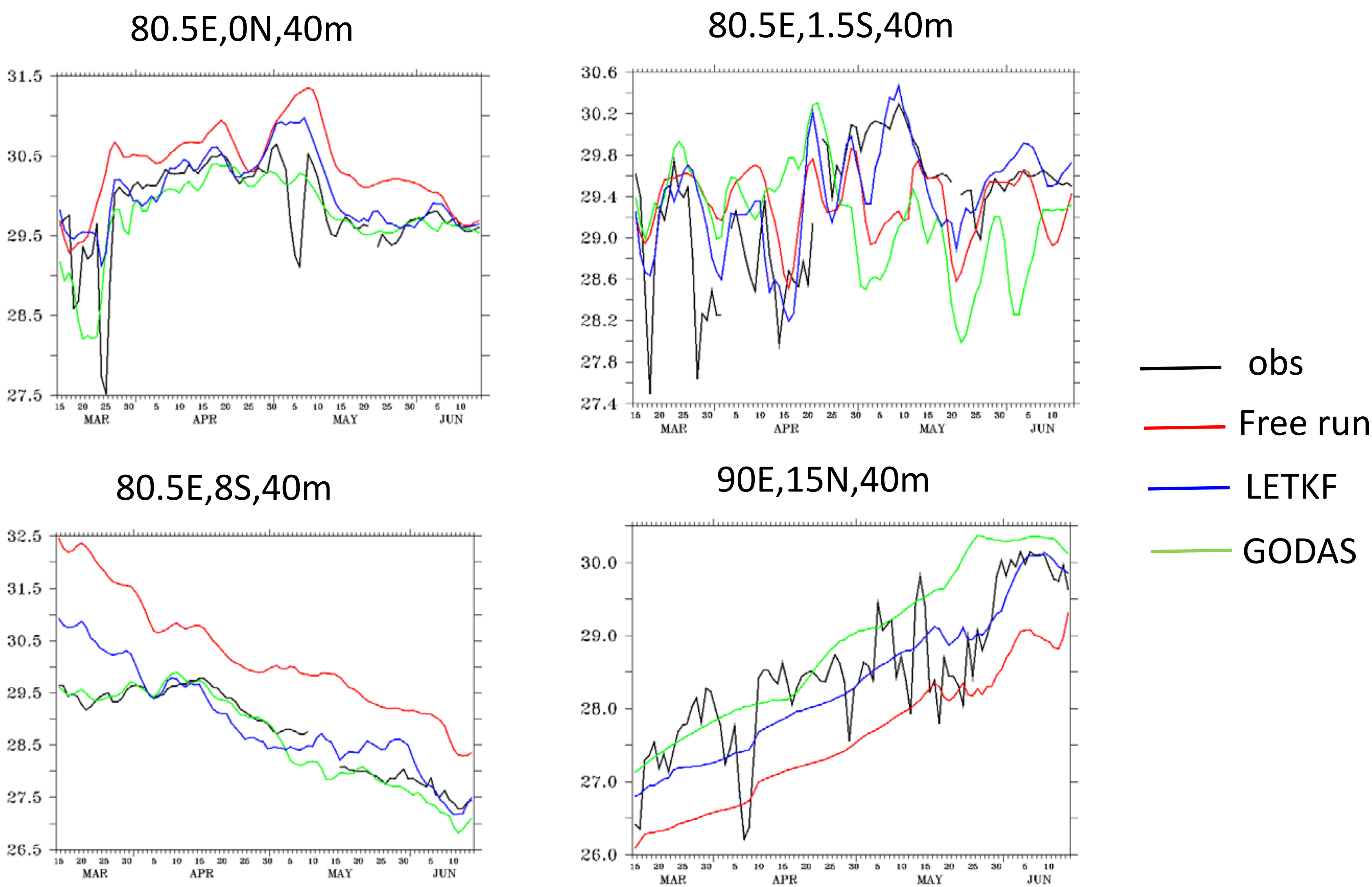
L2 SSS



In-situ T & S

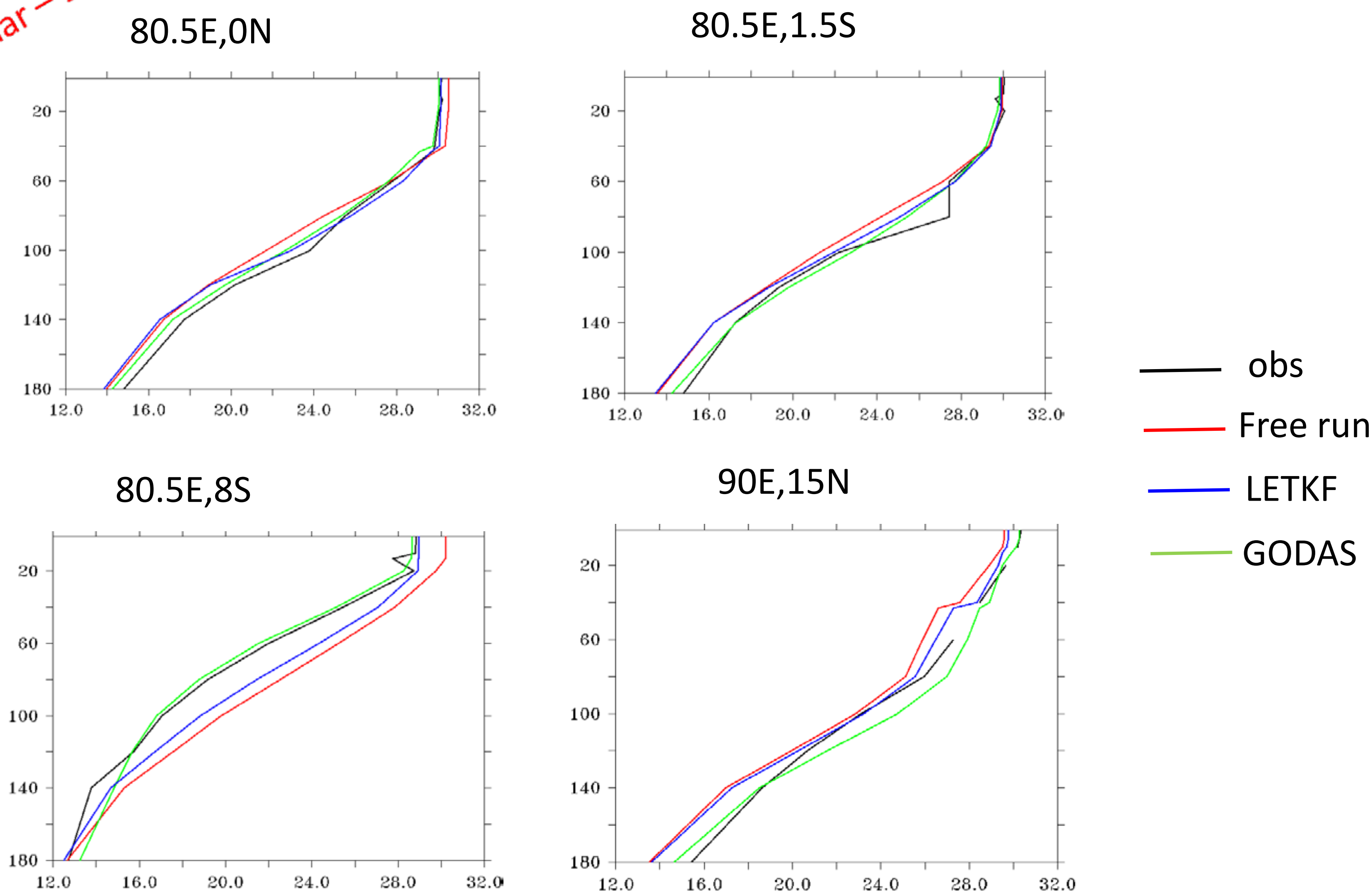


TEMP Time Series Results

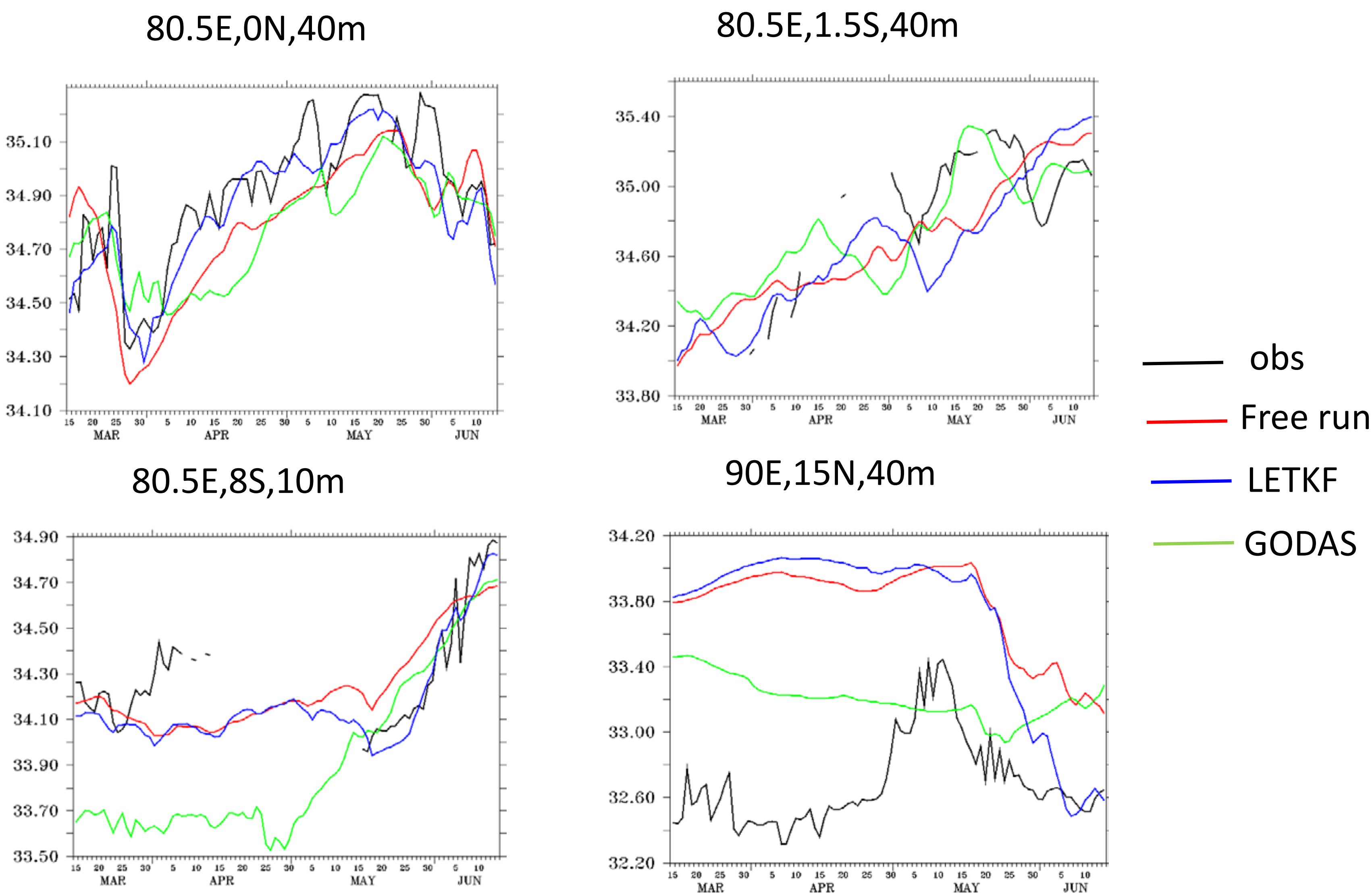


TEMP Vertical Profiles

Averaged from
15 mar – 15 jun, 2010

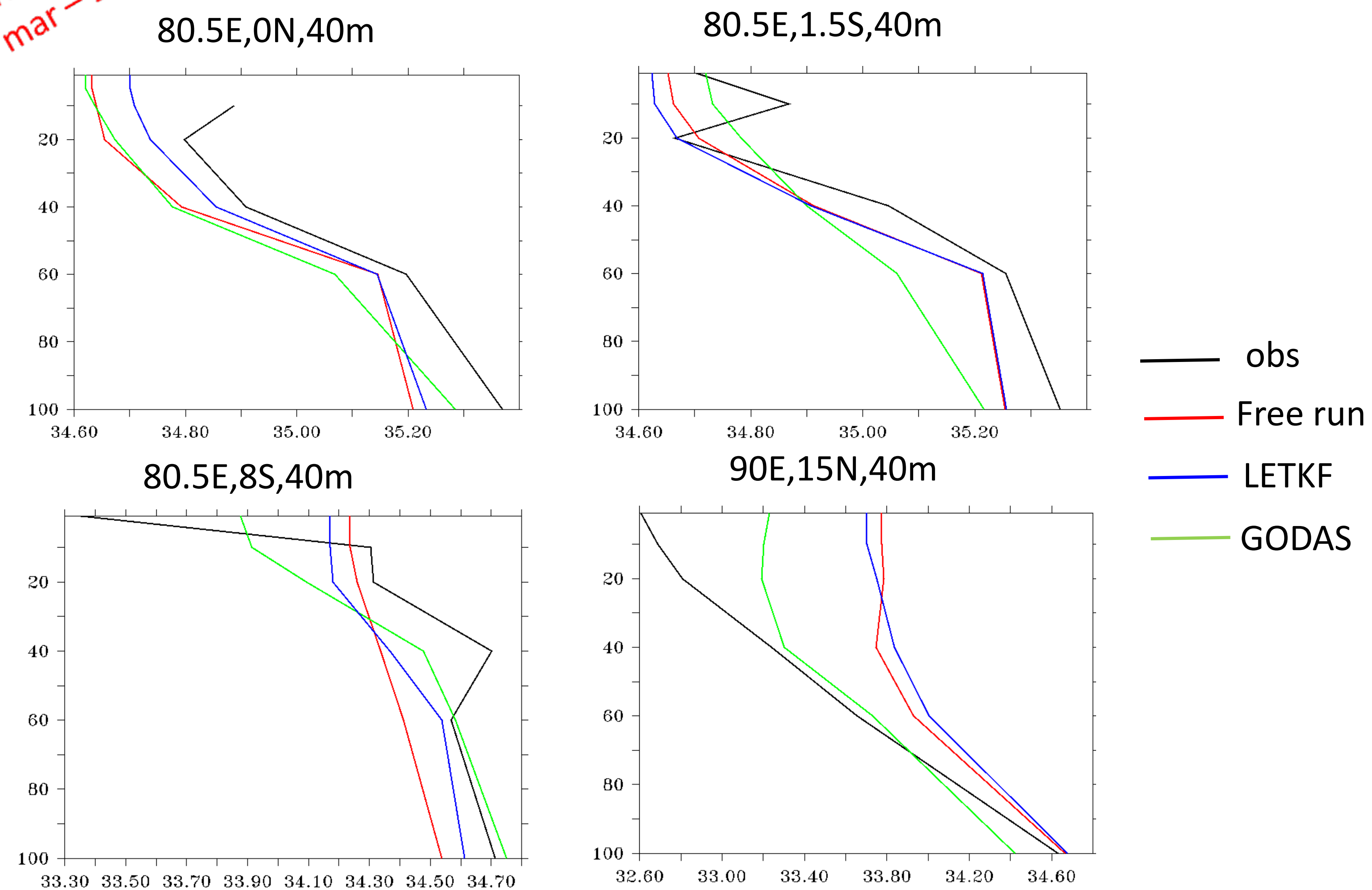


SALT Time Series Results



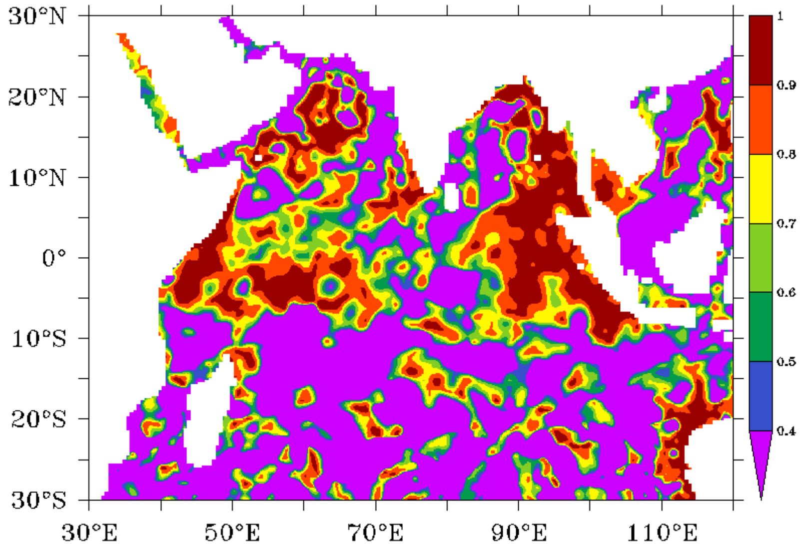
SALT Vertical Profiles

Averaged from
15 mar – 15 jun, 2010

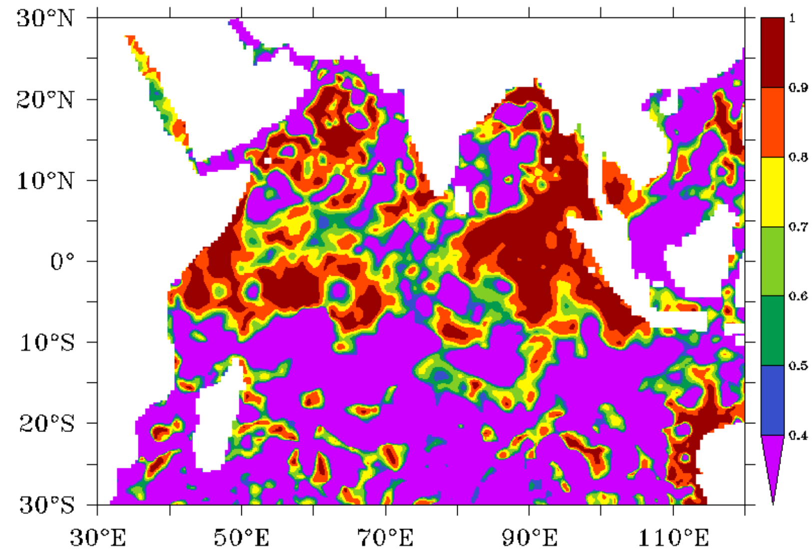


SLA Correlations with AVISO

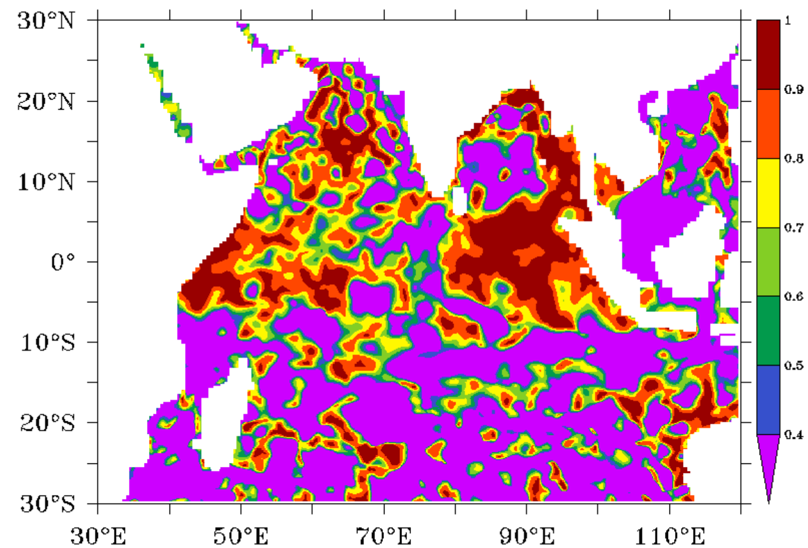
Free Model



LETKF-MOM

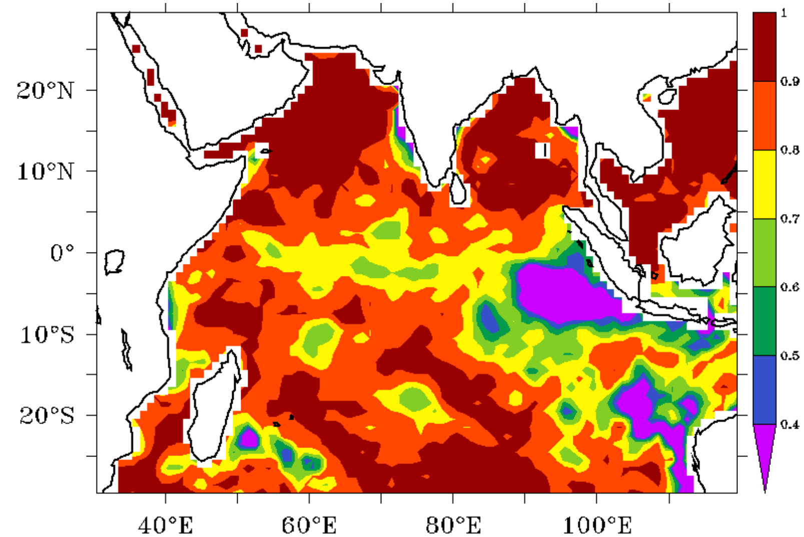


GODAS

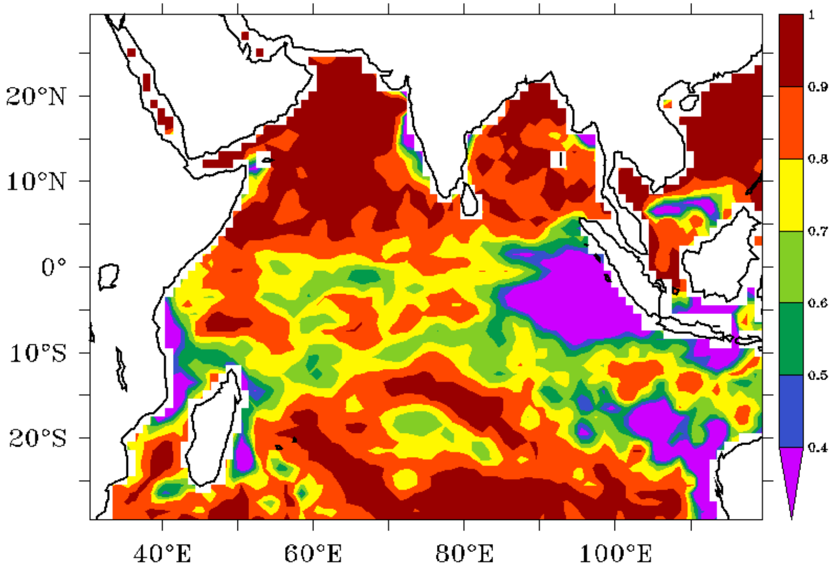


SST Correlations with REYNOLDS

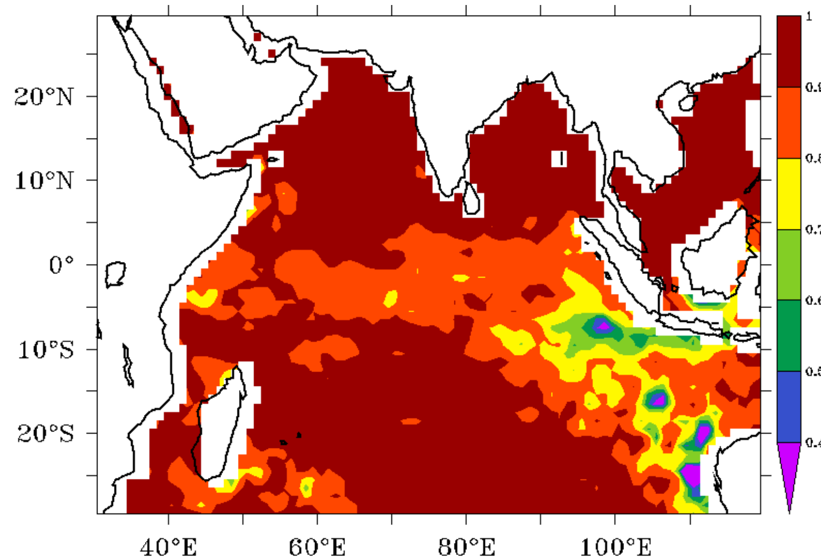
LETKF-MOM



Free Model



GODAS



Conclusions

- LETKF-MOM performance is comparable to GODAS in many aspects even though it starts from a gross initial condition.

- SST assimilation is not good because of large R. Decrease R to get better assimilation.

- It's premature to say whether LETKF can outperform GODAS. But it's definitely promising.

TAKE HOME MESSAGE

- The truth is not known.
- Neither observation nor model is devoid of errors.
- Assimilate these two to get a best estimate.
- Estimating maximum likelihood = Minimizing cost function.
- The model error covariance propagates information from one place to another.
- Covariance inflation is necessary for Ensemble based schemes.
- Localize observations to get rid of spurious correlations.


THANK
YOU!