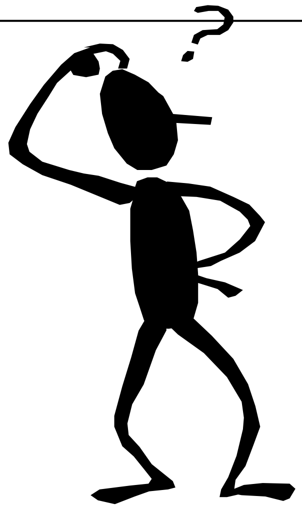
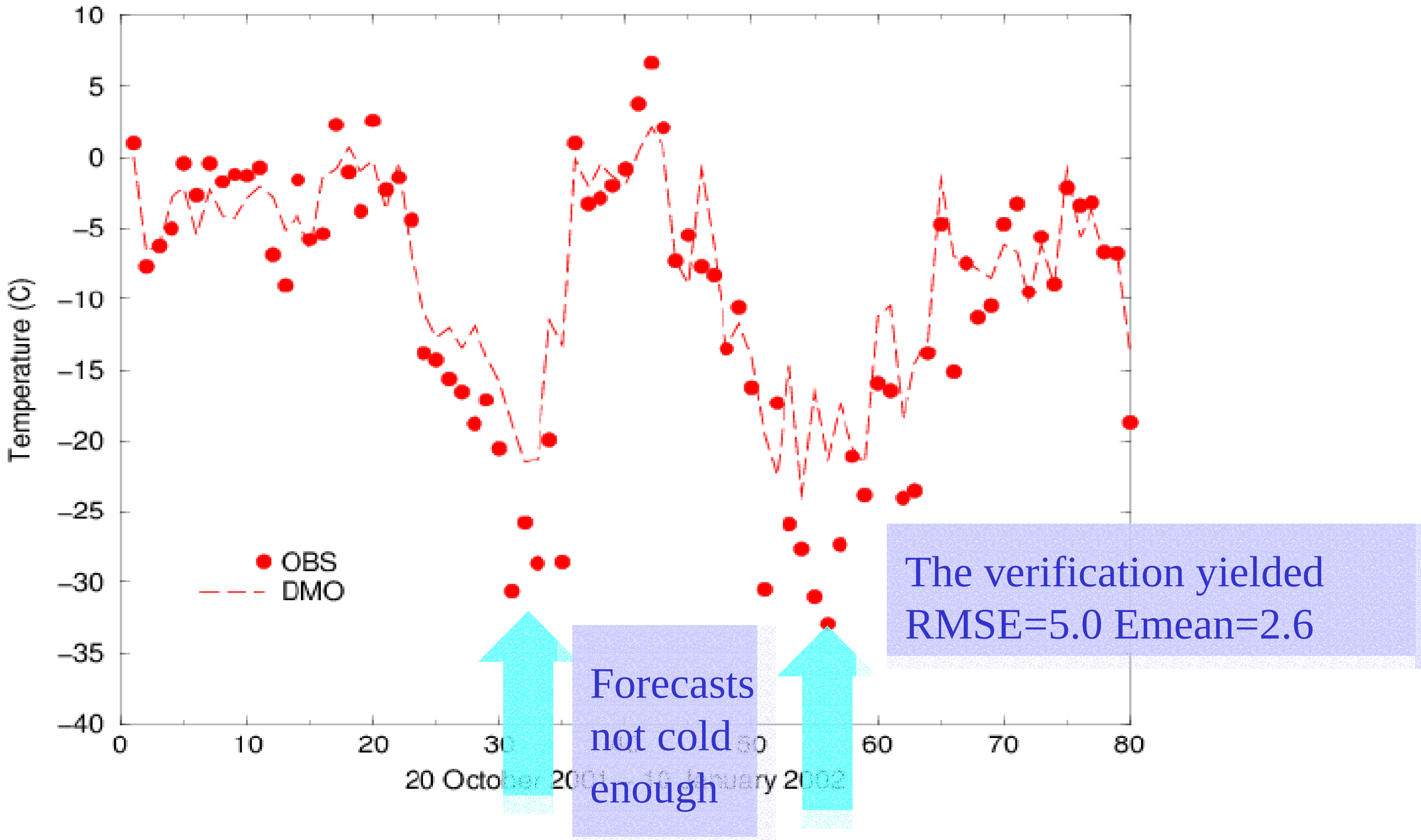
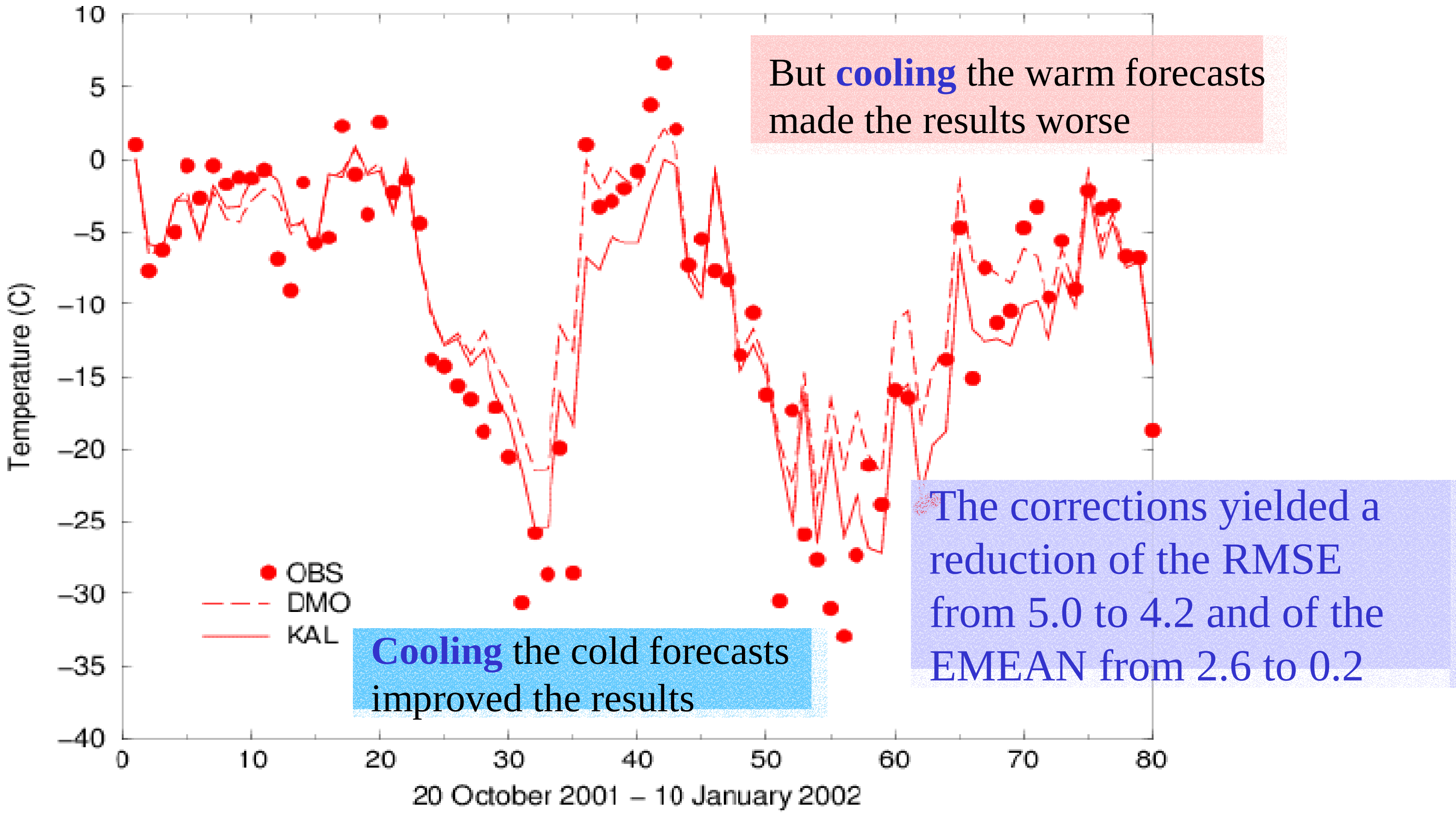
How can an improved model yield higher RMS errors?



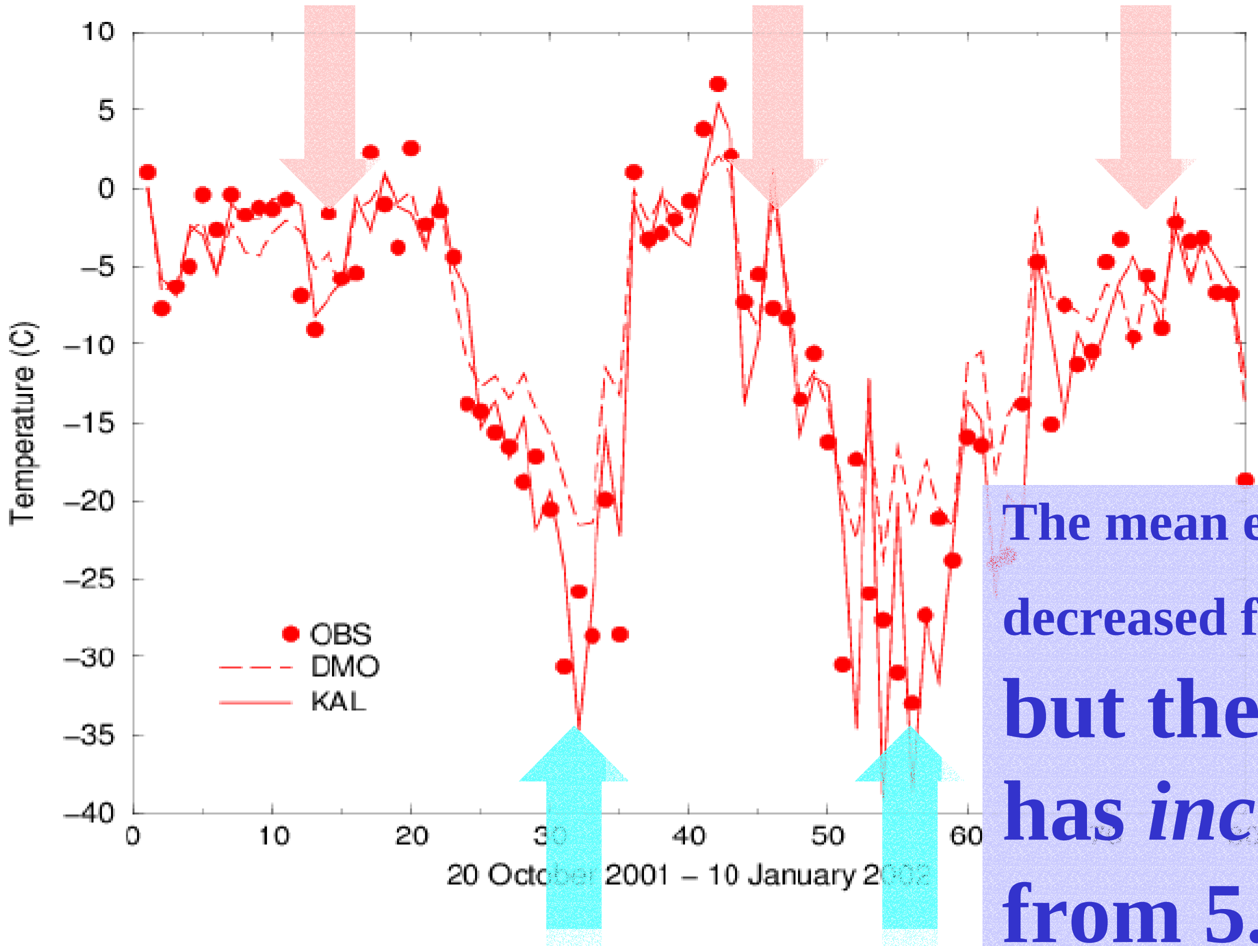
HIRLAM-44 24 hour 2 m temperature forecast for Kiruna in Lapland winter 2001-2002



A 1-dimensional Kalman filter can reduce an overall bias



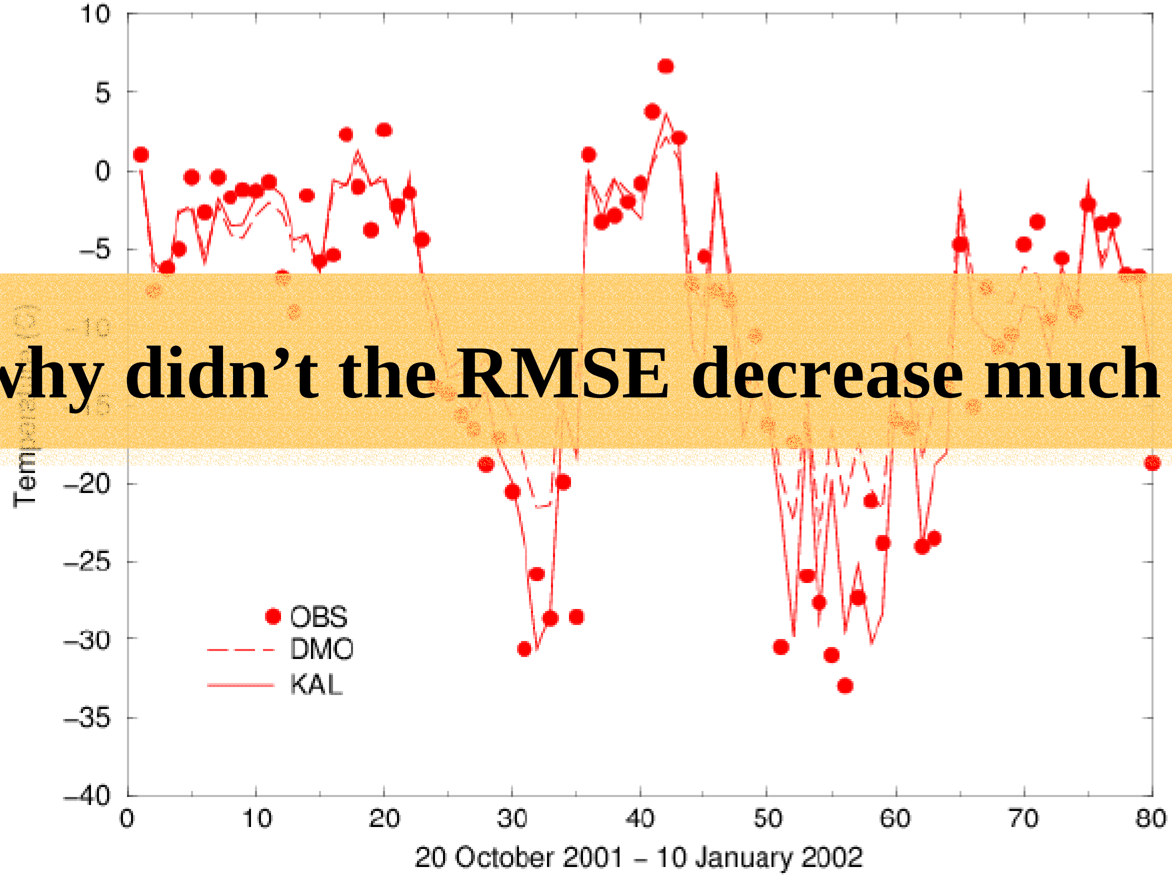
A 2-dimensional Kalman filter can provide different corrections to different regimes



The mean error has decreased from 2.6 to 0.0

but the RMSE has *increased* from 5.0 to 5.1

After proper parameter setting the RMSE reduced from 5.0 to 4.6 while the mean error stayed at 0.3, a reduction by 2.3



But why didn't the RMSE decrease much more?

The full mathematical expression for the RMS error (E_j) of a **j**-day forecast issued on day **i** verified over **N** gridpoints over a period of **T** days

$$E_j = \sqrt{\frac{1}{T}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}(f_{i,j} - a_{i+j})^2}$$

We make things easier for us by considering the ***square*** of the RMSE

$$E_j^2 = \frac{1}{T}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}(f_{i,j} - a_{i+j})^2$$

$$E_j^2 = \frac{1}{T}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}(f_{i,j} - a_{i+j})^2$$

The notation is further simplified by replacing the Σs with an overbar symbolising all temporal and spatial averages. We also skip all the indices.

$$E^2 = \overline{(f-a)^2}$$

The equation looks trivial,

$$E^2 = \overline{(f - a)^2}$$

but reveals its deeper implication when considered in connection with the apparently equally “trivial”

$$(a + b)^2 = a^2 + b^2 + 2ab$$

Decomposing the RMSE

$$E^2 = \overline{(f-a)^2}$$

Introduce **c** as the climate value of the verifying day

$$E^2 = \overline{(f-c+c-a)^2}$$

Reposition **c** to form **f-c** and **a-c**

$$E^2 = \overline{((f-c)-(a-c))^2}$$

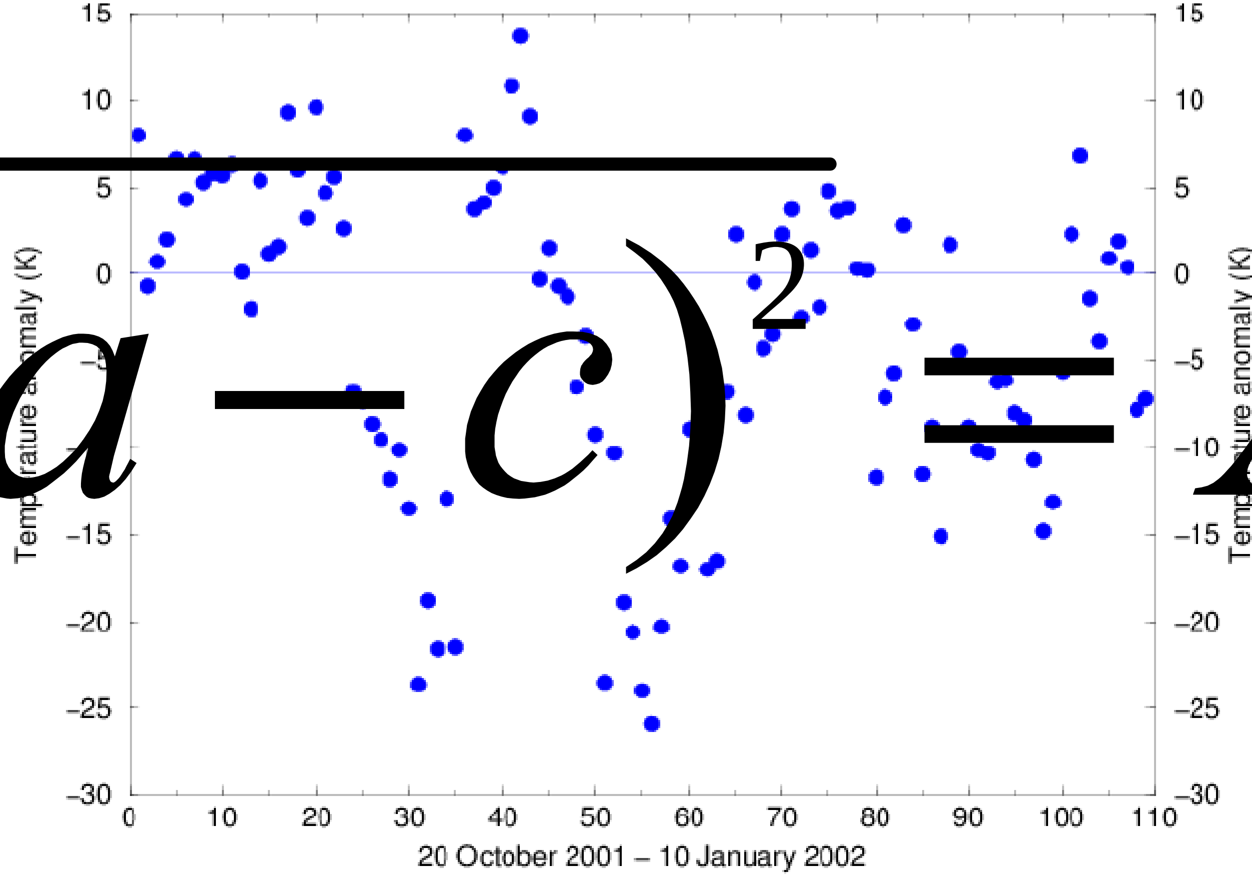
Apply $\mathbf{(a+b)^2=a^2+b^2+2ab}$

$$E^2 = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)}$$

Each of these three terms has its own story to tell

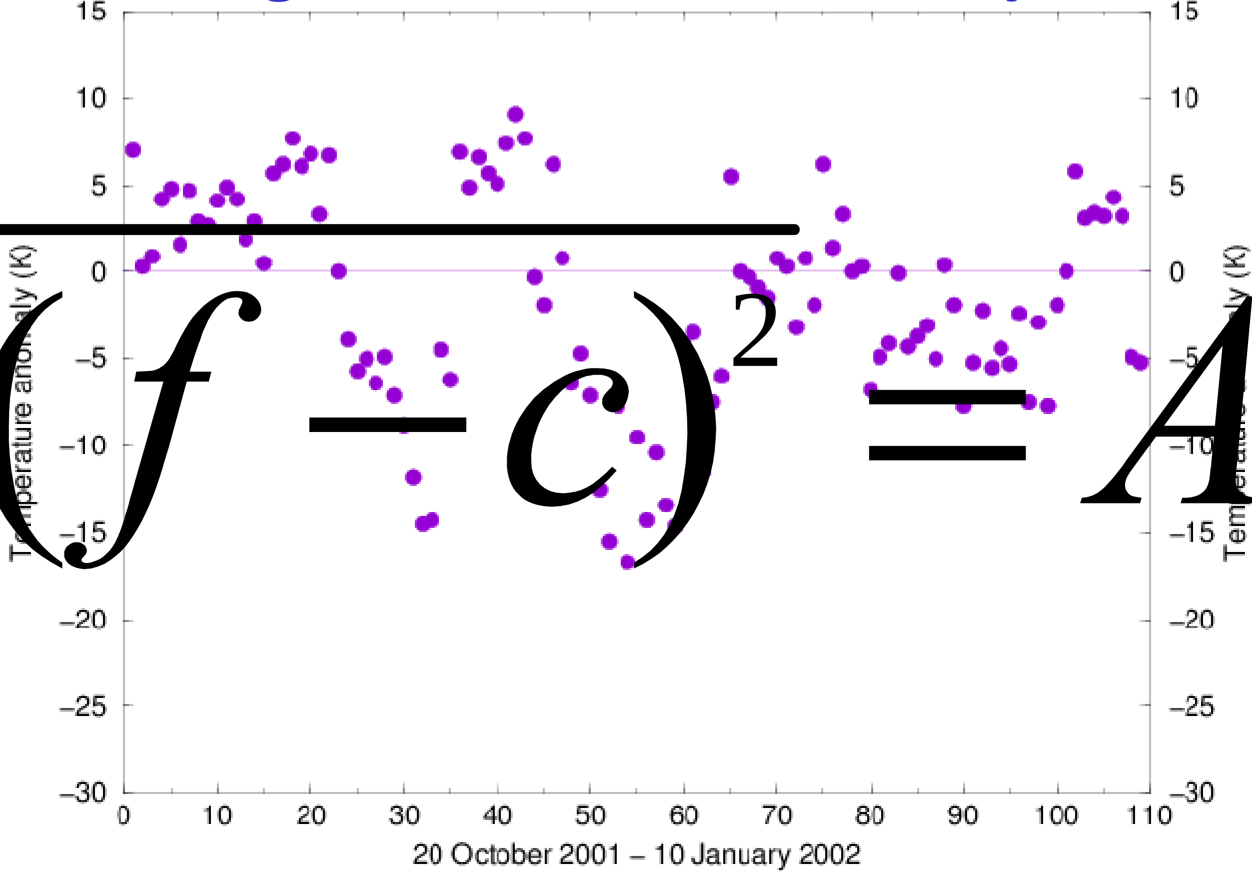
The *observed* variability around the climatological mean is expressed by the term

$$\overline{(a-c)^2} = A_a^2$$



The magnitude of this term can not be affected by human intervention

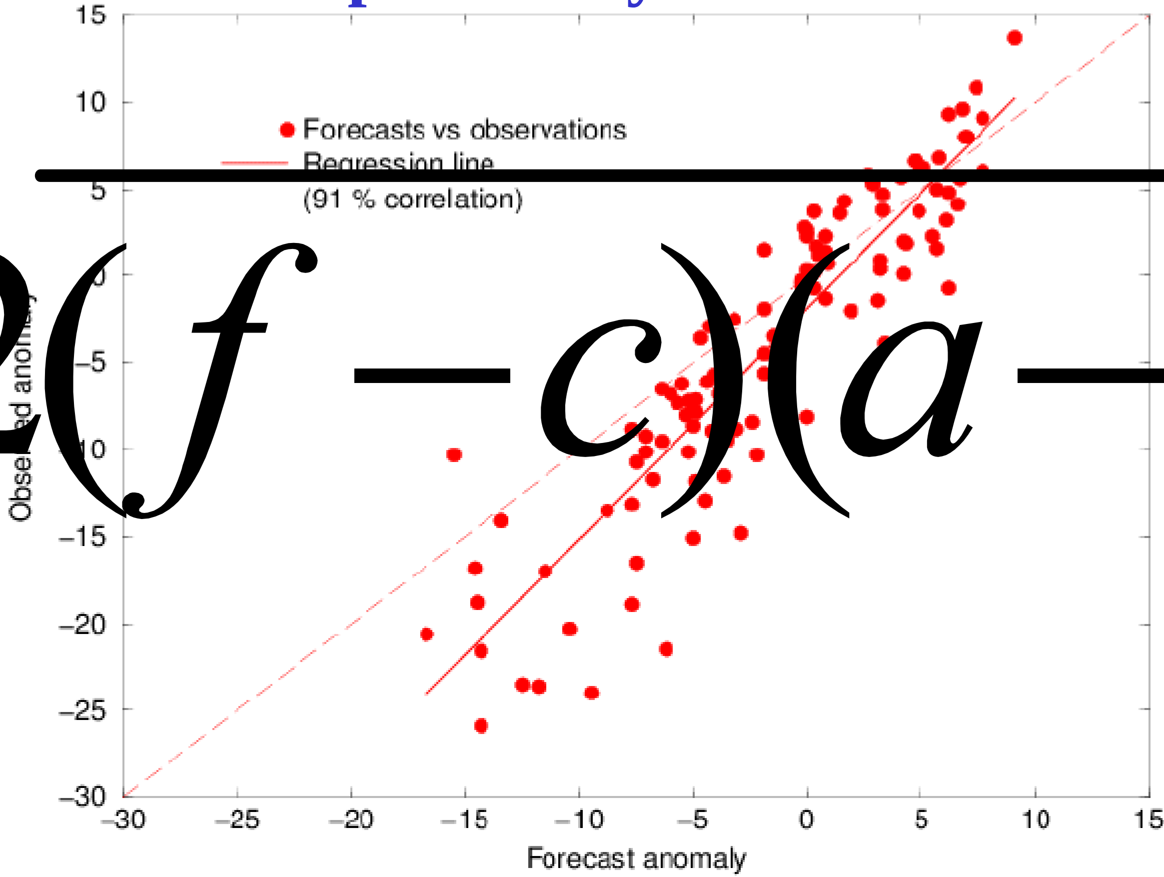
The *forecast* variability around the climatological mean is measured by the term



$$\overline{(f-c)^2} = A_f^2$$

The magnitude of this term can indeed be affected by human intervention

The correspondence between f-c and a-c is expressed by the term



$$\overline{2(f-c)(a-c)}$$

This is the *only* term in the RMSE decomposition which is related to the predictive skill of the model

The *improvement* of the model, as a simulation of the atmospheric system, may therefore appear as *deterioration* of the quality of the model!

An increase in forecast variability increases $\mathbf{A_f}^2$

...compensates the decrease of the RMSE due to improved forecasts

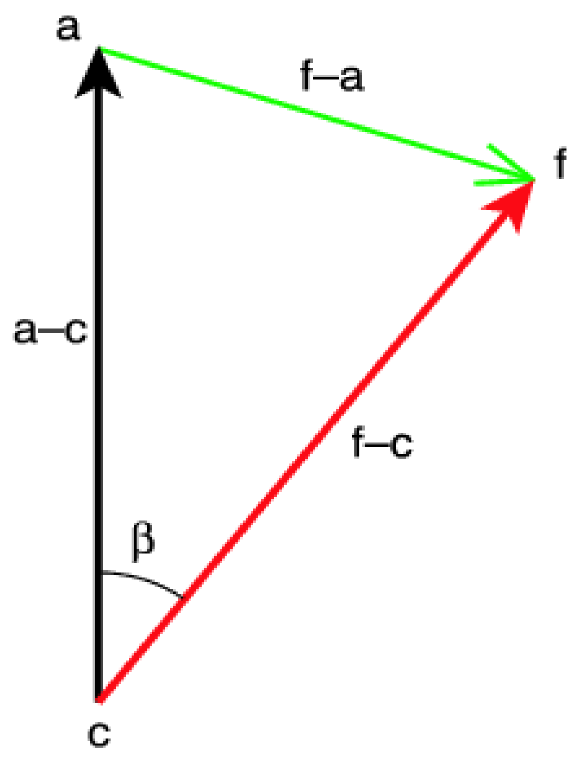
$$E^2 = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)}$$

...to the level of the observed variability $\mathbf{A_a}^2$

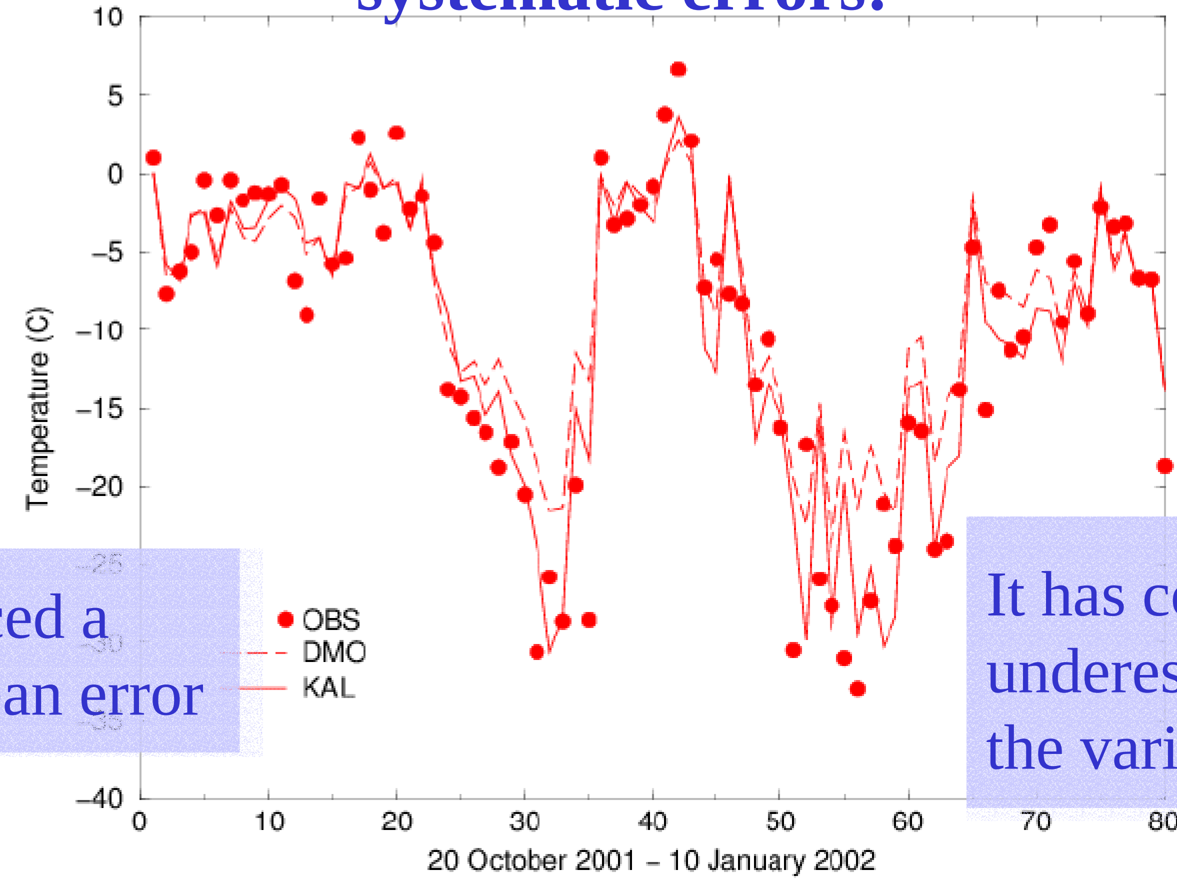
The previous mathematics can also be given a vector algebraic presentation where **a**, **f** and **c** represent states in some phase space

The length of the vectors represent A_a and A_f, and the difference **f-a** is proportional to the RMSE

With an underactive model **f-c** will become somewhat shorter. Also **f-a** will decrease and thus the RMSE



A close inspection of the Kalman filtered forecasts show that the filtering has reduced two systematic errors:

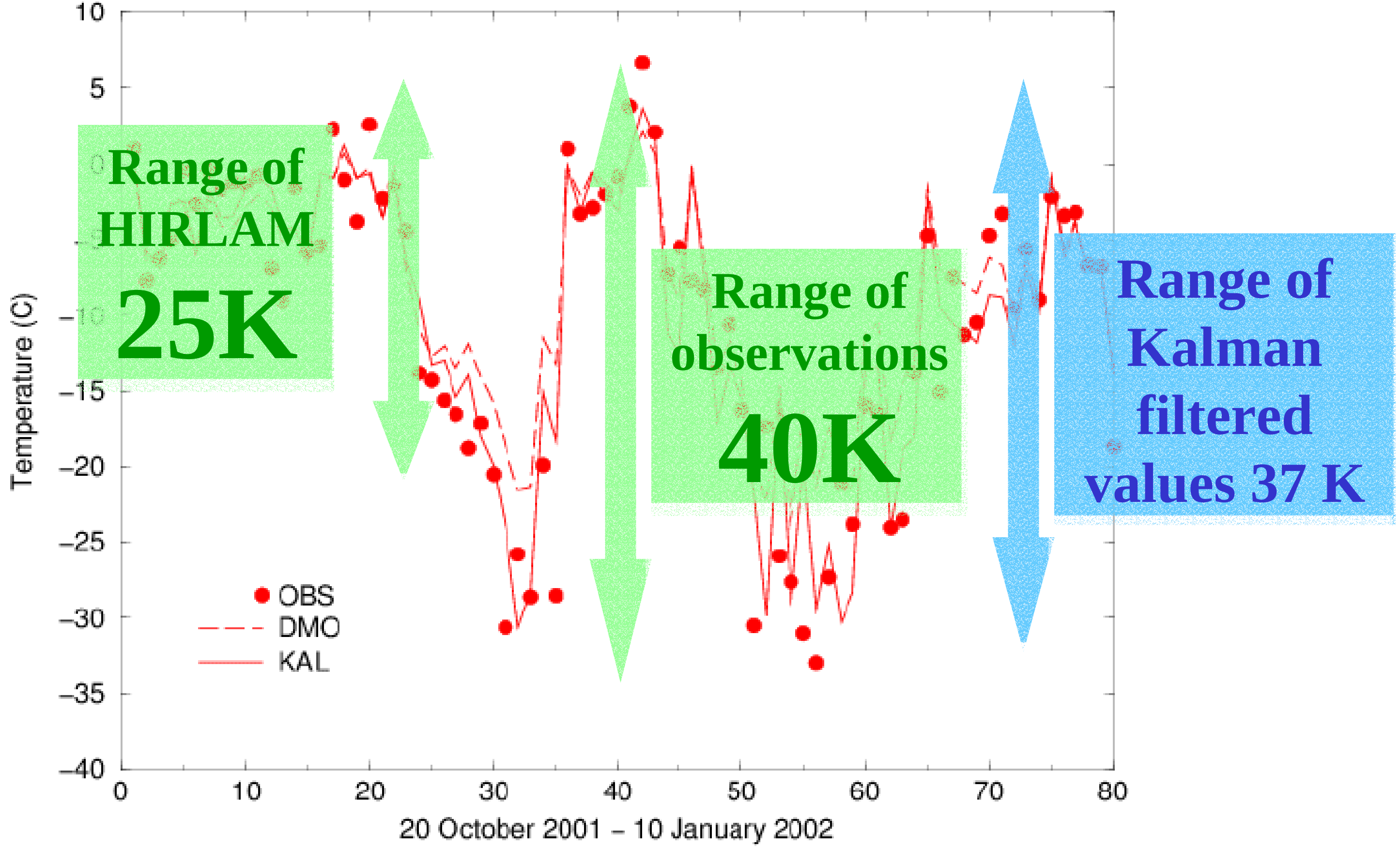


It has reduced a positive mean error

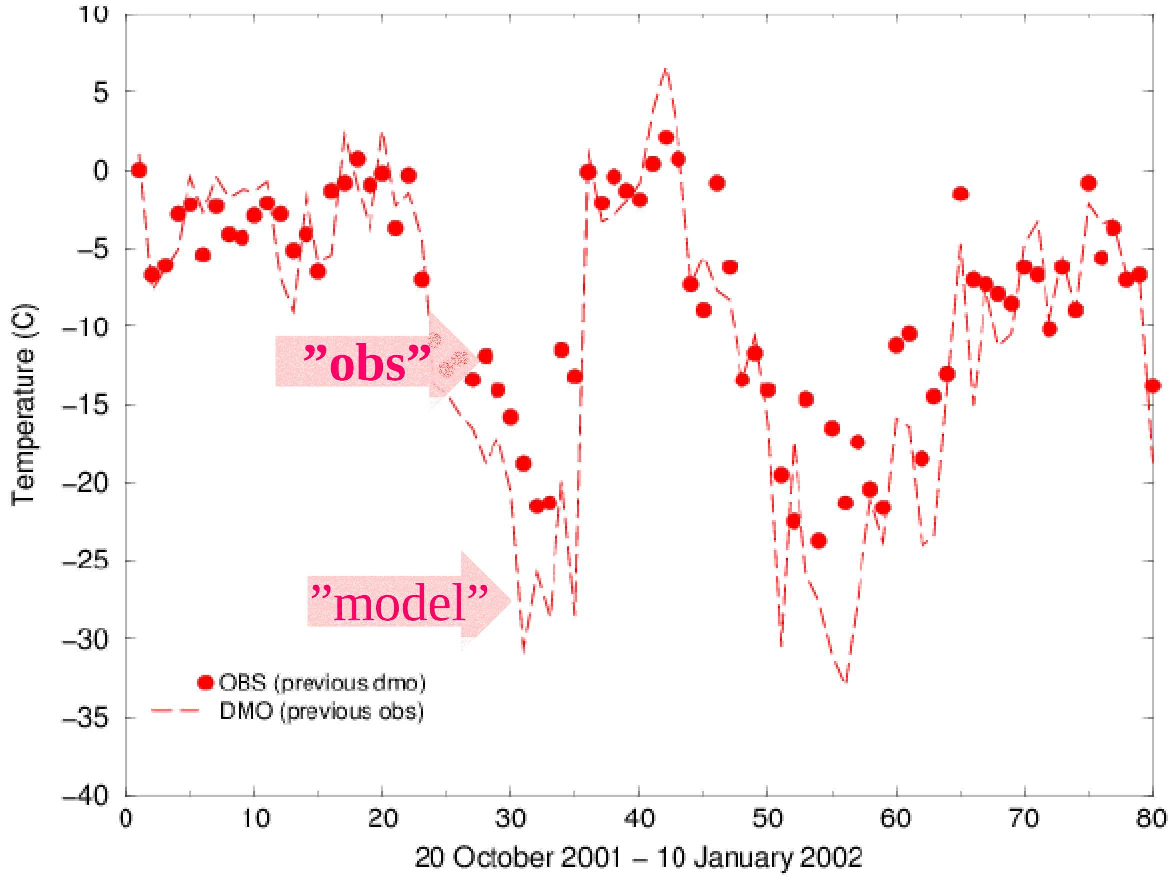
It has corrected an underestimation of the variability

2nd SRNWP Workshop on Statistical and Dynamical Adaptation Wien 5-6 May 2003 Anders Persson SMHI

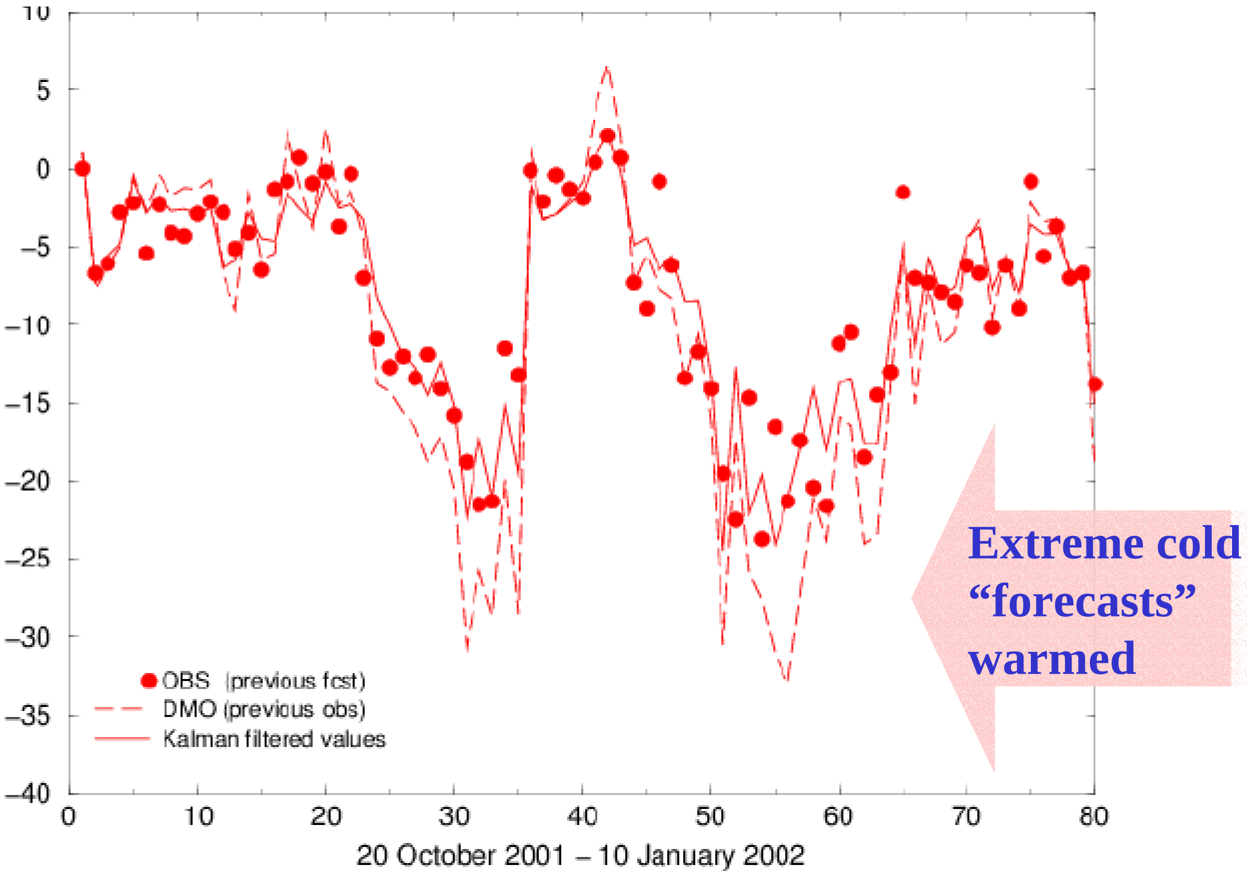
The range of the Kalman filtered values is about the same as the observed



24 hour 2 m temperature forecast for Kiruna in Lapland winter 2001-2002 - with observations and forecasts swopped



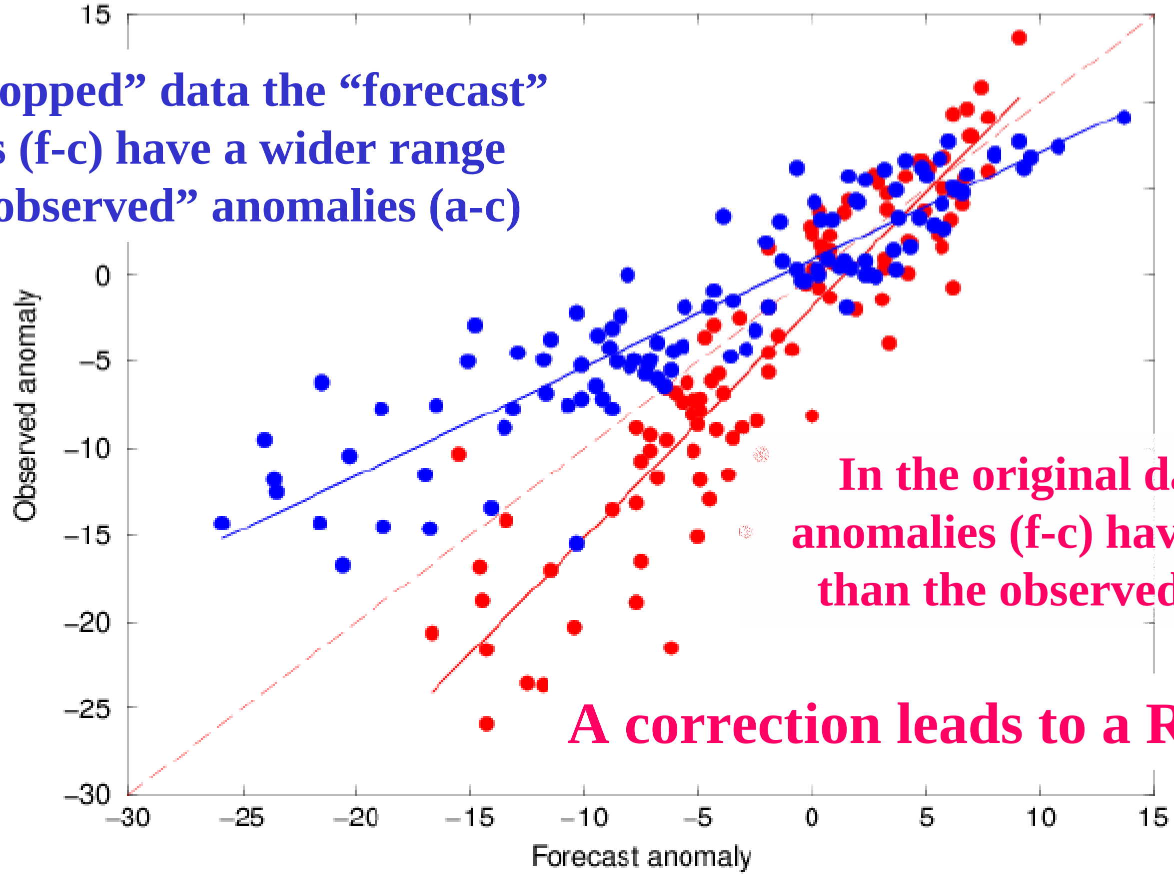
After Kalman filtering the EMEAN is reduced to zero and the RMSE is reduced from 5.0 to 2.9



2 m temperature observed anomalies versus forecast anomalies

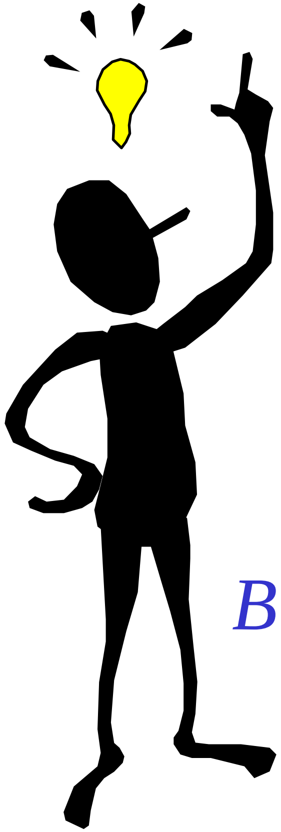
For the ”swopped” data the “forecast” anomalies (f-c) have a wider range than the “observed” anomalies (a-c)

A correction leads to a RMSE *reduction*



In the original data the forecast anomalies (f-c) have a smaller range than the observed anomalies (a-c)

A correction leads to a RMSE *increase*



Depending on if our model is over or underestimating the variability of the atmospheric motions, an improvement may yield decreased or increased values of the RMSE

But there is more to say about the RMSE equation

When **f=c** the first and last terms disappear

$$E^2 = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)}$$

$$E^2 \quad \rightarrow \quad \overline{(a-c)^2}$$

$$E^2 \quad \rightarrow \quad A_a^2$$

We take the square root….

$$E \rightarrow A_a$$

Which is the error level for a purely climatological statement

When the forecasts start to lose skill and the RMSE start to approach high error levels the last term disappears

$$E^2 = \overline{(f-c)^2} + \overline{(a-c)^2} - 2\overline{(f-c)(a-c)}$$

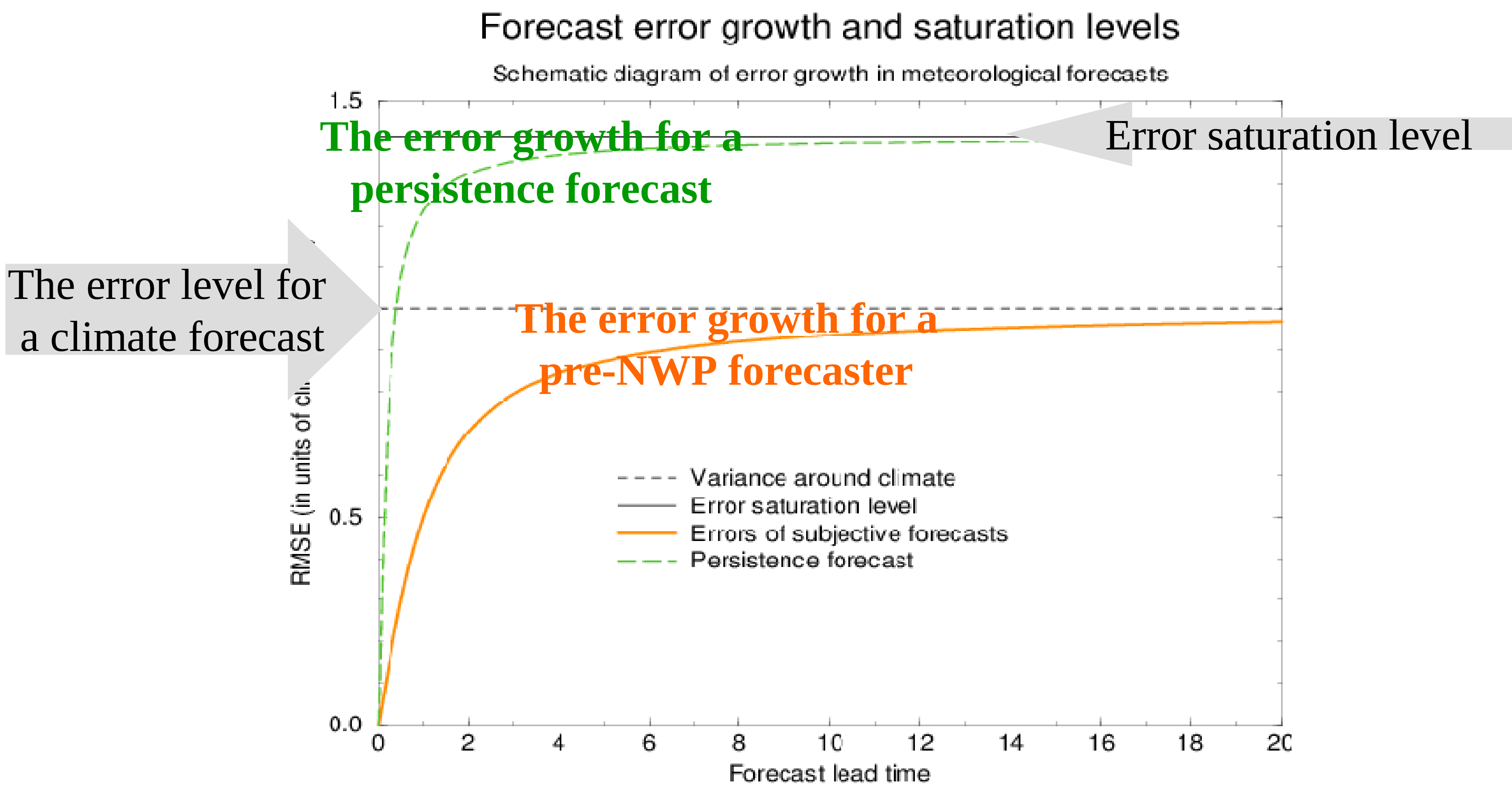
$$E^2 \rightarrow A_f^2 + A_a^2$$

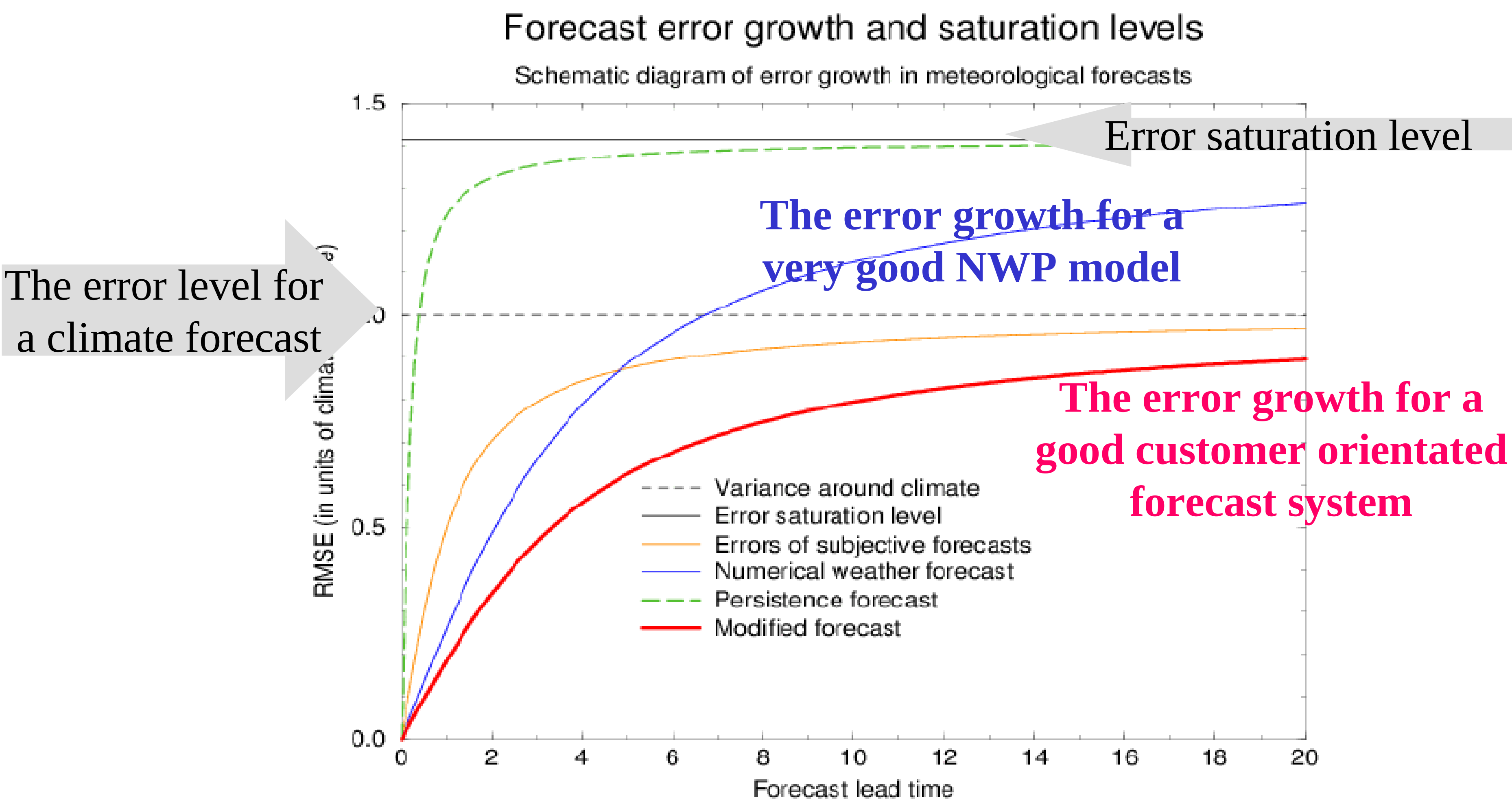
$$E^2 \rightarrow 2A_a^2$$

$$E \rightarrow A_a\sqrt{2}$$

Which is the uppermost error level for a realistic NWP model, 41% above the error level of a purely climatological statement

It is also called *The Error Saturation Level (ESL)*


Forecast error growth and saturation levels
Schematic diagram of error growth in meteorological forecasts
The error growth for a persistence forecast
Error saturation level
The error level for a climate forecast
The error growth for a pre-NWP forecaster
Variance around climate
Error saturation level
Errors of subjective forecasts
Persistence forecast
1.5
0.5
0.0
0
2
4
6
8
10
12
14
16
18
20
Forecast lead time


Forecast error growth and saturation levels
Schematic diagram of error growth in meteorological forecasts
1.5
0.5
0.0
0
2
4
6
8
10
12
14
16
18
20
Forecast lead time
Variance around climate
Error saturation level
Errors of subjective forecasts
Numerical weather forecast
Persistence forecast
Modified forecast
Error saturation level
The error growth for a very good NWP model
The error level for a climate forecast
The error growth for a good customer orientated forecast system

Conflict of interest:

Model developers want a physically realistic model with a natural range of variability reflecting all scales the numerical resolution can resolve and describe. The RMSE will approach the ESL.

Users of the models want to eliminate features which may be realistic as such, but lack predictability beyond a certain forecast range. The RMSE will approach a level much below the ESL.

Should we abandon the RMSE and replace it by a new #@*-score?

- RMSE is well established in all sciences and relates well to general mathematics
- A new score will have difficulties to become accepted (-”You have changed the goal posts”)
- A new score might have its own artefacts and open up for other, more subtle misinterpretations
- RMSE should be used in tandem with other statistical measures like correlation, mean errors etc.