Model bias in data assimilation

Patrick Laloyaux

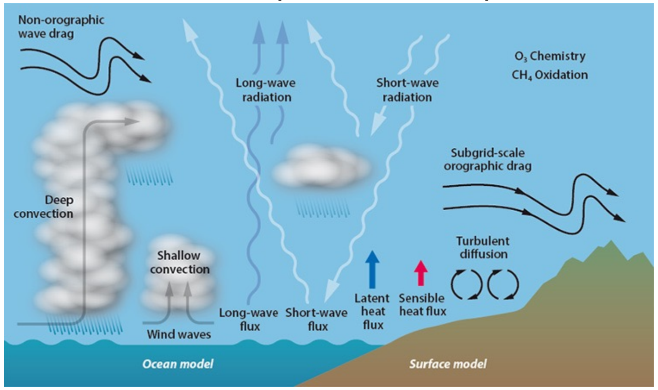
Identify systematic errors in model (biases)

Learn how to develop bias correction methods

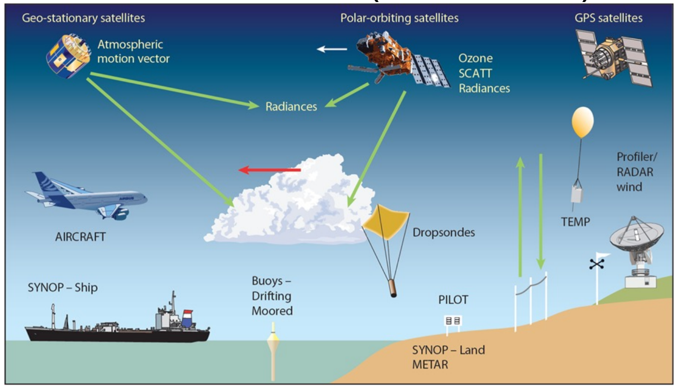
Prospect for future developments

What you have seen so far on data assimilation

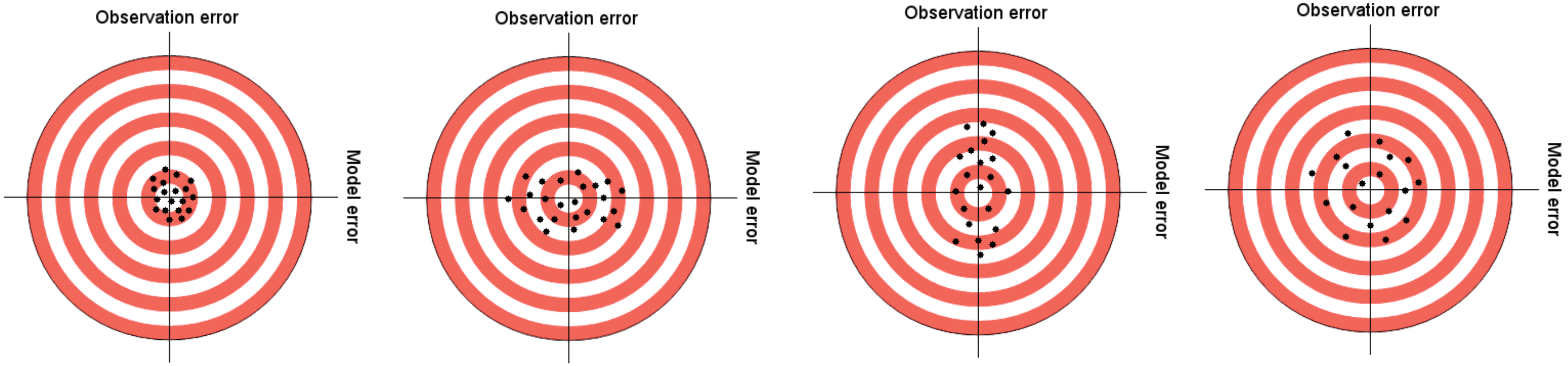
Model (with errors)



Observations (with errors)

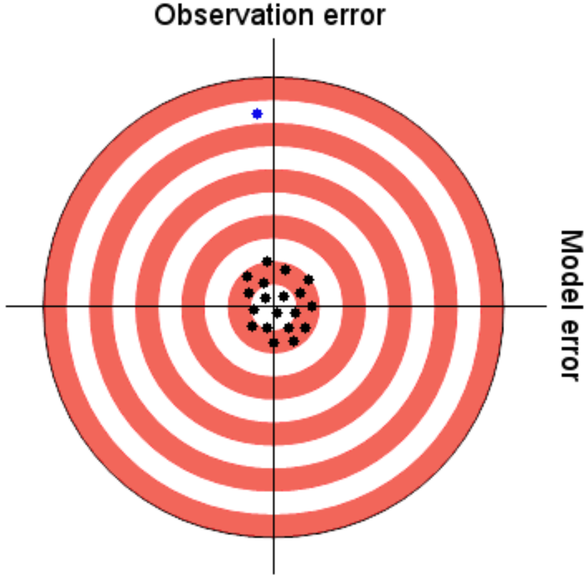


If you are lucky, model and observations are not biased

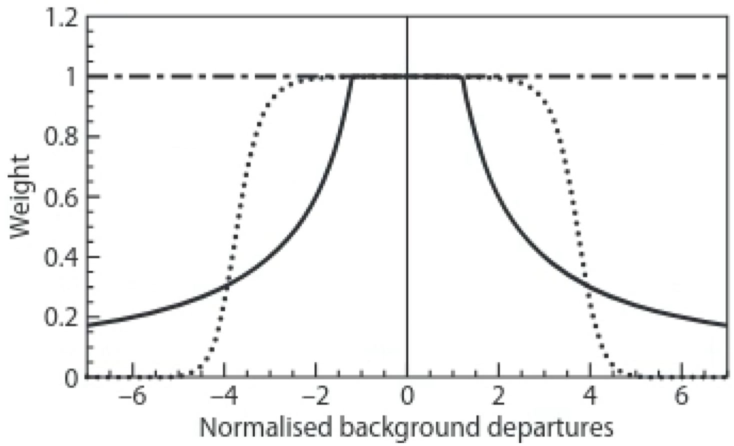


$$
\begin{aligned}
J(x_0) &= \frac{1}{2}(x_0 - x_b)^T \mathbf{B}^{-1}(x_0 - x_b) \\
&+ \frac{1}{2}\sum_{k=0}^{K}[y_k - \mathcal{H}(x_k)]^T \mathbf{R}_k^{-1}[y_k - \mathcal{H}(x_k)]
\end{aligned}
$$

What you have seen so far on data assimilation



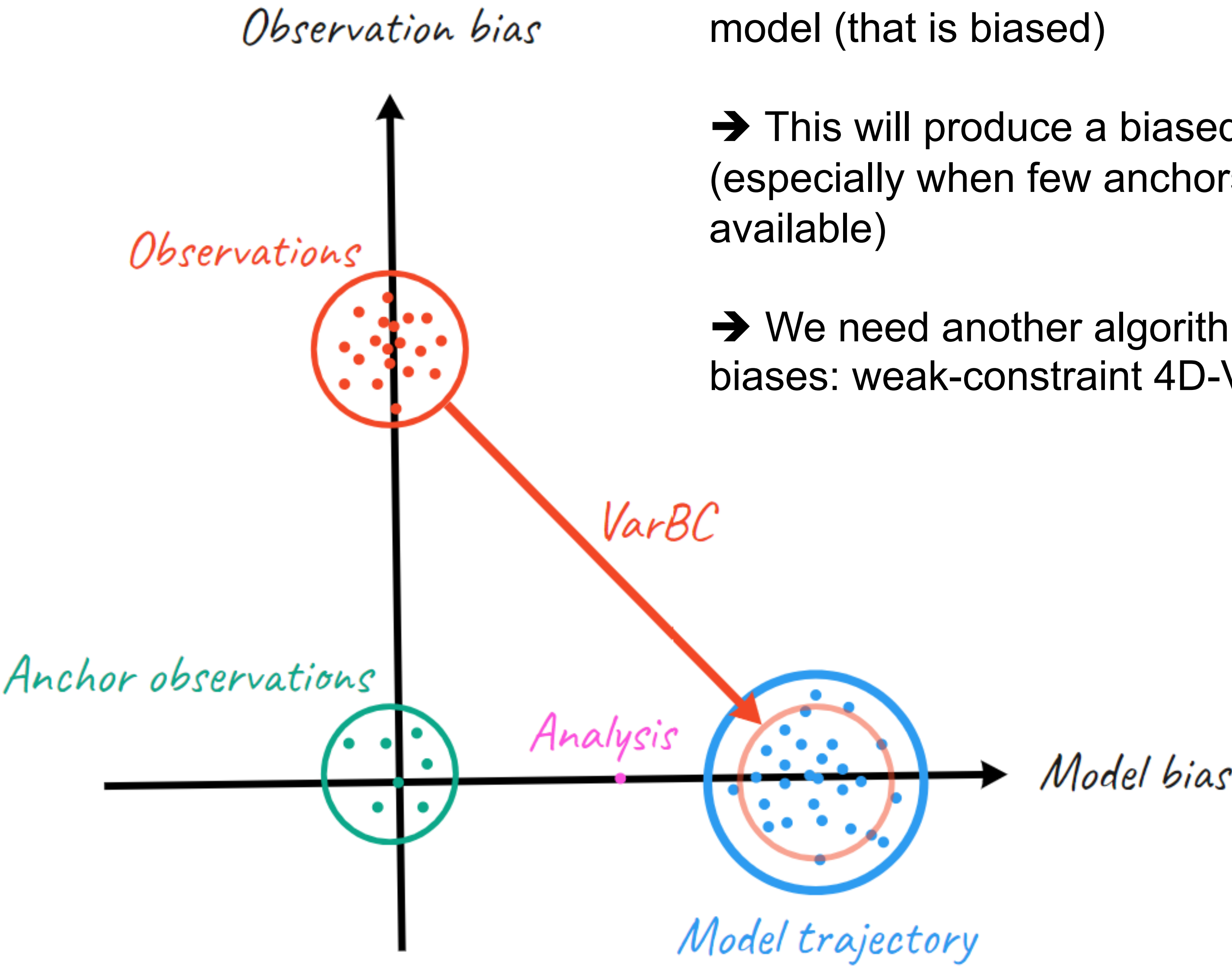
- ➔ Outliers
- ➔ Variational Quality Control (VarQC)



- ➔ Precise but not accurate
- ➔ Variational Bias Control (VarBC)

$$
\begin{aligned}
J(x_0,\beta) &= \frac{1}{2}(x_0-x_b)^T\mathbf{B}^{-1}(x_0-x_b) \\
&+ \frac{1}{2}(\beta-\beta_b)^T\mathbf{B}_\beta{}^{-1}(\beta-\beta_b) \\
&+ \frac{1}{2}\sum_{k=0}^{\text{Radiosonde}}[y_k-\mathcal{H}(x_k)]^T\mathbf{R}_k^{-1}[y_k-\mathcal{H}(x_k)] \\
&+ \frac{1}{2}\sum_{k=0}^{\text{GPSRO}}[y_k-\mathcal{H}(x_k)]^T\mathbf{R}_k^{-1}[y_k-\mathcal{H}(x_k)] \\
&+ \frac{1}{2}\sum_{k=0}^{\text{Others}}[y_k-b(x_k,\beta)-\mathcal{H}(x_k)]^T\mathbf{R}_k^{-1}[y_k-b(x_k,\beta)-\mathcal{H}(x_k)]
\end{aligned}
$$

What happens when VarBC is used with a biased model

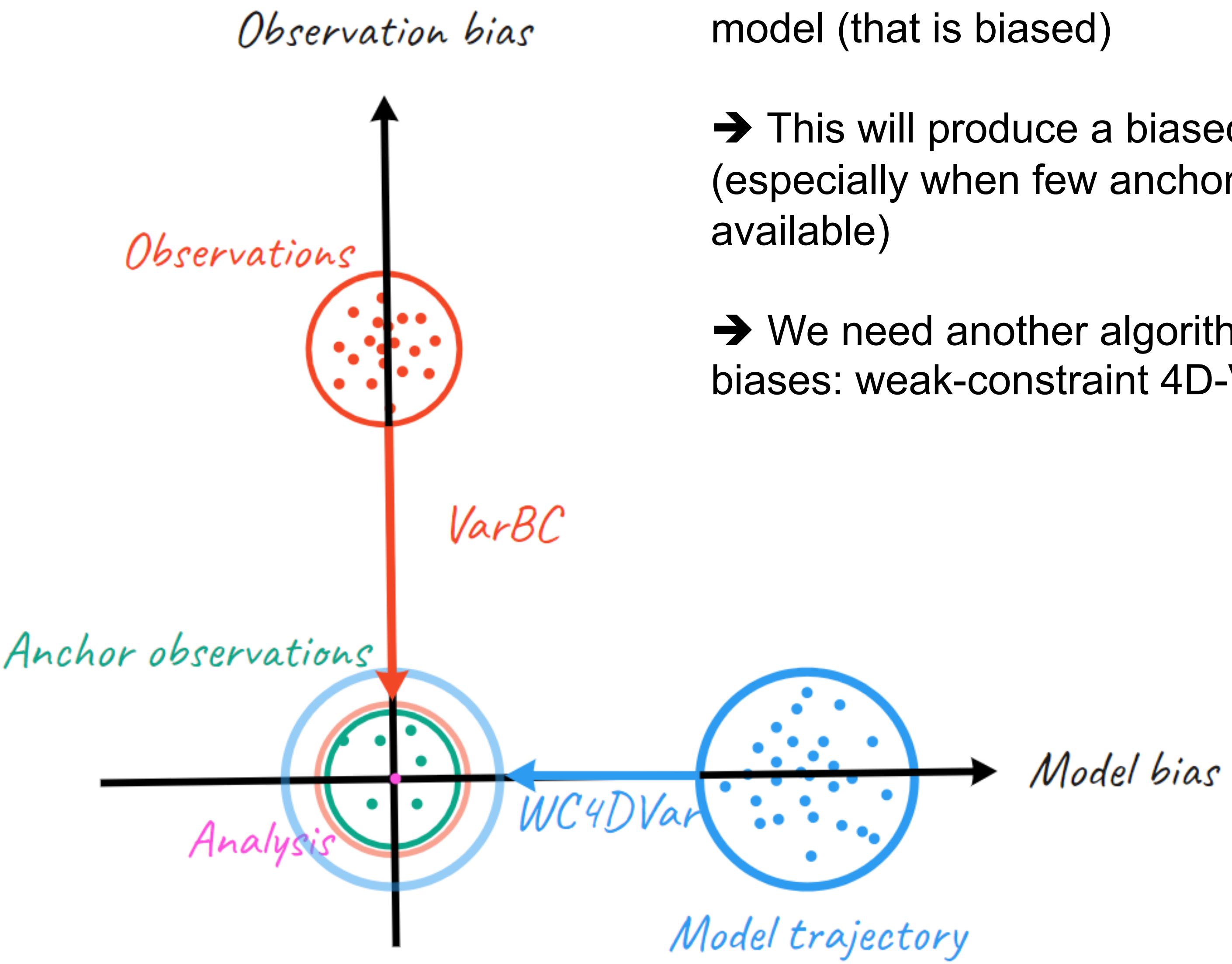


VarBC corrects the observations towards the model (that is biased)

➔ This will produce a biased analysis (especially when few anchors observations are available)

➔ We need another algorithm to handle model biases: weak-constraint 4D-Var

What happens when VarBC is used with a biased model



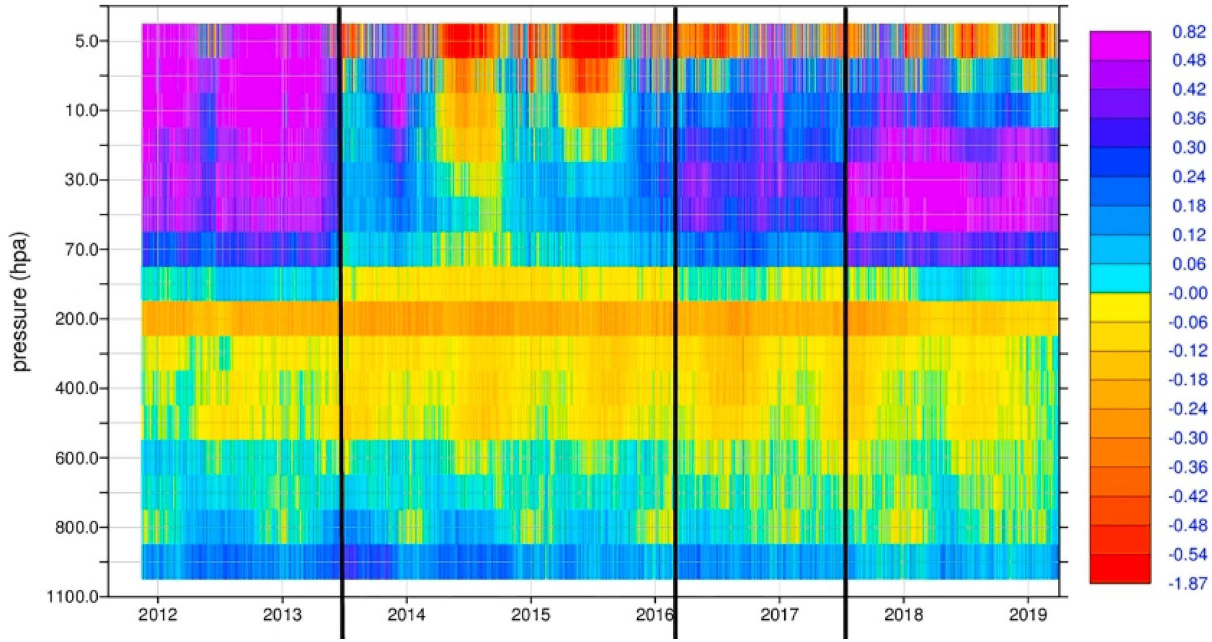
VarBC corrects the observations towards the model (that is biased)

➔ This will produce a biased analysis (especially when few anchors observations are available)

➔ We need another algorithm to handle model biases: weak-constraint 4D-Var

How to estimate model biases

The first-guess trajectory of the model can be compared to accurate observations



Difference between radiosonde temperature observations and the IFS first-guess trajectory (O-B)

Errors in models are often systematic rather than random, zero-mean

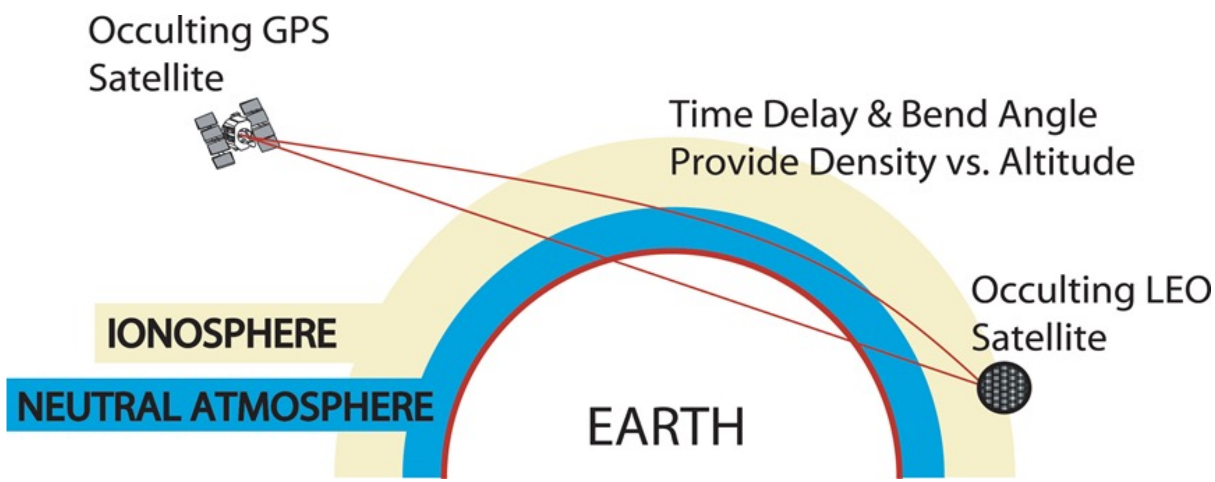
→ Largest bias in the stratosphere

→ Model has a temperature cold bias in the lower/mid stratosphere

→ Model has a warm bias in the upper stratosphere

How to estimate model biases

The GPS satellites are used for positioning and navigation. GPS-RO (Radio Occultation) is based on analysing the bending caused by the atmosphere along paths between a GPS satellite and a receiver placed on a low-earth-orbiting satellite.



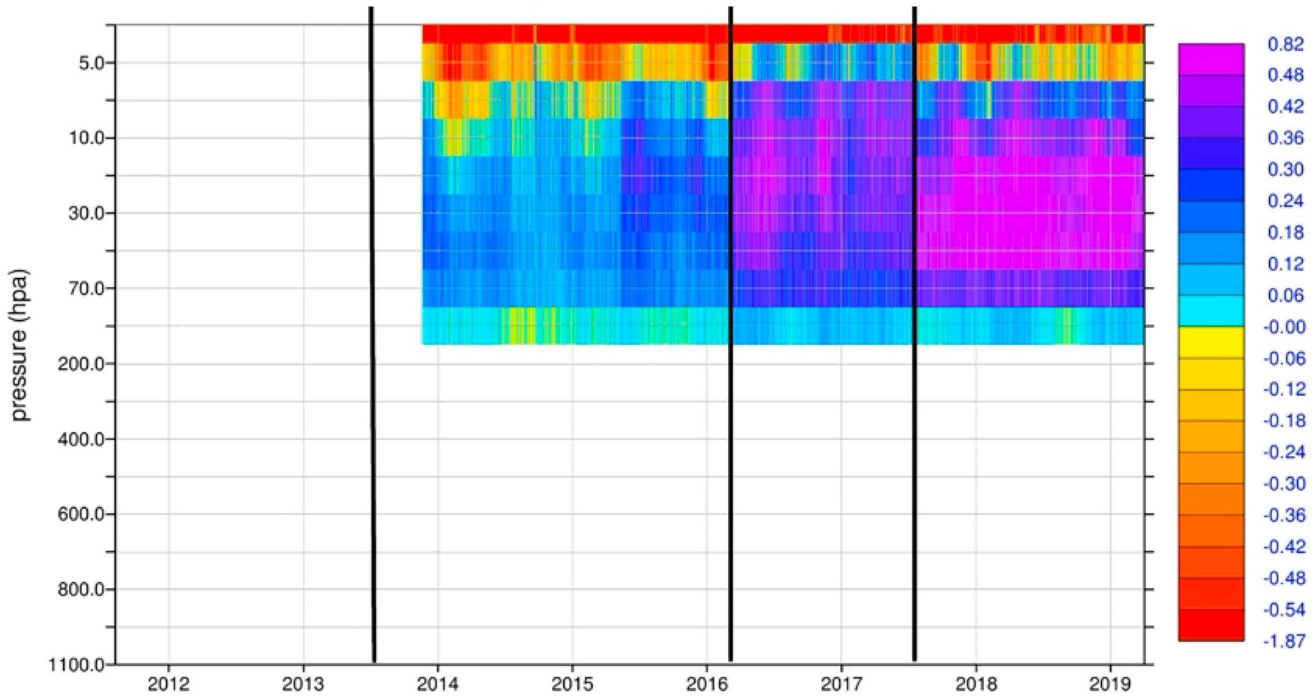
→ As the LEO moves behind the earth, we obtain a profile of bending angles

→ Temperature profiles can then be derived

→ GPS-RO can be assimilated without bias correction. They are good for highlighting errors/biases

How to estimate model biases

The first-guess trajectory of the model can be compared to accurate observations



Difference between GPS-RO temperature retrievals and the IFS first-guess trajectory (O-B)

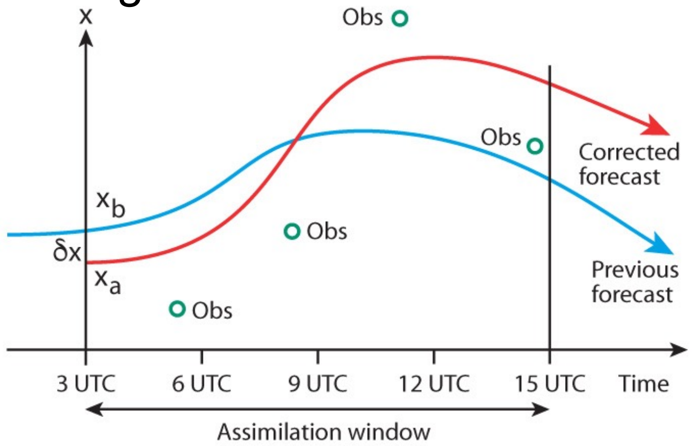
Errors in models are often systematic rather than random, zero-mean

→ Model has a temperature cold bias in the lower/mid stratosphere

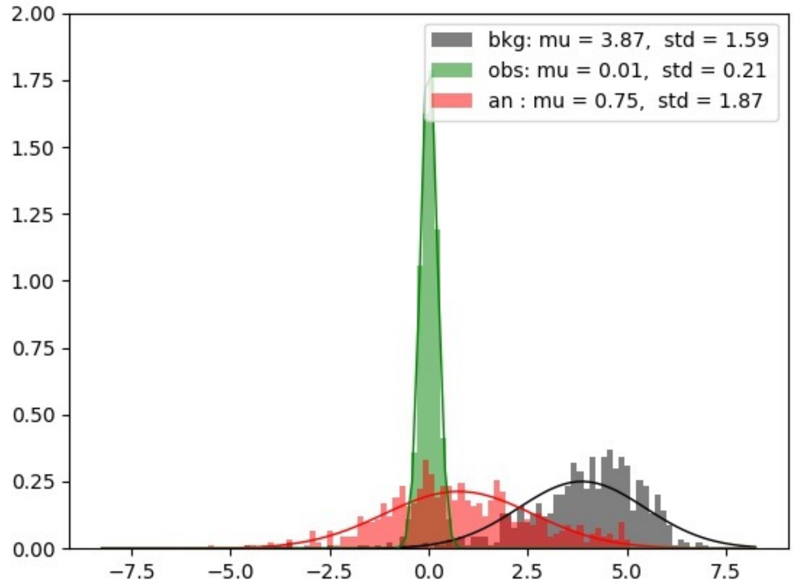
→ Model has a warm bias in the upper stratosphere

How to deal with model biases in data assimilation

Strong constraint 4D-Var

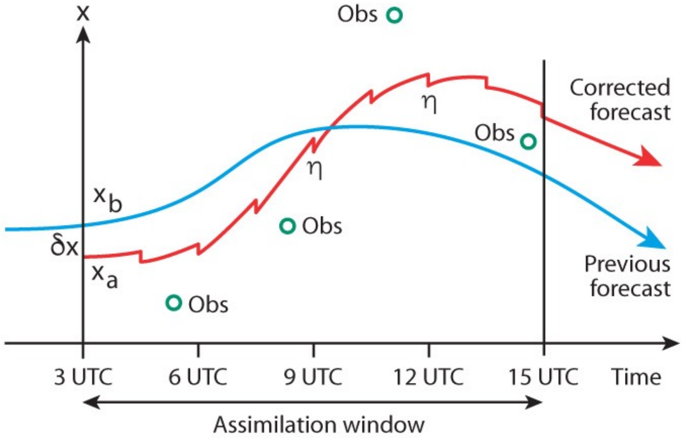


$$x_k = \mathcal{M}_k(x_{k-1})$$

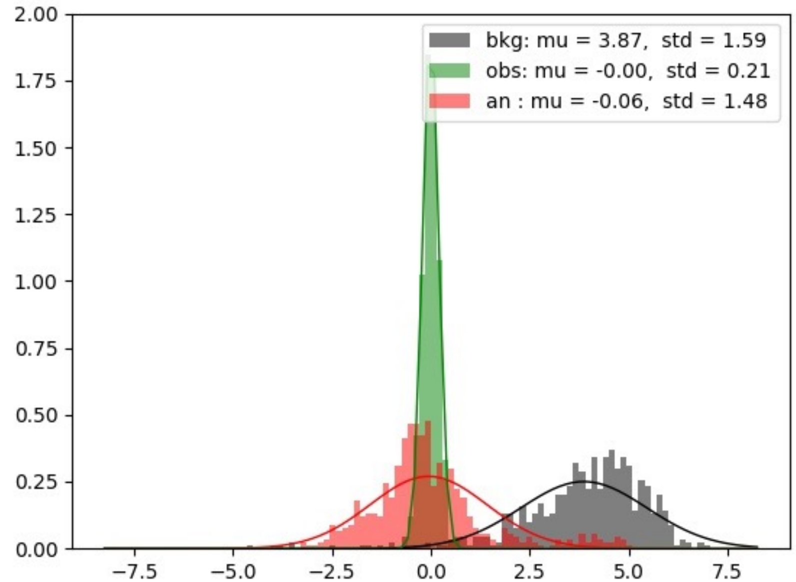


→ Large bias and standard deviation in the analysis

Weak constraint 4D-Var



$$x_k = \mathcal{M}_k(x_{k-1}) + \eta \quad \text{for } k = 1, 2, \cdots, K$$



→ Bias in the analysis has been reduced, standard deviation as well

Weak constraint 4D-Var

We assume that the model is not perfect, adding an error term η in the model equation

$$x_k = \mathcal{M}_k(x_{k-1}) + \eta \quad \text{for } k = 1, 2, \cdots, K$$

The model error estimate η contains 3 physical 3D fields

- temperature
- vorticity
- divergence

Constant model error forcing over the assimilation window to correct the model bias

$$\begin{aligned} J(\underbrace{x_0}_{\text{Model state}}, \underbrace{\beta}_{\text{Observation bias}}, \underbrace{\eta}_{\text{Model bias}}) &= \frac{1}{2}(x_0 - x_b)^T \mathbf{B}^{-1}(x_0 - x_b) \\ &+ \frac{1}{2}\sum_{k=0}^{K}[y_k - \mathcal{H}(x_k) - b(x_k, \beta)]^T \mathbf{R}_k^{-1}[y_k - \mathcal{H}(x_k) - b(x_k, \beta)] \\ &+ \frac{1}{2}(\beta - \beta_b)^T \mathbf{B}_\beta{}^{-1}(\beta - \beta_b) \\ &+ \frac{1}{2}(\eta - \eta_b)^T \mathbf{Q}^{-1}(\eta - \eta_b) \end{aligned}$$

→ Introduce additional controls to target an unbiased analysis

→ The model error covariance matrix Q constrains the model error field

→ This looks very much like VarBC with a constant predictor, but in the model space!

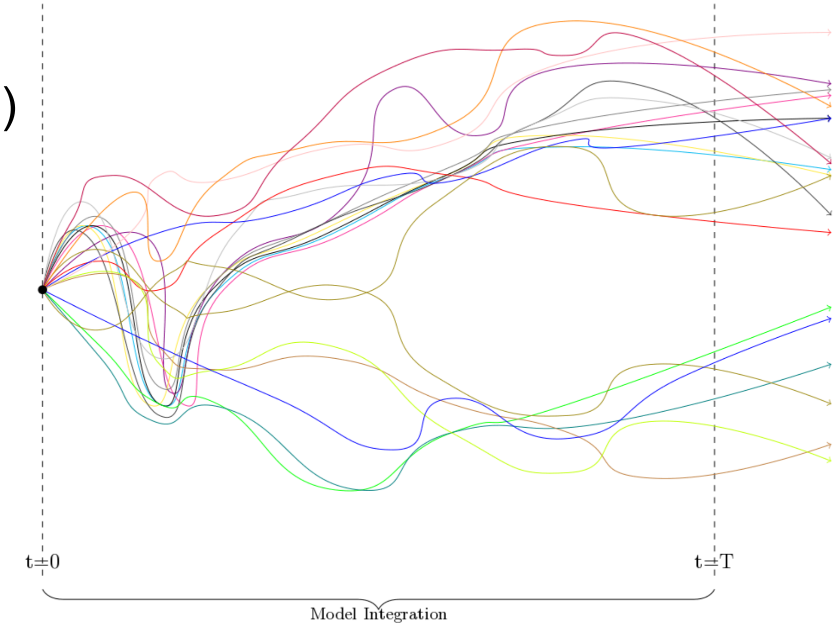
How to estimate the model error covariance matrix (Q)

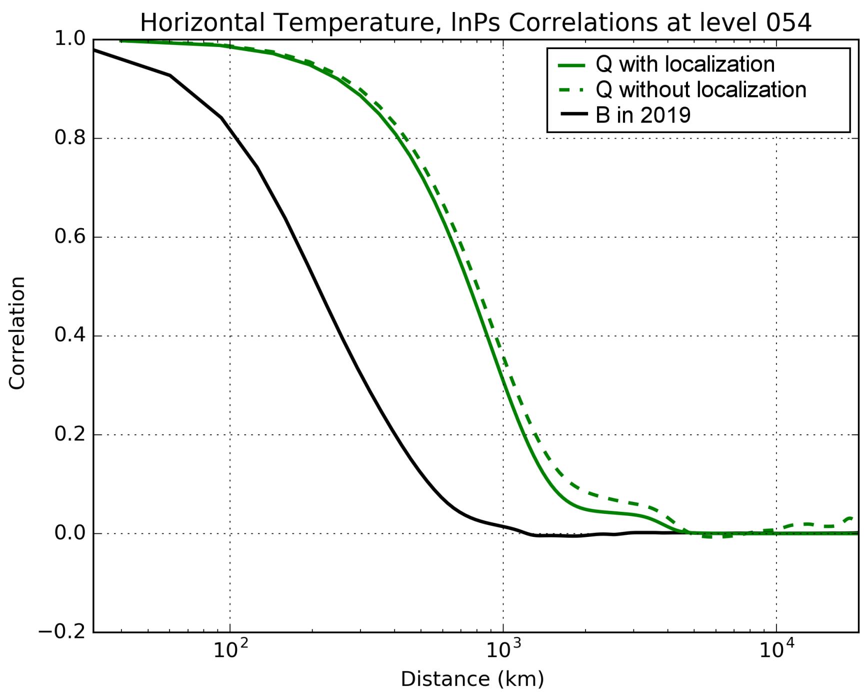
Estimate the model error covariance matrix

➔ run the ensemble forecasting system (ENS) with perturbed physics (51 members with the same initial condition for different days)

➔ differences after 12 hours are used to compute Q

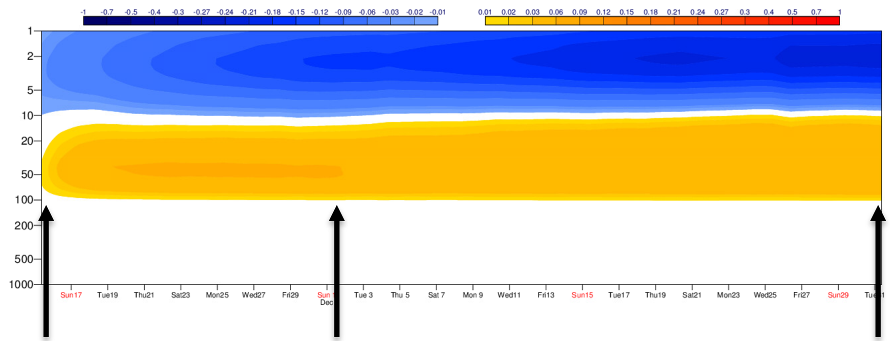
$$Q_f = \frac{1}{N-1}\sum_{i=1}^{N}\left(f_i^{12} - f_{i+1}^{12}\right)\left(f_i^{12} - f_{i+1}^{12}\right)^{\mathrm{T}}$$

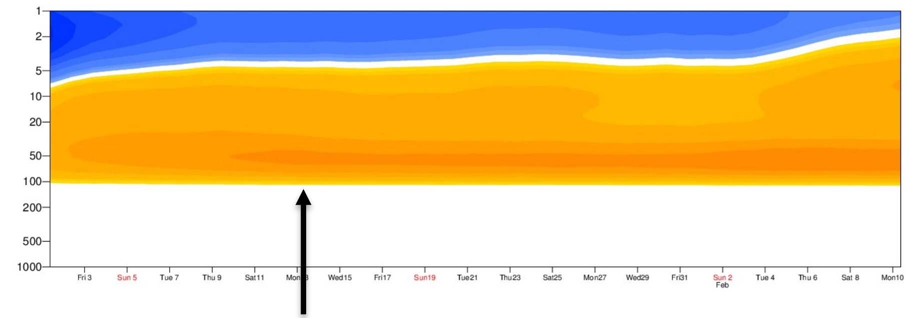




4D-Var corrects small scale errors (background errors) by changing the initial condition and large scale errors (model errors) by changing the model forcing

How fast does weak-constraint 4D-Var learn?



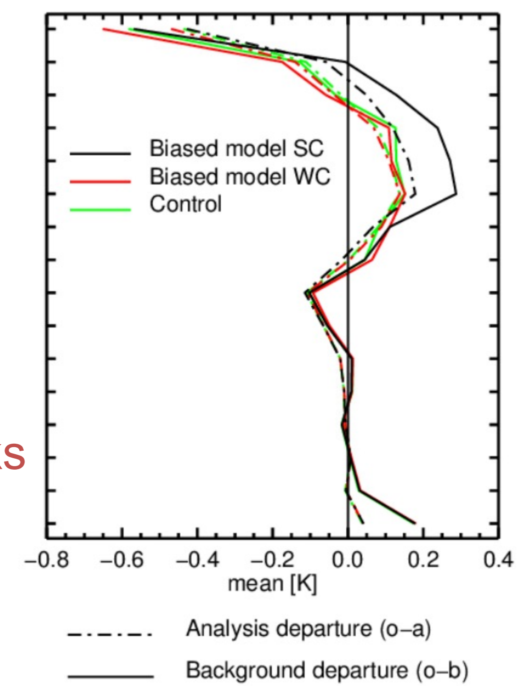


Weak-constraint 4D-Var is cold started (model error is zero)

After 2 weeks, the model error estimate is steady

Introduce a (bad) model change (vertical finite element replaced by vertical finite difference)

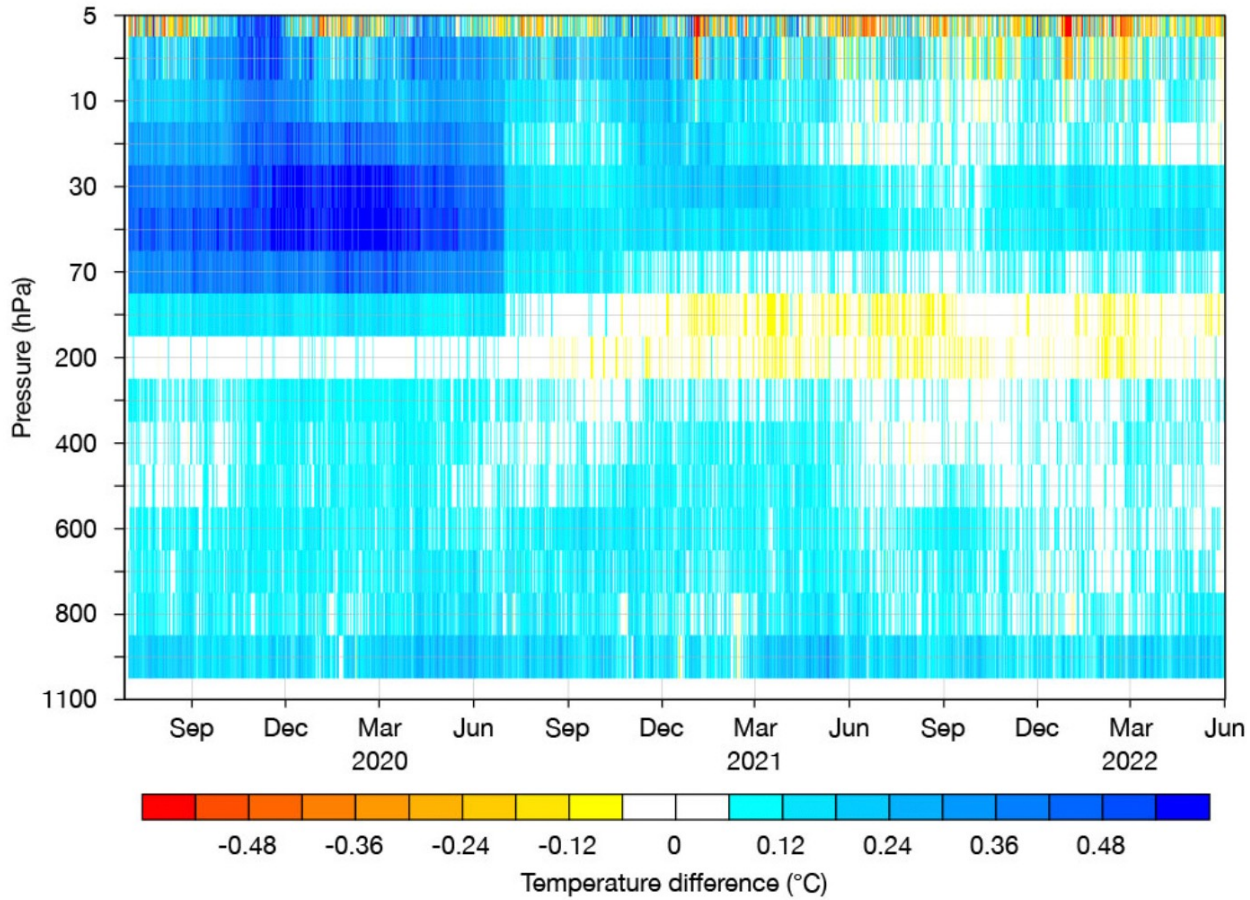
Weak-constraint 4D-Var learns the new bias



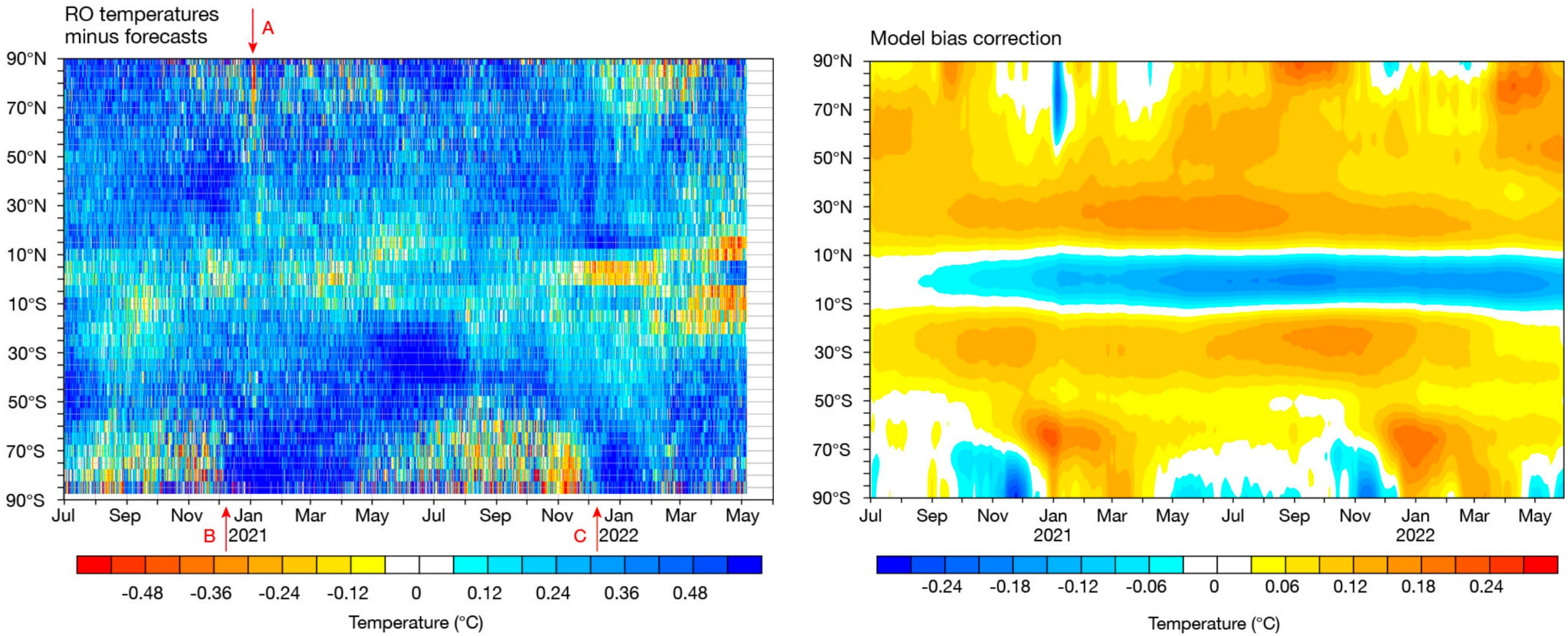
➔The fit to the observations is not degraded since weak-constraint 4D-Var learns the new model error quickly (thanks to anchoring observations)

Weak-constraint 4D-Var in operations for the stratosphere

Time series of the difference between radiosonde temperature observations and model first-guess (47r1 implemented on 30 June 2020)



Weak-constraint 4D-Var in operations for the stratosphere



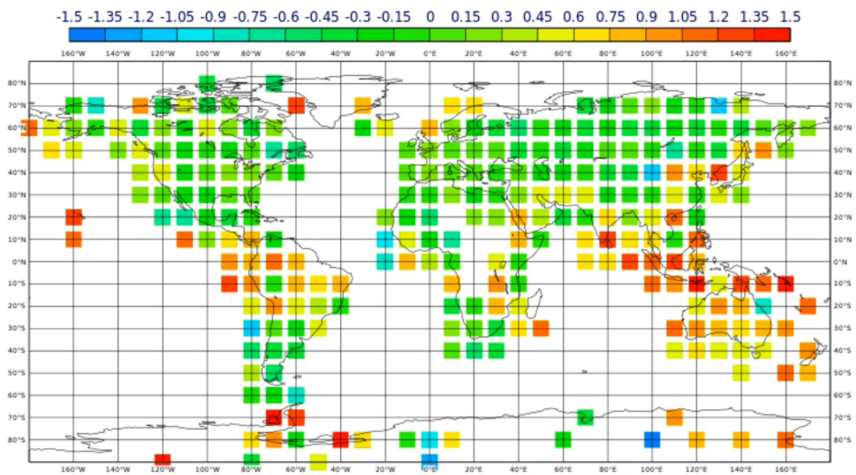
A) On 31 December 2020, a Sudden Stratospheric Warming (SSW) event started over the northern hemisphere

B&C) Clear seasonal cycle in the model bias over the southern hemisphere with a sharp transition in early December 2020 and 2021

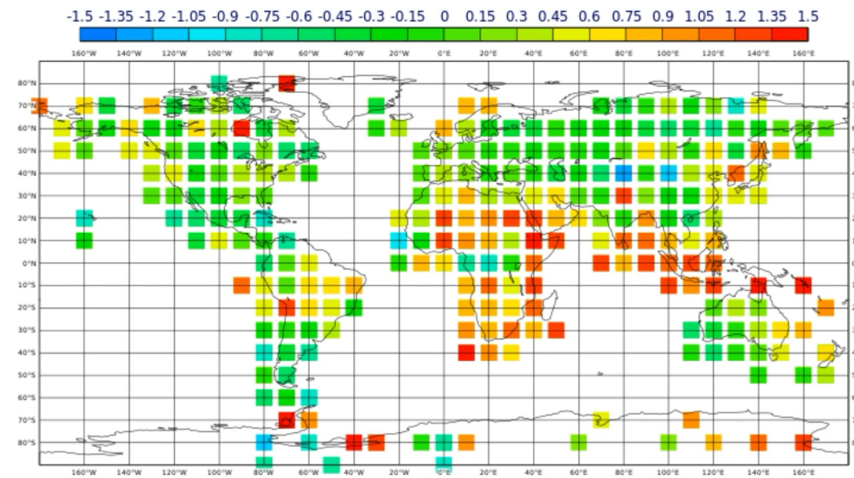
Model biases in the boundary layer

Several diagnostics shows that the structure of model biases is time-correlated

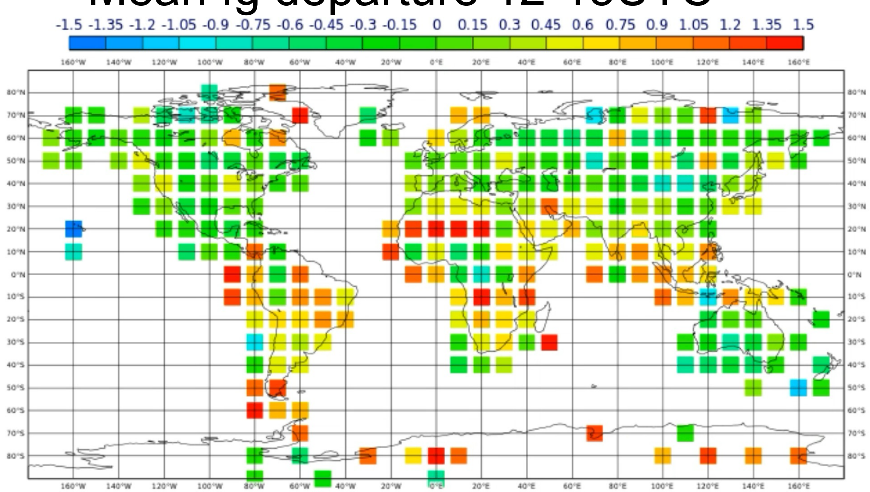
Mean fg departure 00-03UTC



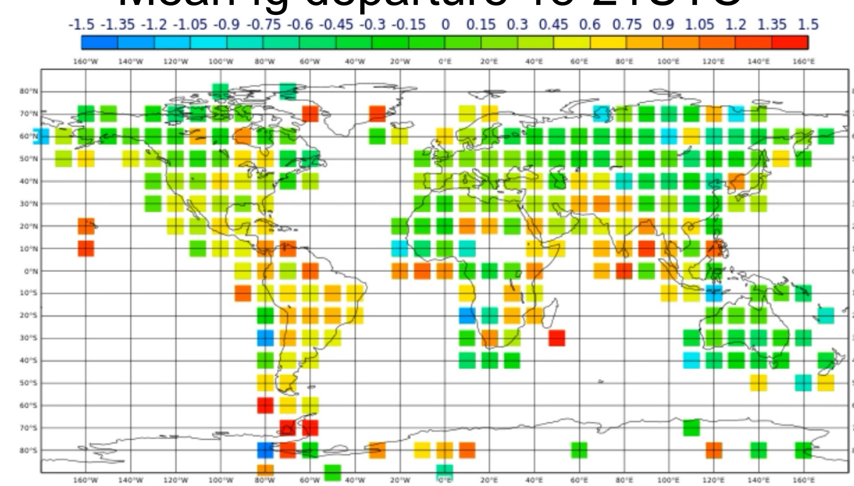
Mean fg departure 06-09UTC



Mean fg departure 12-15UTC

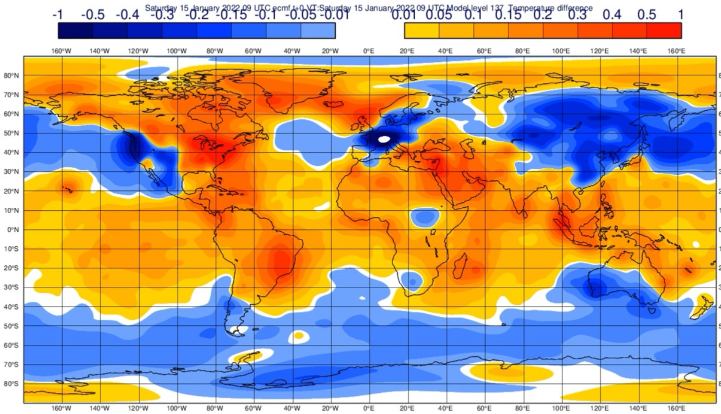


Mean fg departure 18-21UTC

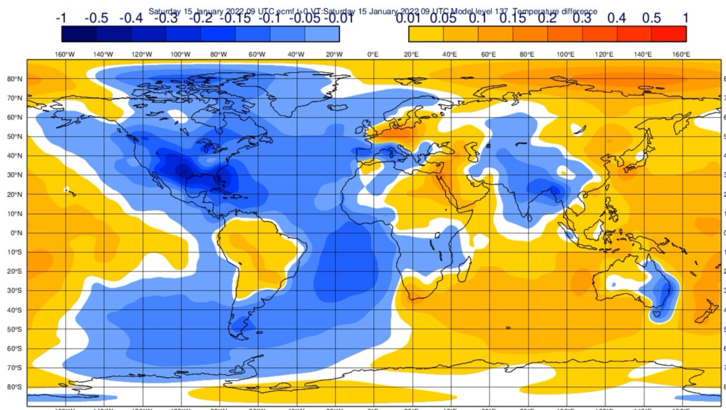


Model biases in the boundary layer

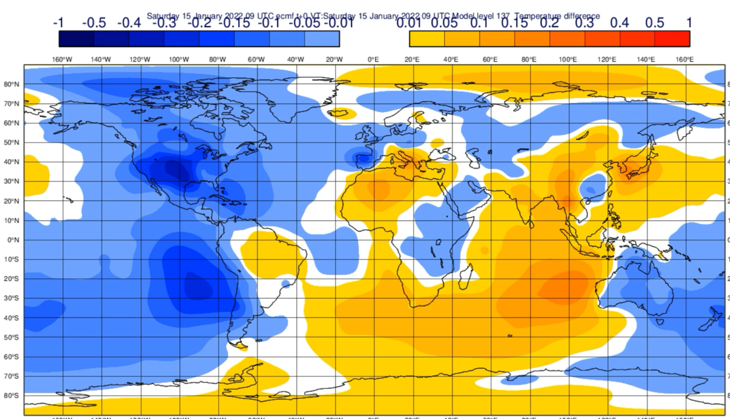
η_0



η_1



η_2



New model for model bias:

$$\eta_0 + \eta_1 \sin(2\pi\frac{t}{24}) + \eta_2 \cos(2\pi\frac{t}{24})$$

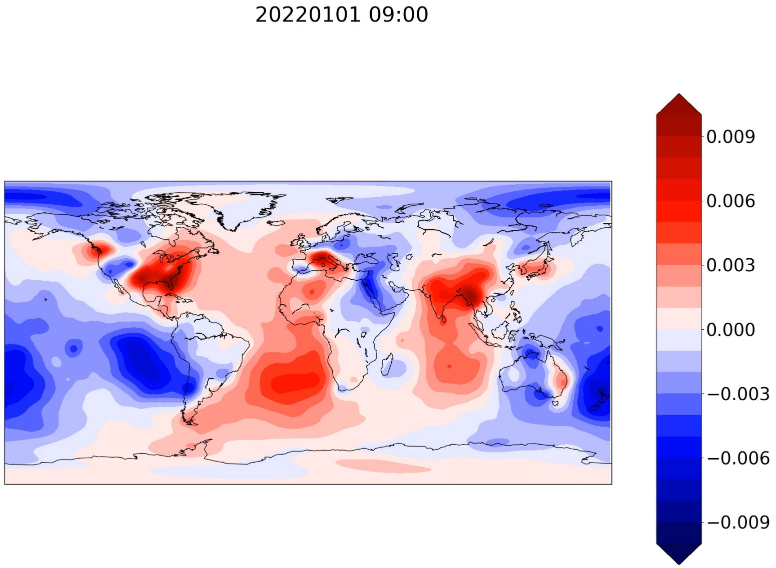
➔ Time-varying within the assimilation window

➔ Designed to capture a diurnal cycle

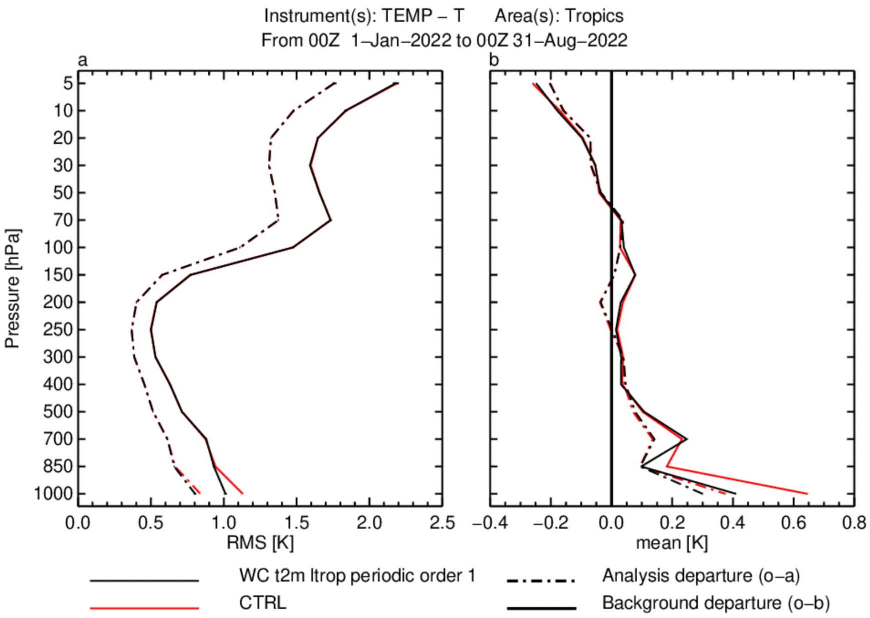
$$\begin{aligned}
J(x_0,\beta,\eta) &= \frac{1}{2}(x_0 - x_b)^T\mathbf{B}^{-1}(x_0 - x_b) \\
&+ \frac{1}{2}\sum_{k=0}^{K}[y_k - \mathcal{H}(x_k) - b(x_k,\beta)]^T\mathbf{R}_k^{-1}[y_k - \mathcal{H}(x_k) - b(x_k,\beta)] \\
&+ \frac{1}{2}(\beta - \beta_b)^T\mathbf{B}_\beta{}^{-1}(\beta - \beta_b) \\
&+ \frac{1}{2}(\eta - \eta_b)^T\mathbf{Q}^{-1}(\eta - \eta_b)
\end{aligned}$$

Model biases in the boundary layer

Model bias correction (level 137)



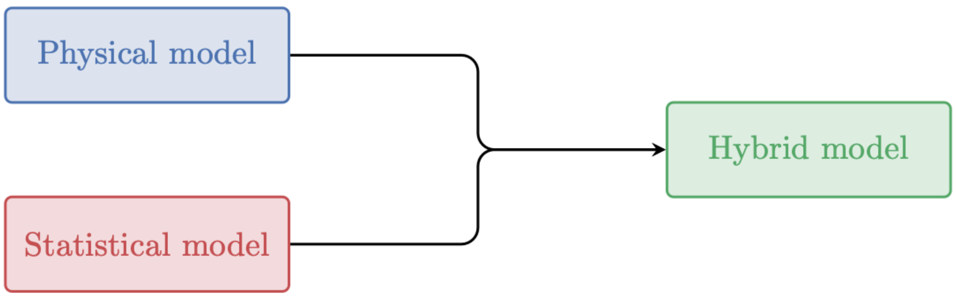
Impact in the mean state against radiosondes



Possible future application of WC4DVar

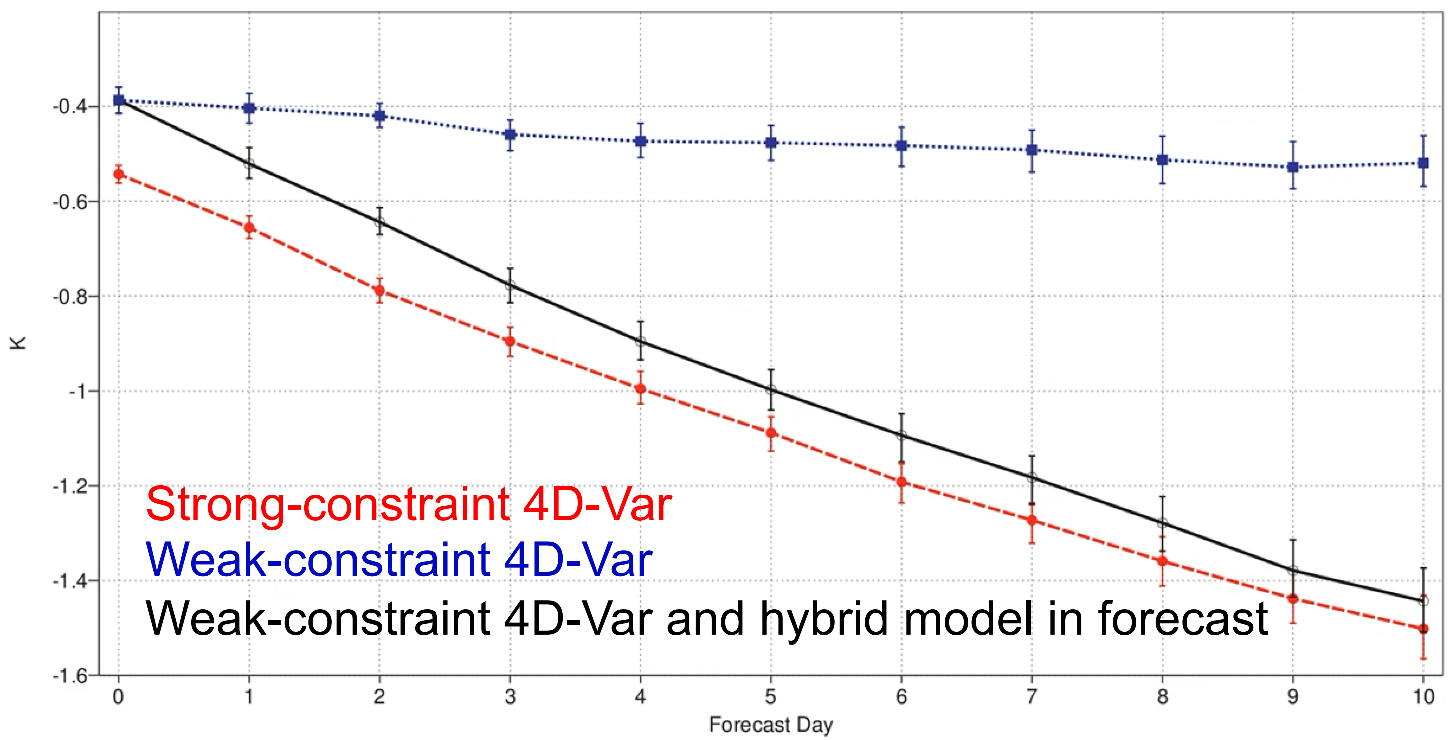
We can see weak-constraint 4D-Var as a tool to build hybrid models

$$x_k = \mathcal{M}_k(x_{k-1}) + \eta \quad \text{for } k = 1, 2, \cdots, K$$

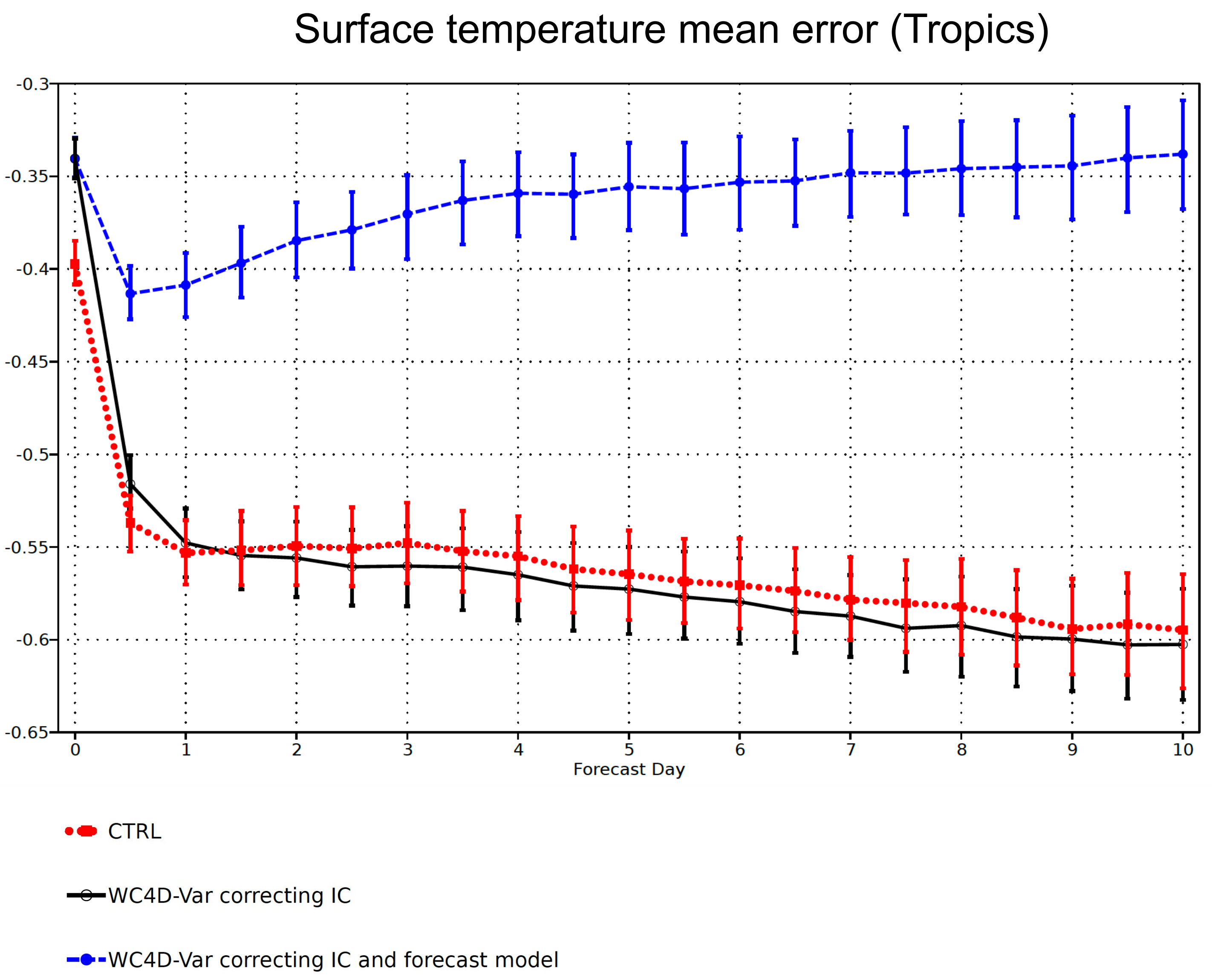


➔ introducing a statistical model that is data-driven to correct for model errors

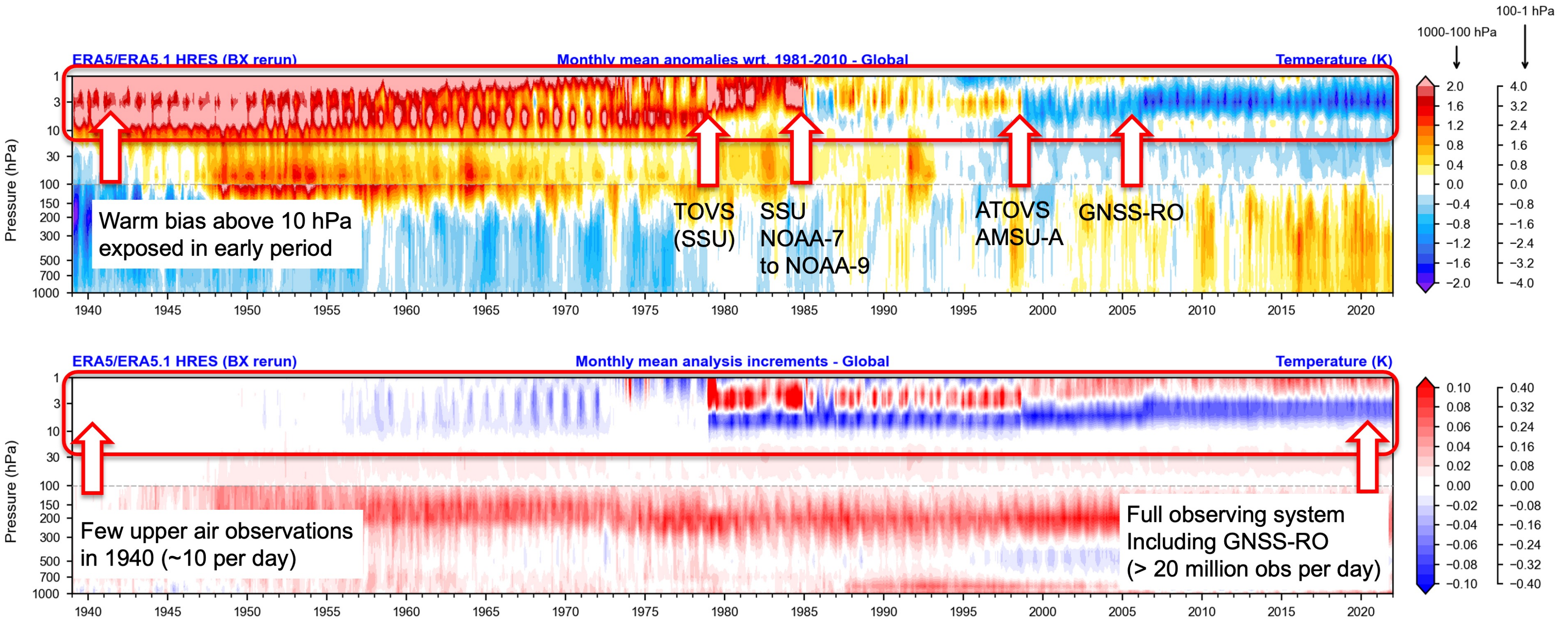
Mean error of the 10-day forecast at 50hPa with respect to the radiosonde observations



Possible future application of WC4DVar



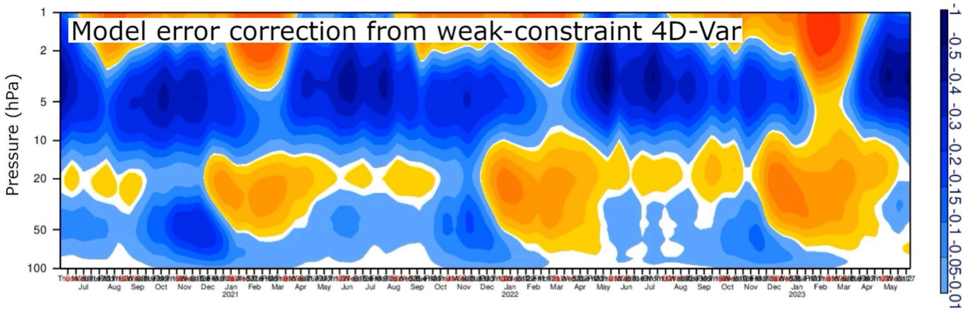
Using hybrid models for reanalysis



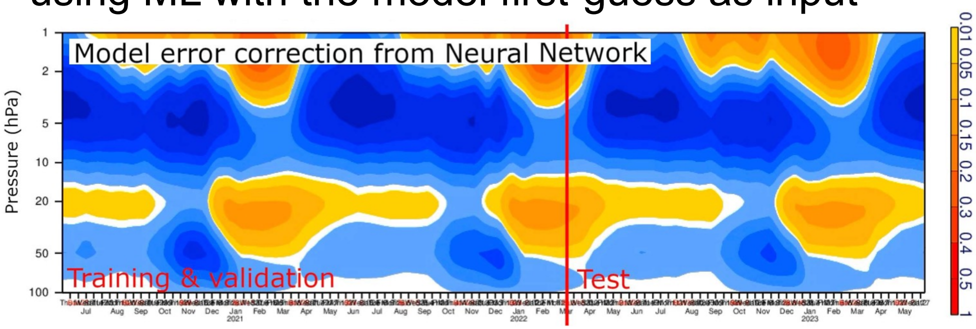
Challenge: reduce artefacts in the stratosphere coming from model biases while preserving climate trends. Amplitude of current spurious signal can be large (>1K)

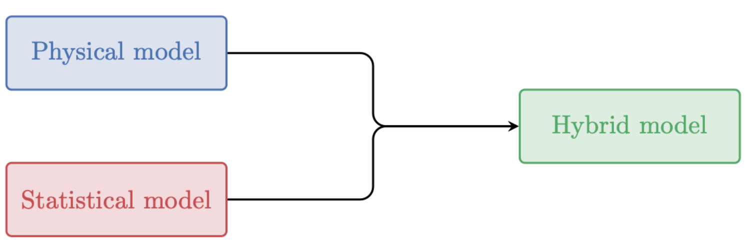
Using hybrid model for reanalysis

1. Weak-constraint 4D-Var estimates model biases effectively over recent periods (2021/2023)

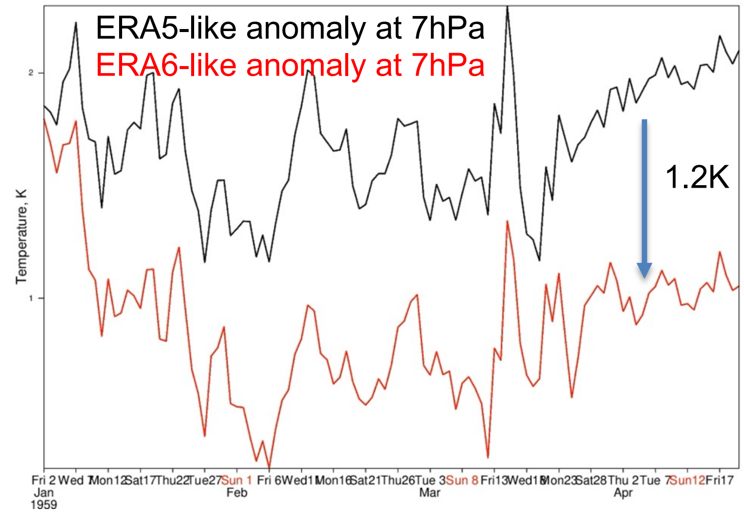


2. This model bias correction is emulated using ML with the model first-guess as input





3. The ML correction can be applied over any reanalysis period (e.g. Jan 1959 to May 1959)



4. Emulator cools down the upper stratosphere to account for the warm bias

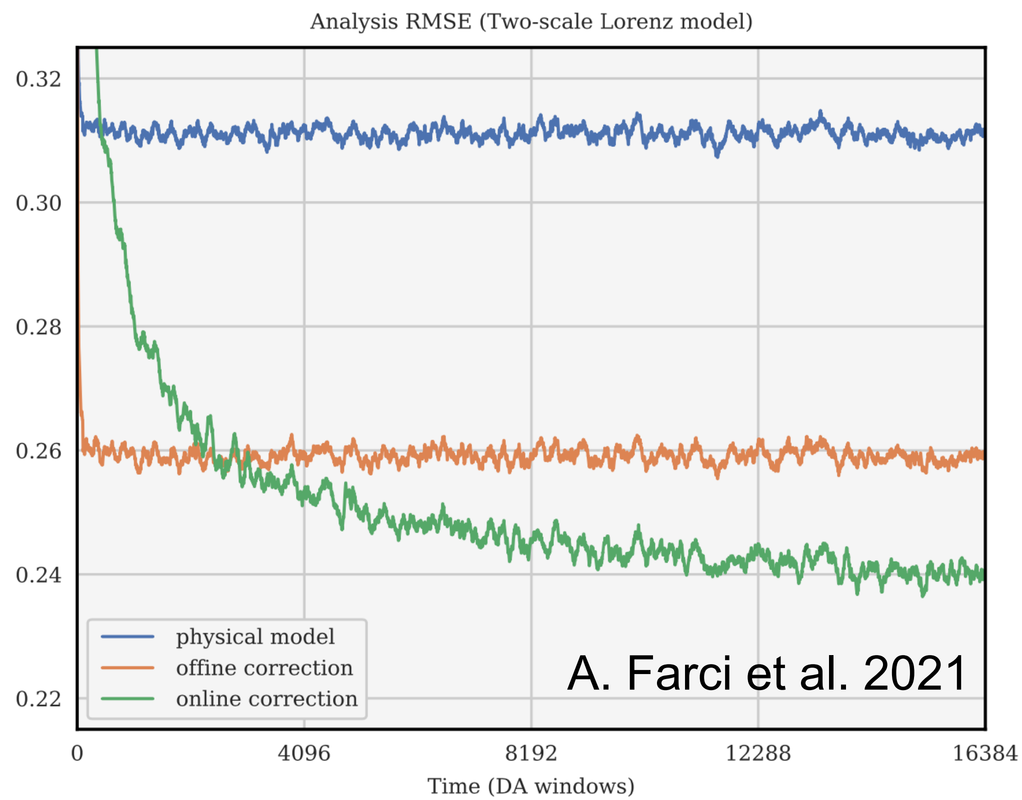
Another possibility for a hybrid model

The hybrid model (physical model + NN correction) is estimated inside 4D-Var

$$\mathbf{x}_{k+1} = \mathcal{M}^{\mathsf{nn}}_{k+1:k}\left(\mathbf{p}, \mathbf{x}_k\right) = \mathcal{M}_{k+1:k}\left(\mathbf{x}_k\right) + \mathcal{F}\left(\mathbf{p}, \mathbf{x}_k\right)$$

NN online loss function

$$\mathcal{J}^{\mathsf{nn}}\left(\mathbf{p}, \mathbf{x}_0\right) = \frac{1}{2}\left\|\mathbf{x}_0 - \mathbf{x}_0^{\mathsf{b}}\right\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\left\|\mathbf{p} - \mathbf{p}^{\mathsf{b}}\right\|^2_{\mathbf{P}^{-1}} + \frac{1}{2}\sum_{k=0}^{L}\left\|\mathbf{y}_k - \mathcal{H}_k \circ \mathcal{M}^{\mathsf{nn}}_{k:0}\left(\mathbf{p}, \mathbf{x}_0\right)\right\|^2_{\mathbf{R}_k^{-1}}$$



➔ learn both model state and NN parameters from observations

➔ TL and ADJ are available

➔ the online correction steadily improves the model, learning from observations

Another possibility for a hybrid model

NN online loss function

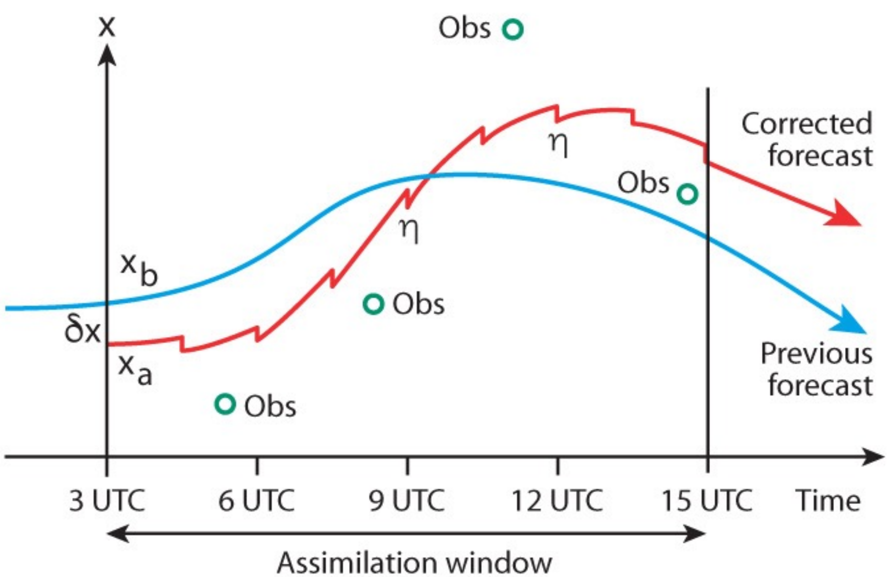
$$\mathcal{J}^{\mathrm{nn}}\left(\mathbf{p},\mathbf{x}_0\right)=\frac{1}{2}\left\|\mathbf{x}_0-\mathbf{x}_0^{\mathrm{b}}\right\|^2_{\mathbf{B}^{-1}}+\frac{1}{2}\left\|\mathbf{p}-\mathbf{p}^{\mathrm{b}}\right\|^2_{\mathbf{P}^{-1}}+\frac{1}{2}\sum_{k=0}^{L}\left\|\mathbf{y}_k-\mathcal{H}_k\circ\mathcal{M}^{\mathrm{nn}}_{k:0}\left(\mathbf{p},\mathbf{x}_0\right)\right\|^2_{\mathbf{R}_k^{-1}}$$

In weak-constraint 4D-Var, an error term is introduced in the model equation

$$\mathbf{x}_k=\mathcal{M}_{k+1:k}\left(\mathbf{x}_k\right)+\mathbf{w}$$

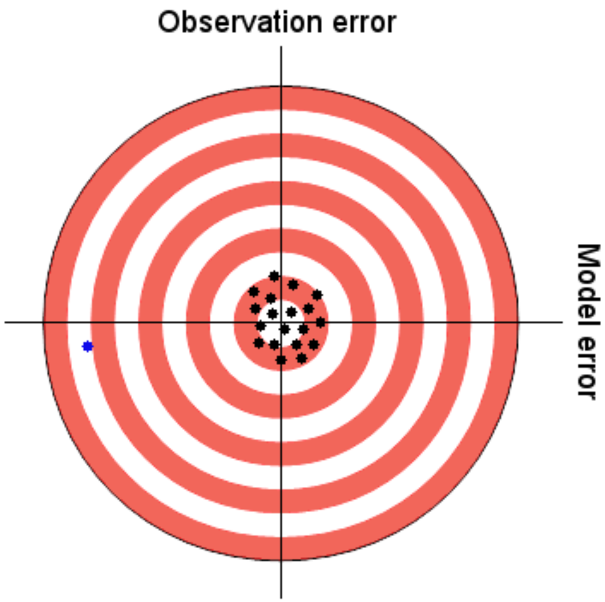
WC4D-Var cost function

$$\mathcal{J}^{\mathrm{wc}}\left(\mathbf{w},\mathbf{x}_0\right)=\frac{1}{2}\left\|\mathbf{x}_0-\mathbf{x}_0^{\mathrm{b}}\right\|^2_{\mathbf{B}^{-1}}+\frac{1}{2}\left\|\mathbf{w}-\mathbf{w}^{\mathrm{b}}\right\|^2_{\mathbf{Q}^{-1}}+\frac{1}{2}\sum_{k=0}^{L}\left\|\mathbf{y}_k-\mathcal{H}_k\circ\mathcal{M}^{\mathrm{wc}}_{k:0}\left(\mathbf{w},\mathbf{x}_0\right)\right\|^2_{\mathbf{R}_k^{-1}}$$

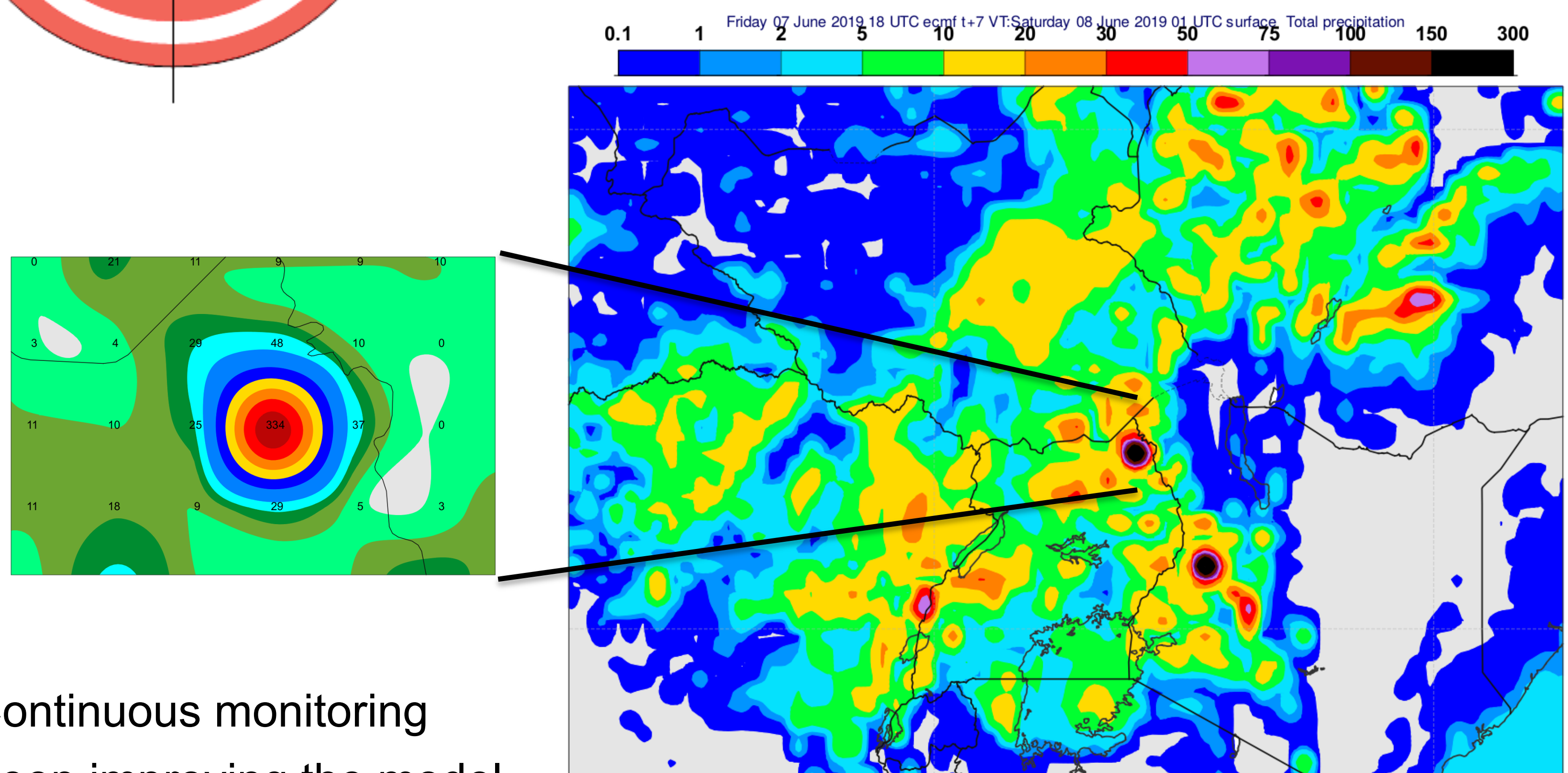


Talk on Friday
by Alan Geer

Not the job of weak-constraint 4D-Var: Model gross errors



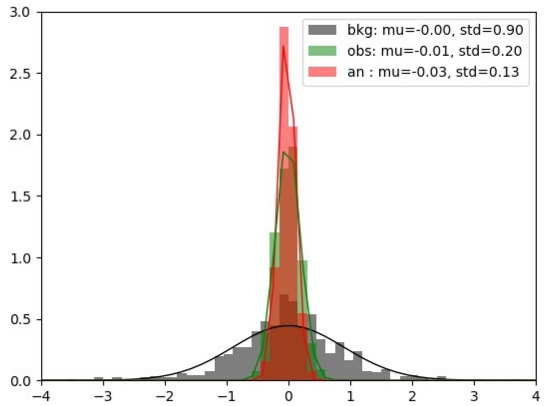
Total precipitation on 07 June 2019
(accumulated over 6 hours)



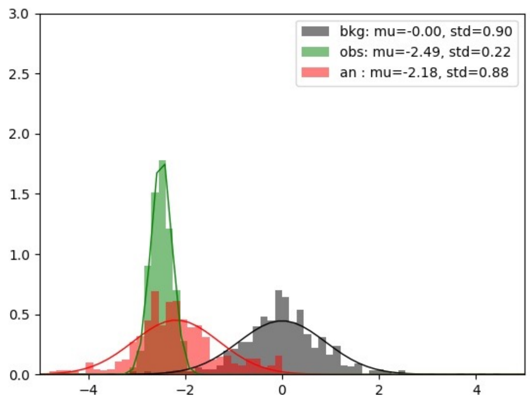
→ Continuous monitoring

→ Keep improving the model

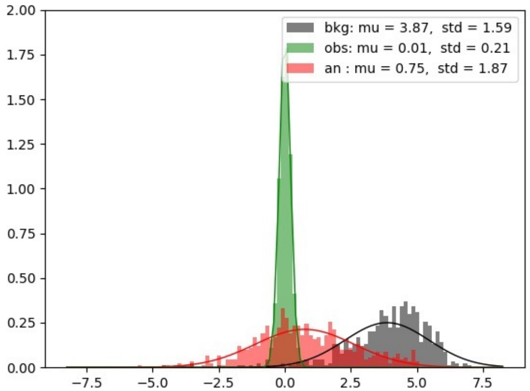
Summary 1/3



Background: unbiased (only random errors)
Observation: unbiased (only random errors)
Standard 4D-Var



Background: unbiased (only random errors)
Observation: biased
Standard 4D-Var & Variational Bias Control (VarBC)



Background: biased
Observation: unbiased (only random errors)
Weak constraint 4D-Var

Summary 2/3

How do I know if my observations are biased?
How do I know if my model is biased?
You don't know the truth, but you have to trust something

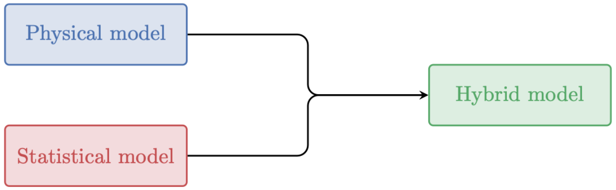
Reference observations are used

Radiosondes

GPS-RO

Summary 3/3

From bias-blind to bias-aware data assimilation



$$
\begin{aligned}
J(x_0, \beta, \eta) &= \frac{1}{2}(x_0 - x_b)^T \mathbf{B}^{-1}(x_0 - x_b) \\
&+ \frac{1}{2} \sum_{k=0}^{\text{Radiosonde}} [y_k - \mathcal{H}(x_k)]^T \mathbf{R}_k^{-1} [y_k - \mathcal{H}(x_k)] \\
&+ \frac{1}{2} \sum_{k=0}^{\text{GPSRO}} [y_k - \mathcal{H}(x_k)]^T \mathbf{R}_k^{-1} [y_k - \mathcal{H}(x_k)] \\
&+ \frac{1}{2} \sum_{k=0}^{\text{Others}} [y_k - \mathcal{H}(x_k) - b(x_k, \beta)]^T \mathbf{R}_k^{-1} [y_k - \mathcal{H}(x_k) - b(x_k, \beta)] \\
&+ \frac{1}{2}(\beta - \beta_b)^T \mathbf{B}_\beta{}^{-1}(\beta - \beta_b) \\
&+ \frac{1}{2}(\eta - \eta_b)^T \mathbf{Q}^{-1}(\eta - \eta_b)
\end{aligned}
$$

Any questions? Feel free to contact me patrick.laloyaux@ecmwf.int