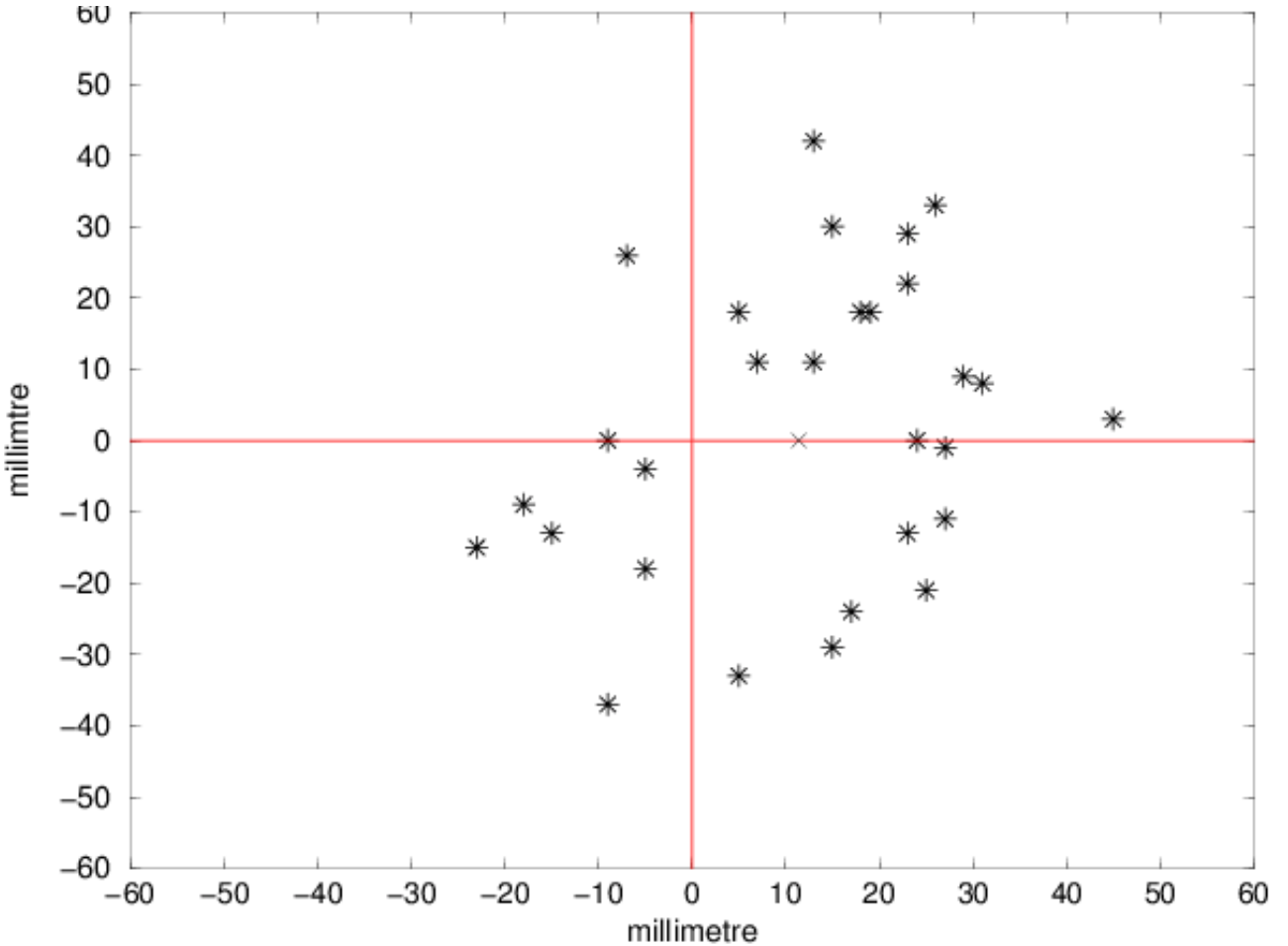
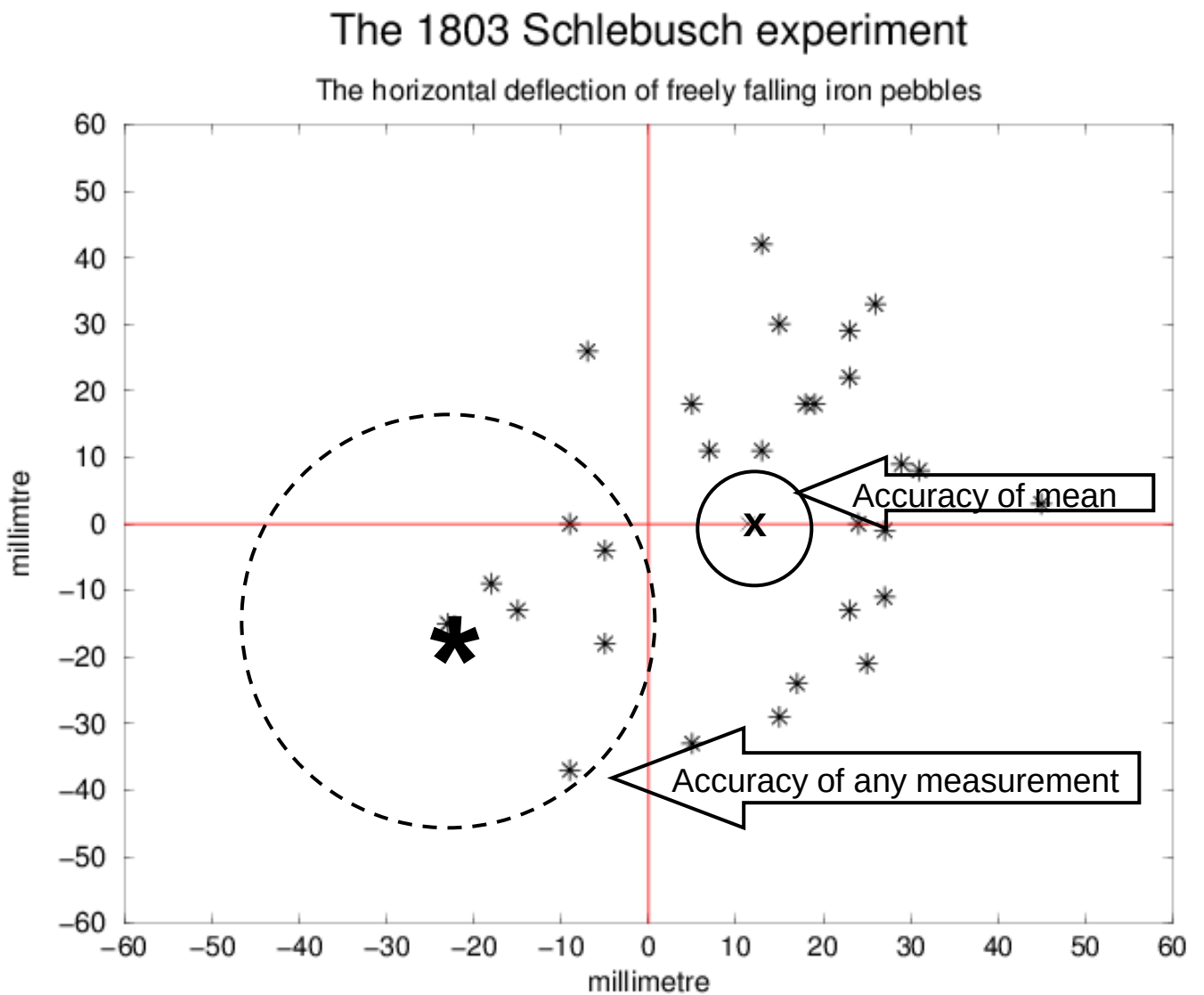
# Forecast accuracy, “jumpiness” and the EPS

# It was the 1803 experiment in the Schlebusch mines that for the first time confirmed that the earth was rotating




In 1803 a group of European scientists tried to estimate the deflection of a vertically falling object by dropping iron pebbles in a mine shaft in Saxony, Germany.

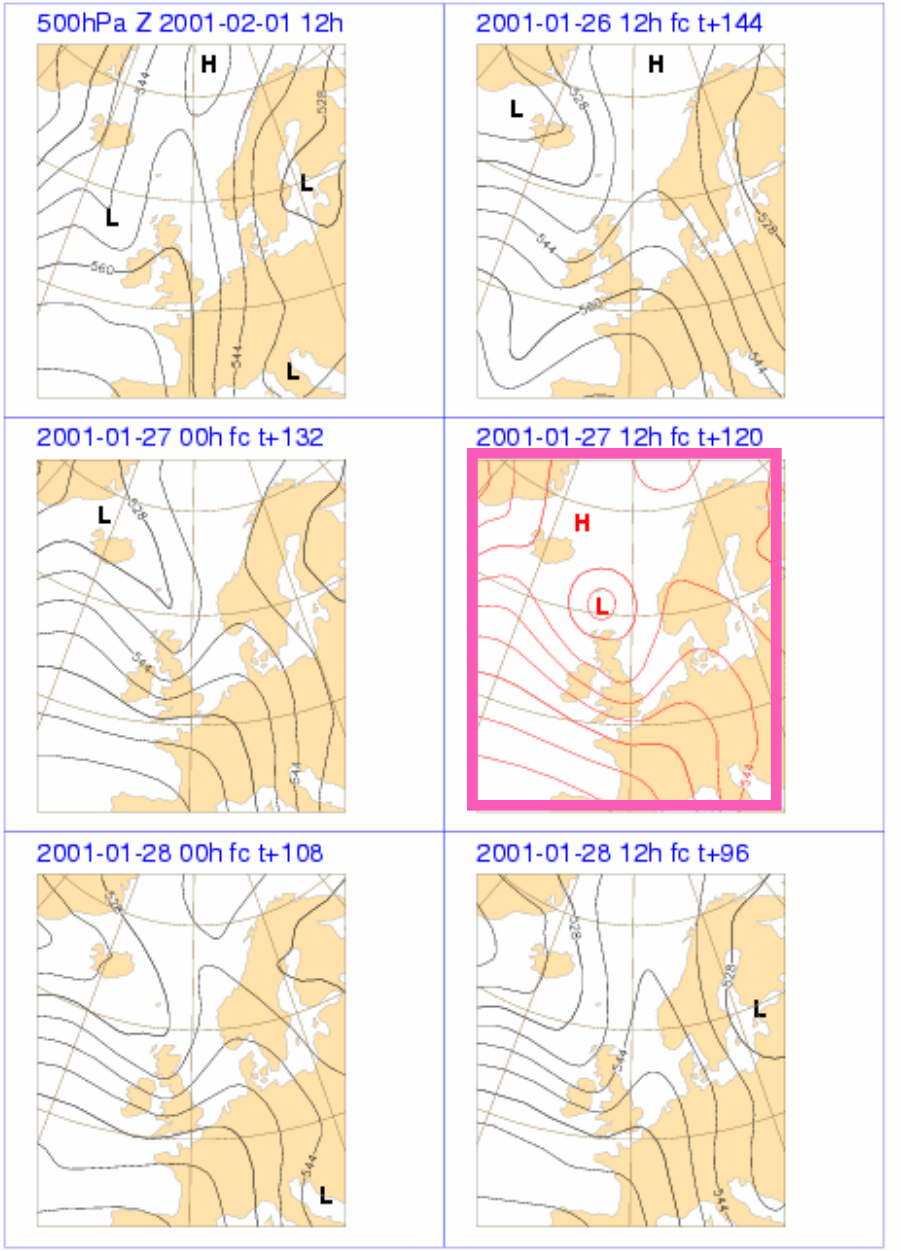



## Interestingly, none of the measurements were correct

The average of the 29 impacts was almost identical to the theoretically expected value (x) computed by Gauss and Laplace.

The accuracy of this mean was 5-6 higher than for any individual measurement

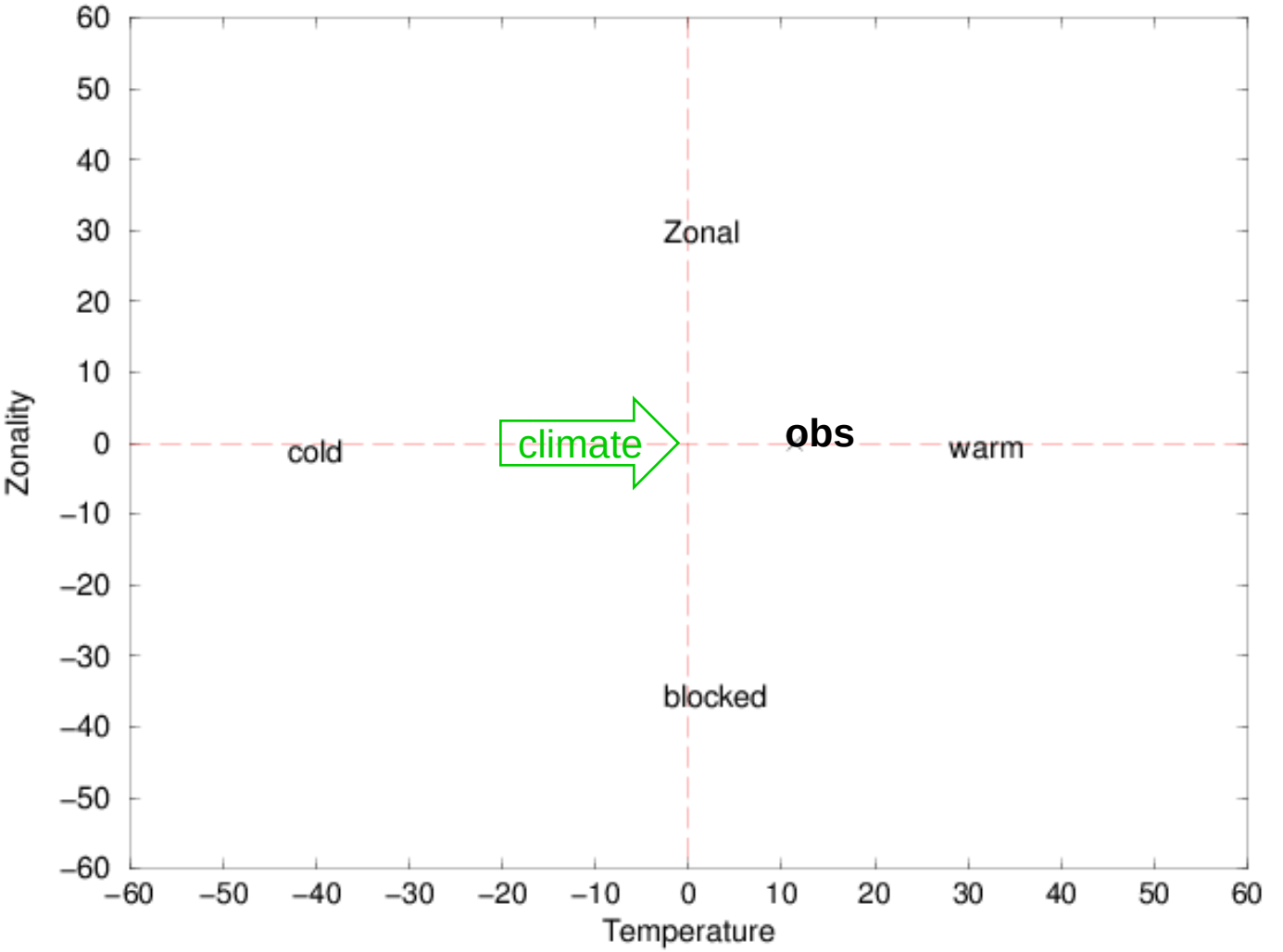
**Let us now use the principles and data from the 1803 Schlebush experiment to acquire a deeper understanding of the weather forecast process...**



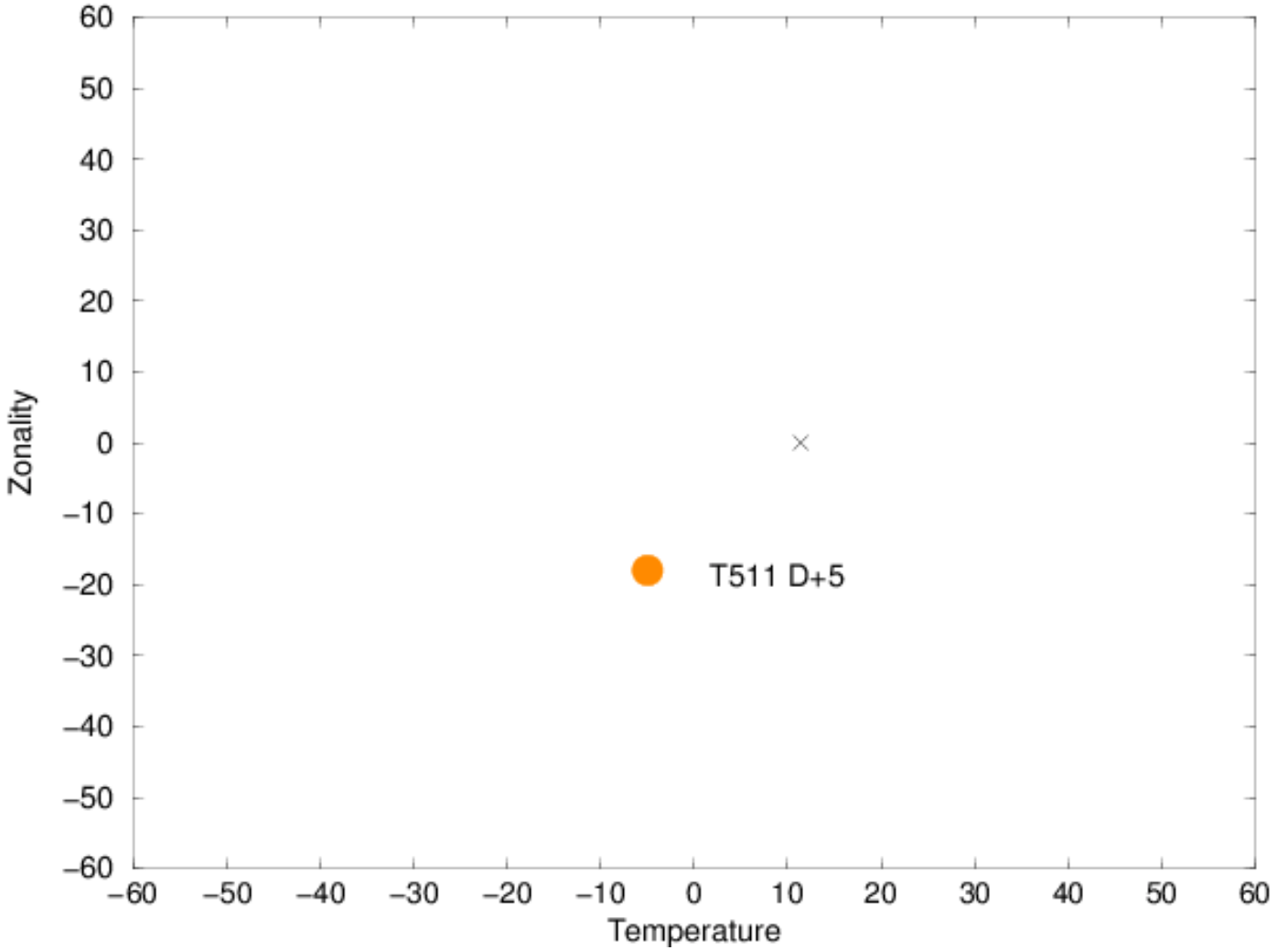

**All forecasts verify 12 UTC 1 Feb 2001:**

**A common situation: a sequence of consistent forecasts is broken by a sudden "jump"**

# Forecasting is in some sense the same as measuring, but now we measure in the future



# A five day forecast from the best model in the world



# Forecast quality depends on the scale of the weather systems

**a week**

heat waves

cold air outbreak

gales and storms

cold and warm fronts

**24 hours**

heavy precipitation

fog and stratus

**3 hours** showers and thunder

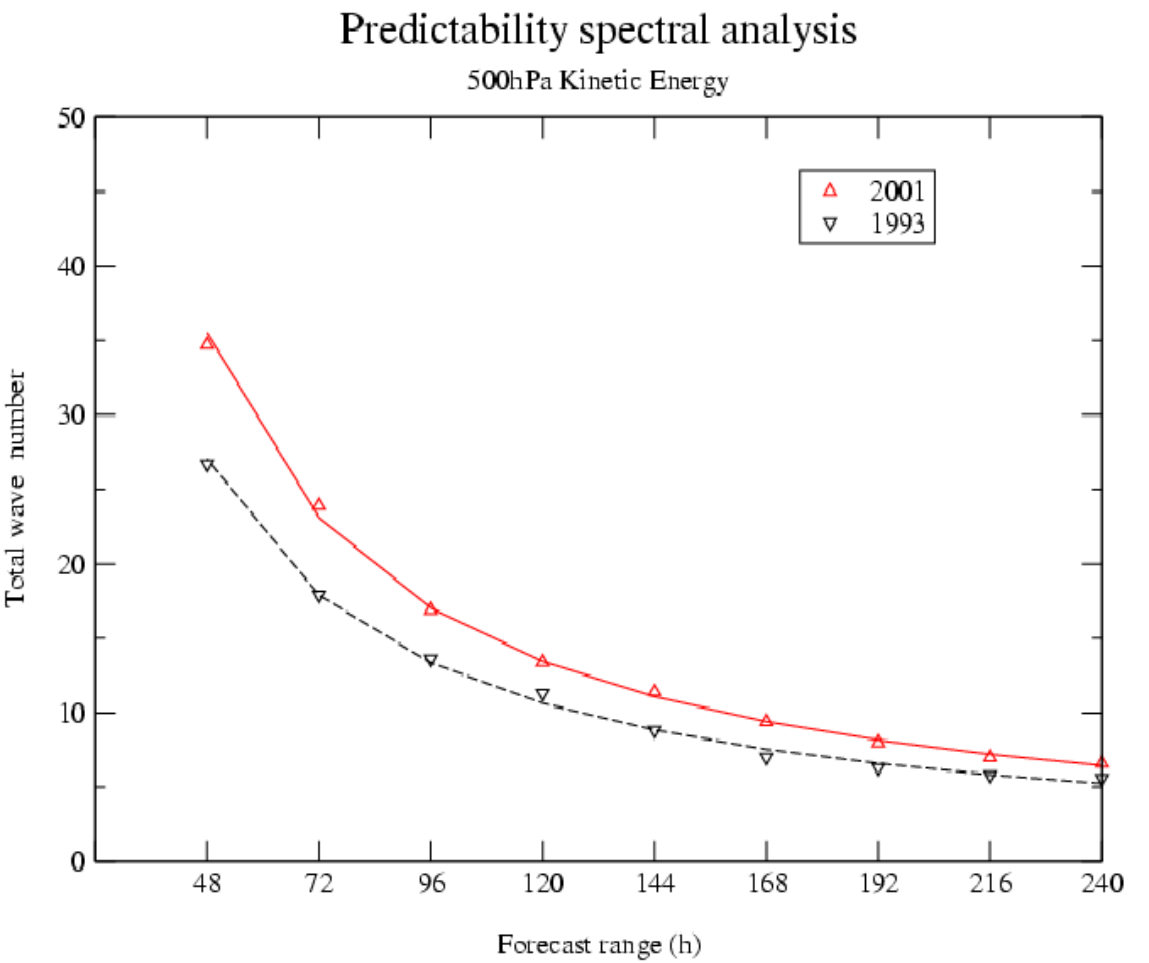
10 km 100 km 1000 km

# Larger scales are more predictable

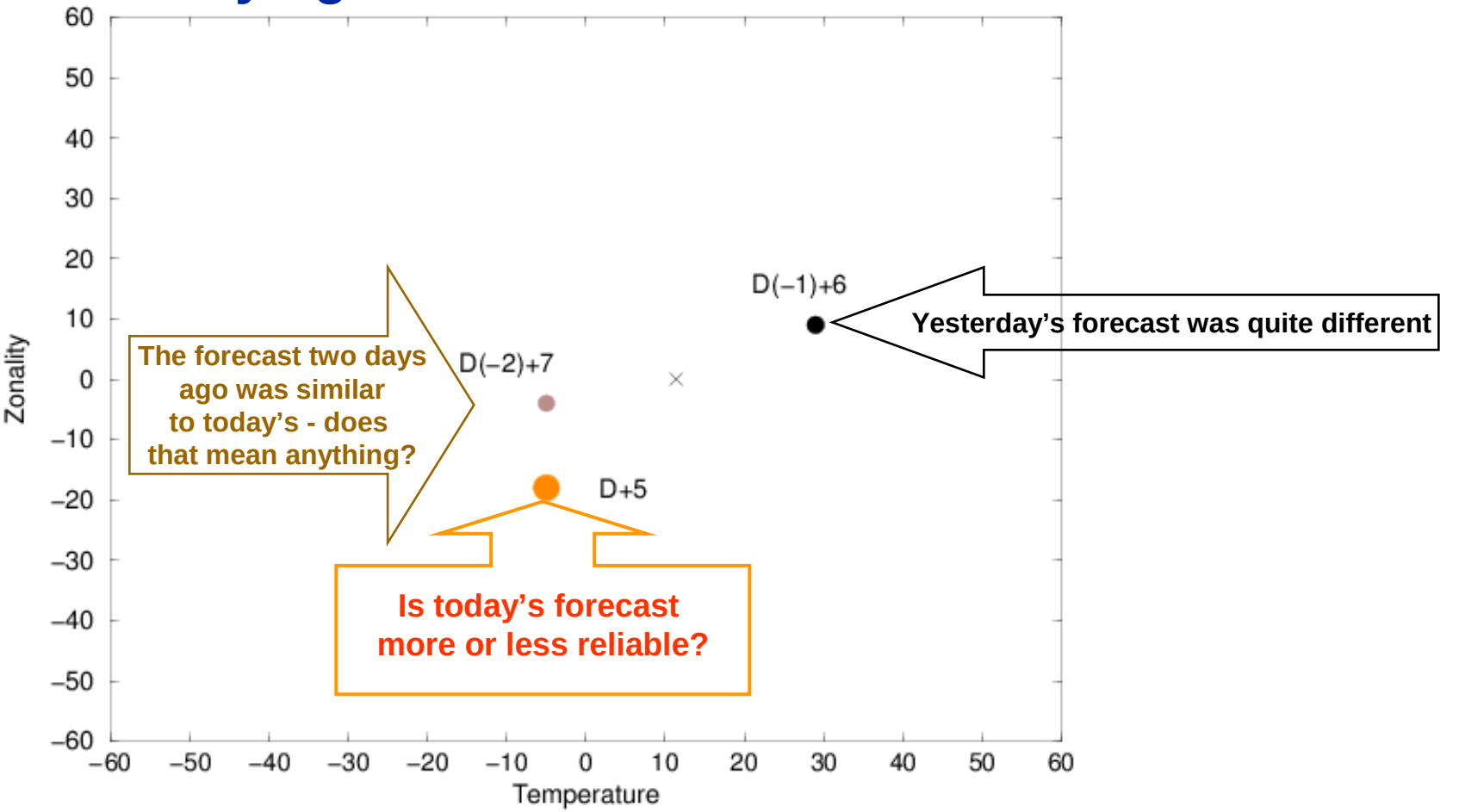
**The relation between the forecast range and the predictable scales**

**A truncation below 15 will be able to describe only the largest planetary (Rossby) waves**

**From 1993 to 2001 the ECMWF forecasts improved by one day**



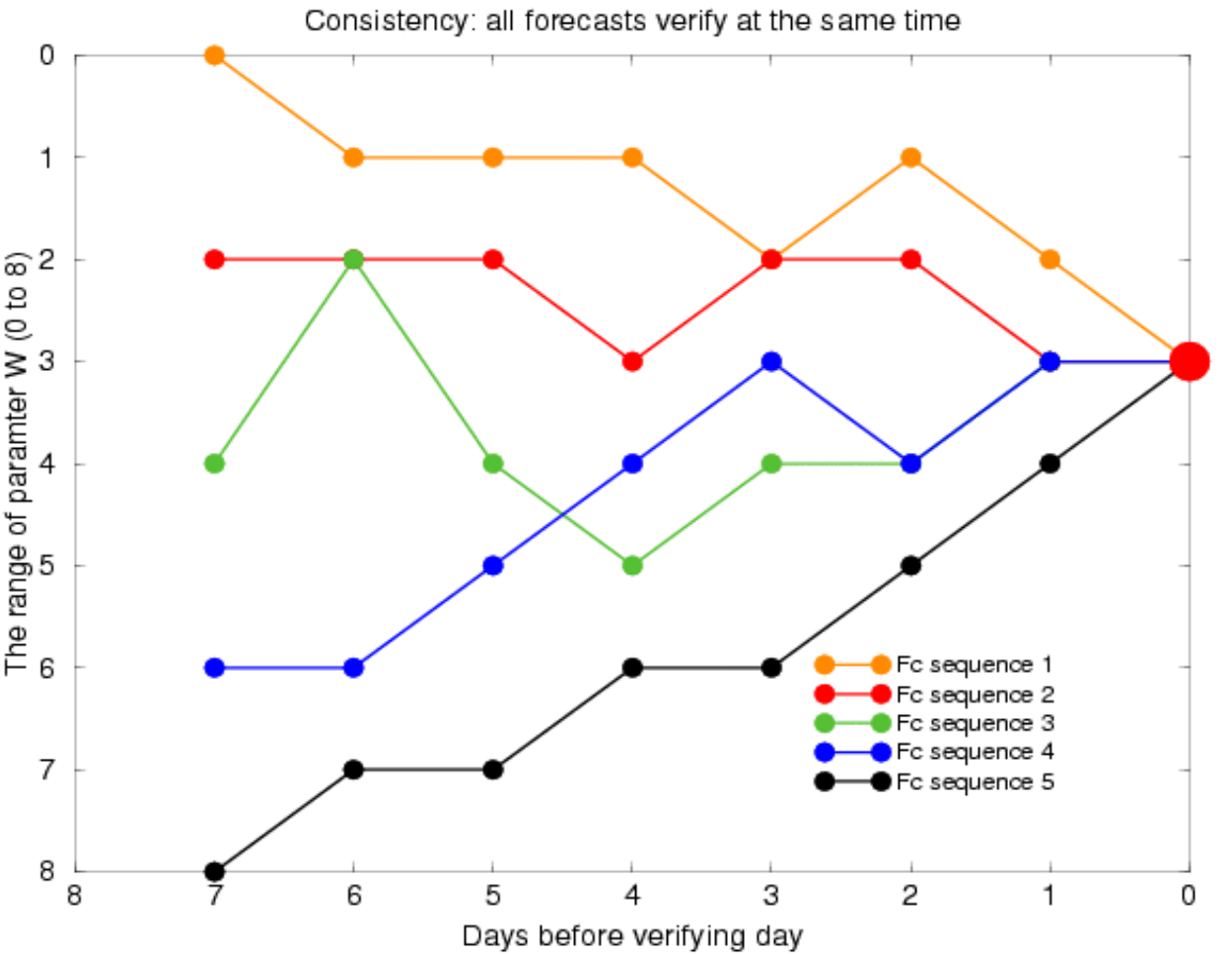
# The most recent forecast is often compared with previous forecasts verifying at the same time



# How we would like forecasts to evolve...

**Let us imagine a system with eight states from 0 to 8, which are forecast seven days ahead**

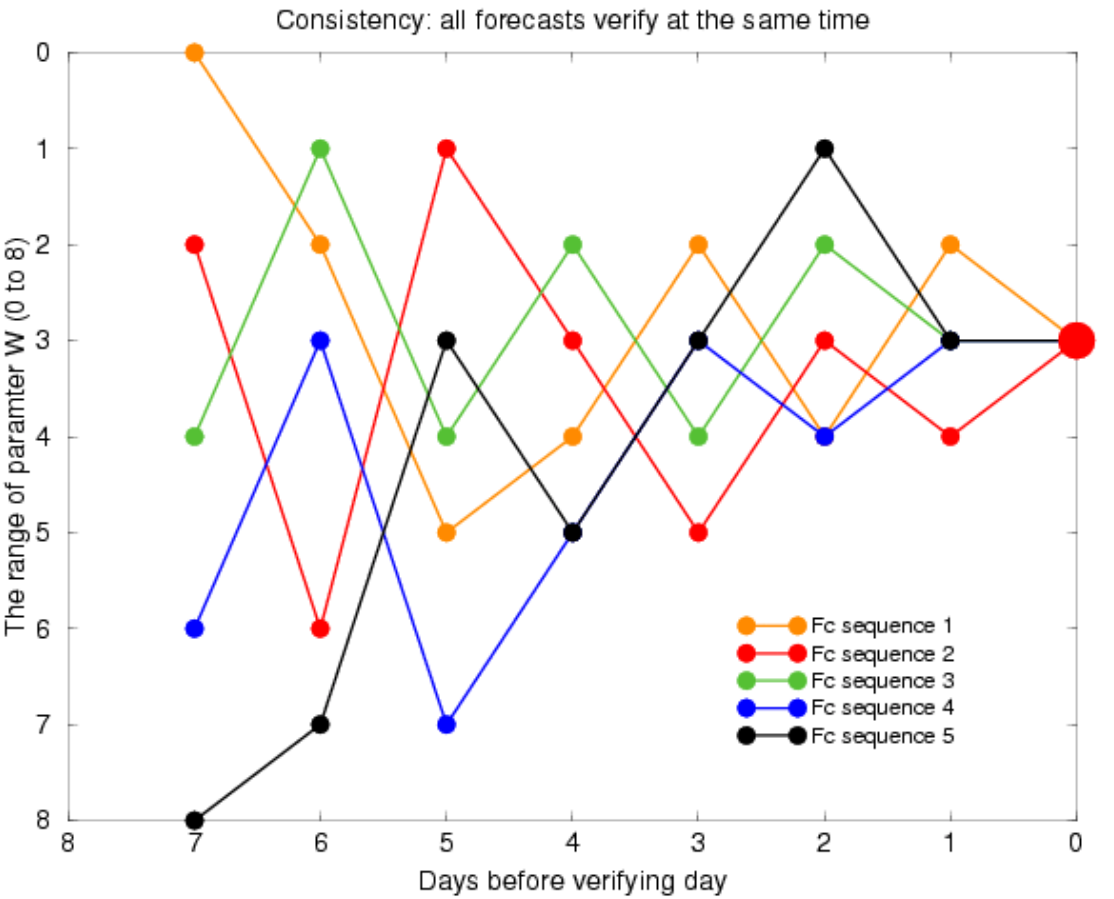
**In five different periods, when the final verification was 3, an automatic forecast system delivered values in this order**




**The first forecast seven days ahead were often wrong, but during subsequent days the system "zoomed in" toward the correct value**

# Instead it is often like this!

**Instead of a day by day orderly and consistent progression toward the truth, the forecast system jumps erratically before it finally, and rather late zooms in on the verifying value**

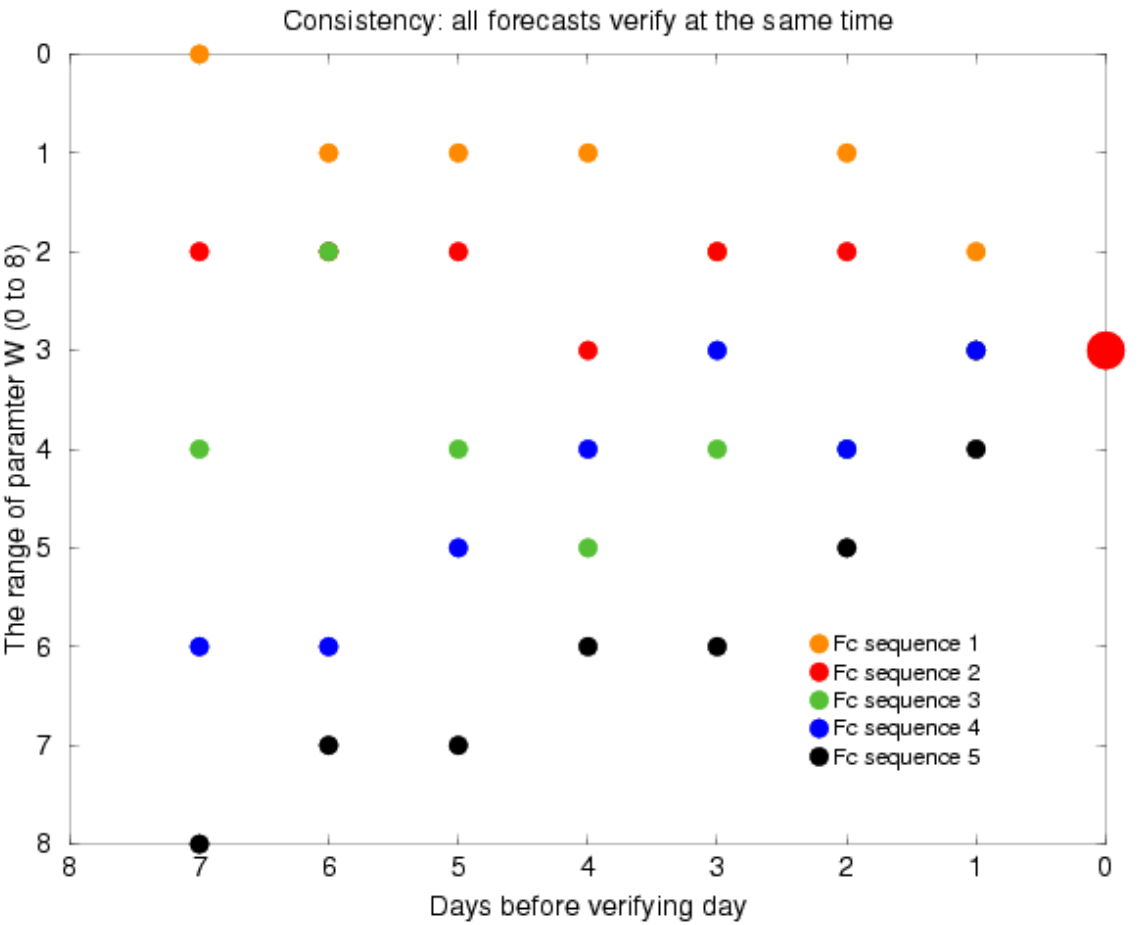


# Would we really like it to be like this?

**Once we remove the connecting lines, indicating the *succession* of forecast values, an other picture emerges**

**Many forecast are actually far off from the verifying value**

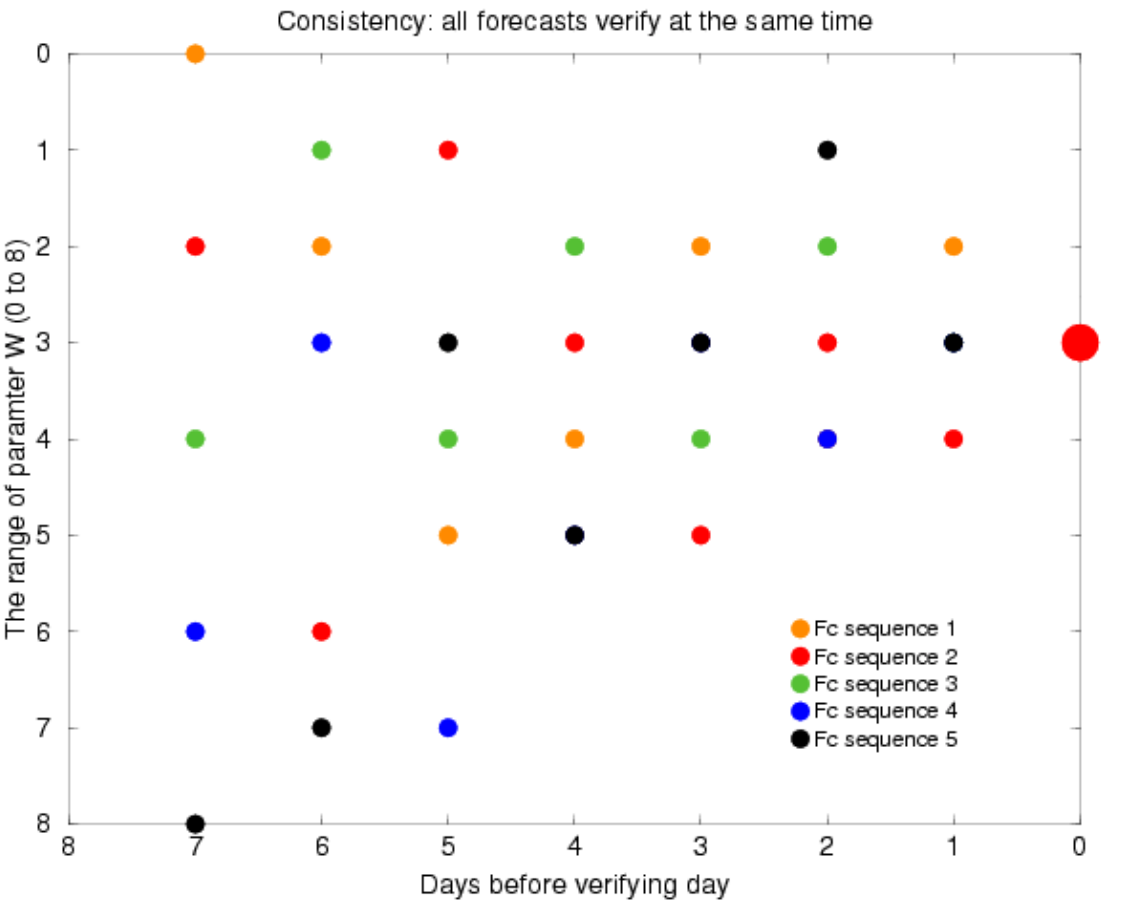
**The mean absolute error is 1.8**



# This isn't that bad after all!

**The forecasts, which looked very "jumpy" and erratic, actually scored somewhat better than the "orderly" or consistent ones**

**The mean absolute error is indeed only 1.4**



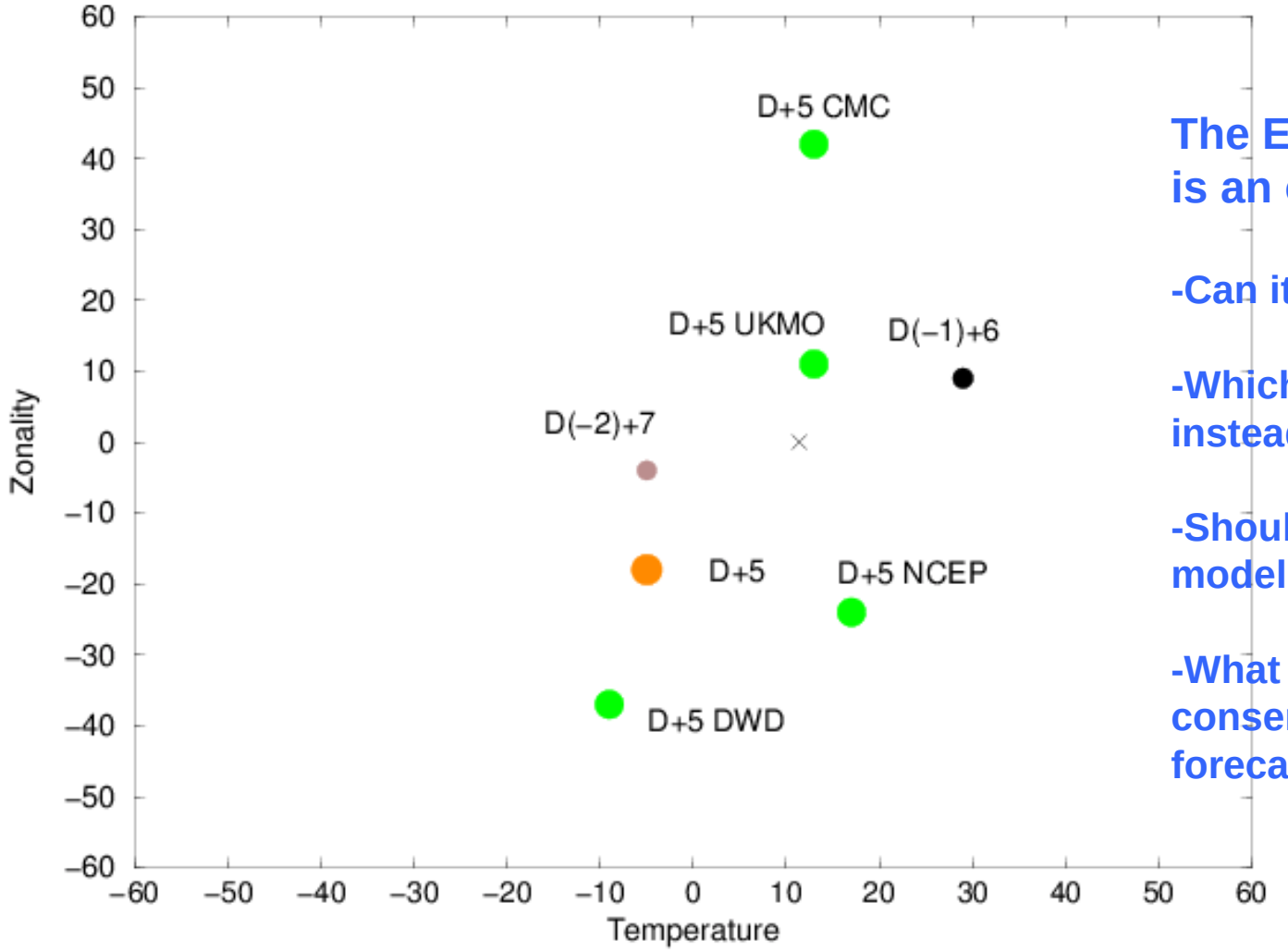

**The high degree of order in the first forecast sequences gave a confident impression which appeared to promise skill, whereas the more accurate system gave the opposite impression**

# Five golden rules

- © **"Jumpiness" is an unavoidable consequence of imperfect forecast skill by a realistic NWP model**
- © **The forecast should avoid over-interpreting details when the forecasts are consistent**
- © **When the forecasts are "jumpy", the forecaster should go for the latest or, preferably a synthesis of the last two or three**
- © **Do not be over-confident when the forecasts are consistent, do not be under-confident when they are "jumpy"**
- © **Treat the "jumpiness" as a source of extra information**

-Do not complain about jumpiness – make use of it!

# The forecast is often compared with output from other models run from slightly different initial conditions




**The ECMWF D+5 is an outlier:**

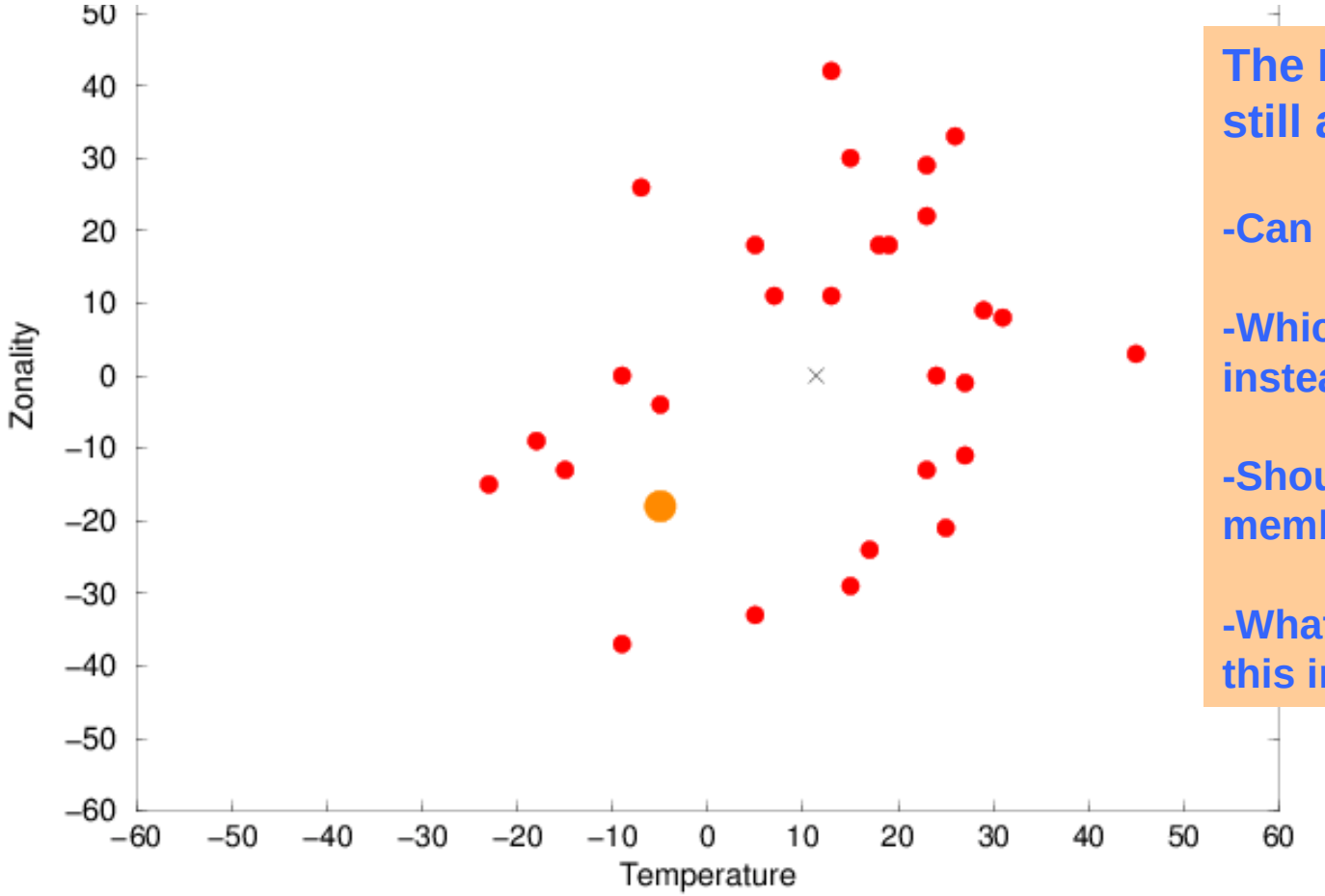
**-Can it be trusted?**

**-Which model should instead be trusted?**

**-Should *any specific* model be trusted?**

**-What does a consensus forecast offer?**

# A logical consequence of this praxis is to run a series of forecasts on the same model from slightly different conditions (EPS)




**The ECMWF D+5 is still an outlier:**

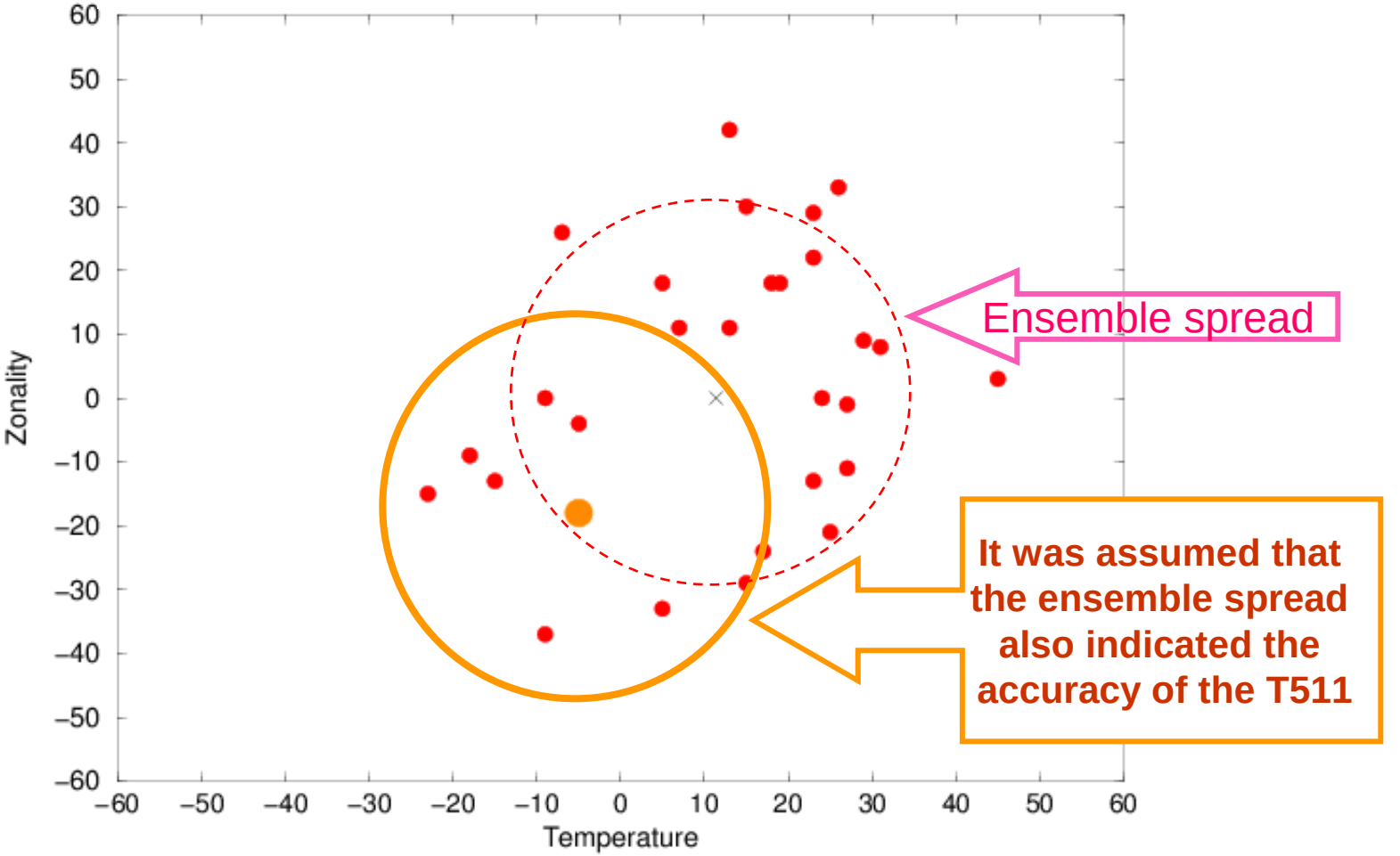
**-Can it be trusted?**

**-Which member should instead be trusted?**

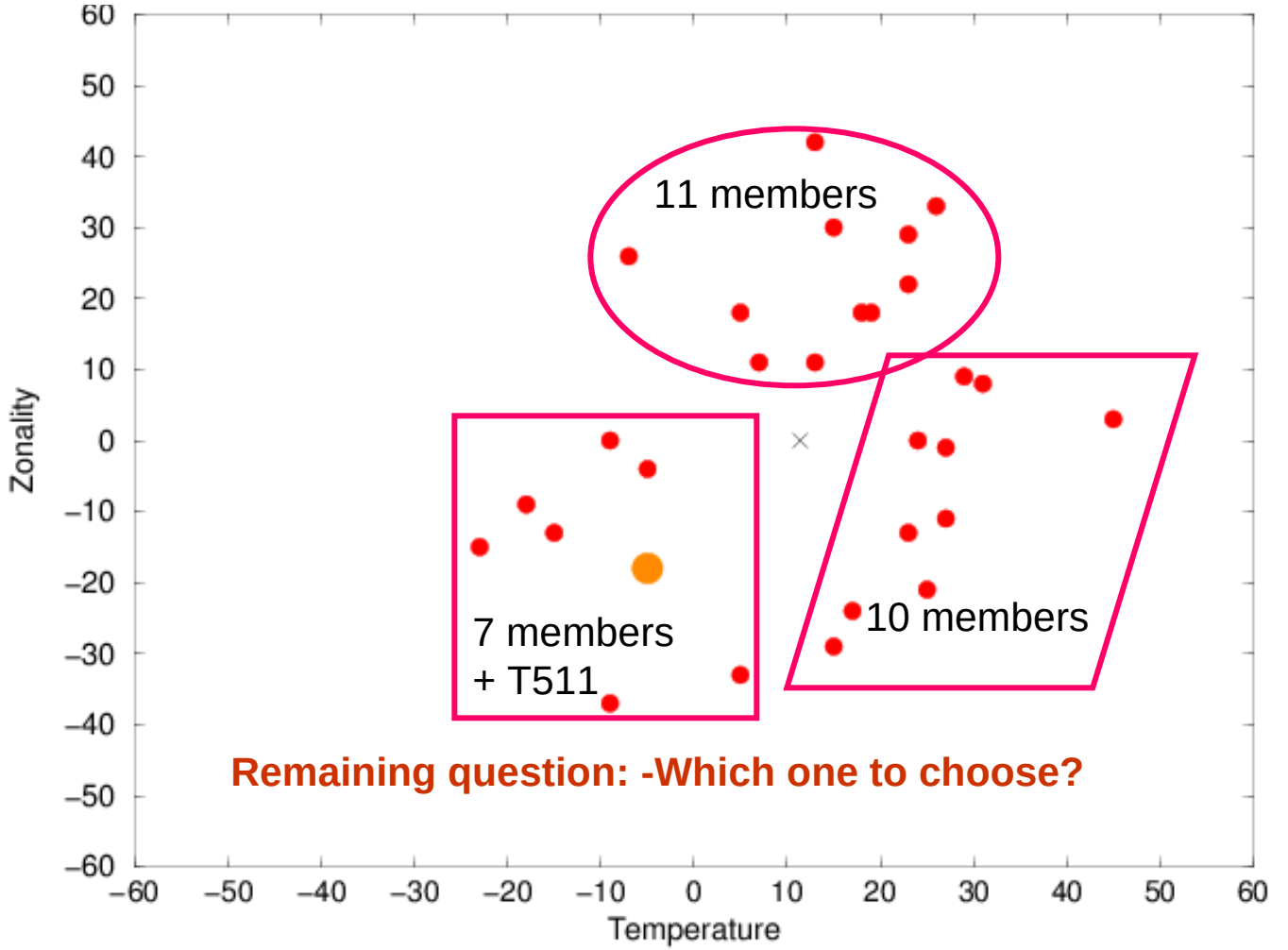
**-Should *any specific* member be trusted?**

**-What shall we do with this information??**

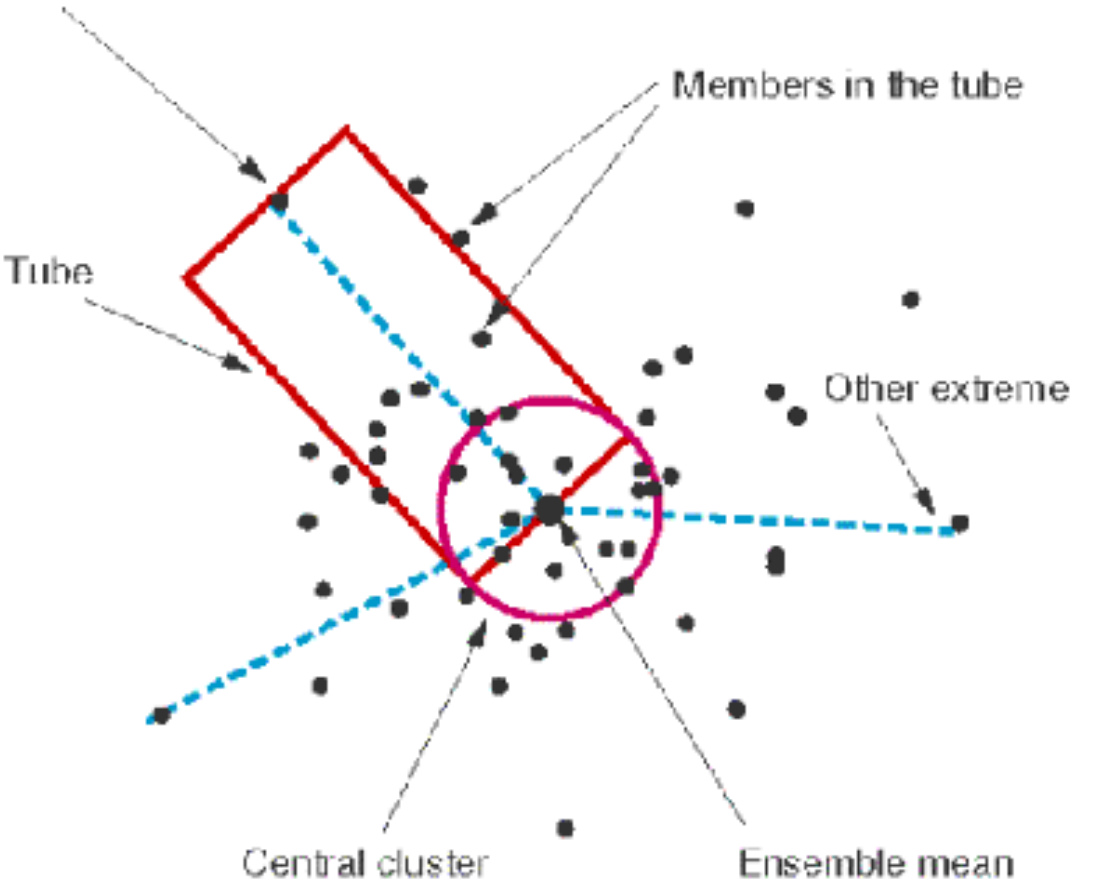
# The first suggestion was to use the spread of the ensemble to estimate the accuracy of the deterministic T511 forecast



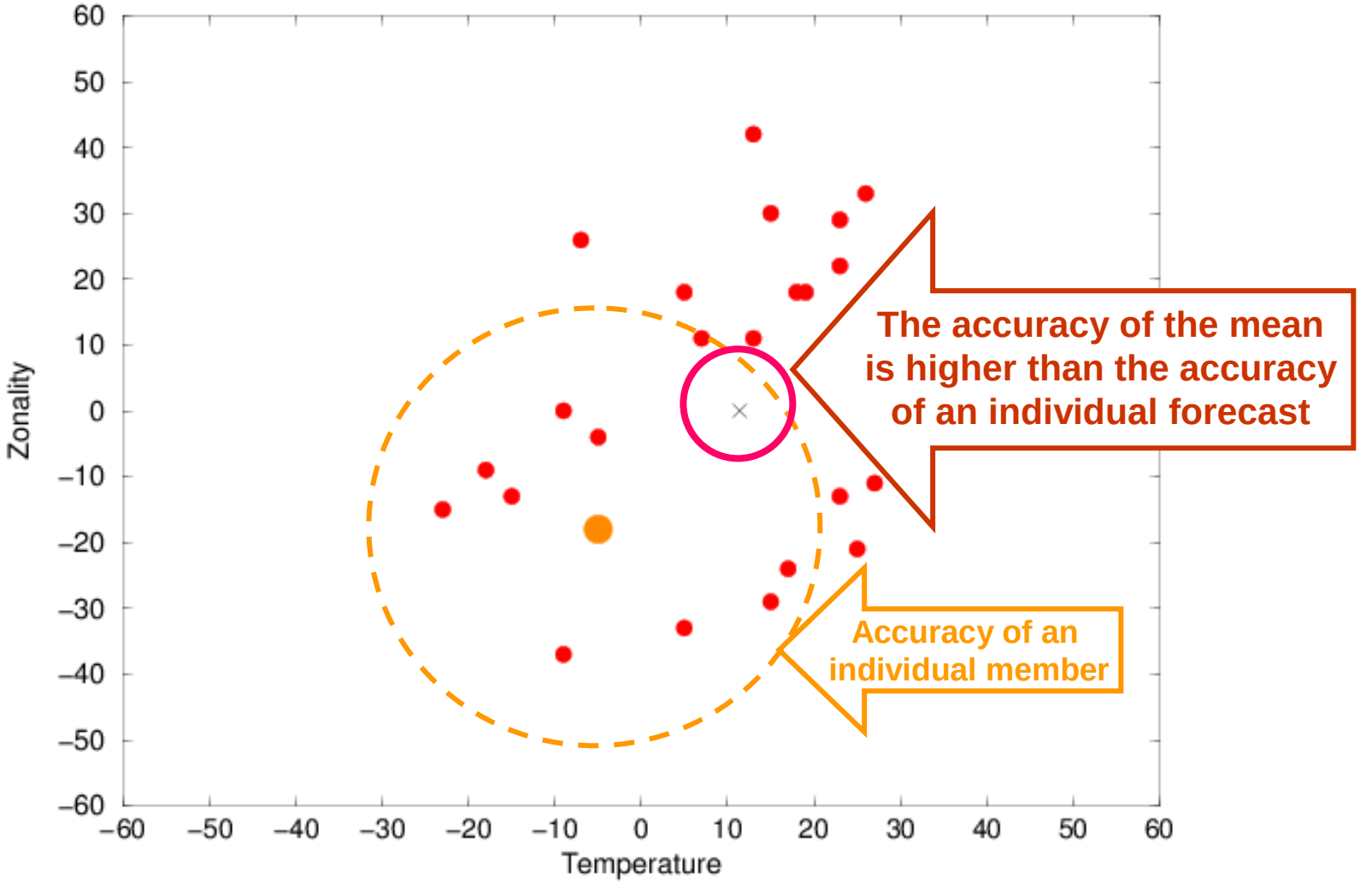
# An other strategy tried to cluster the output to identify different possible alternative regimes



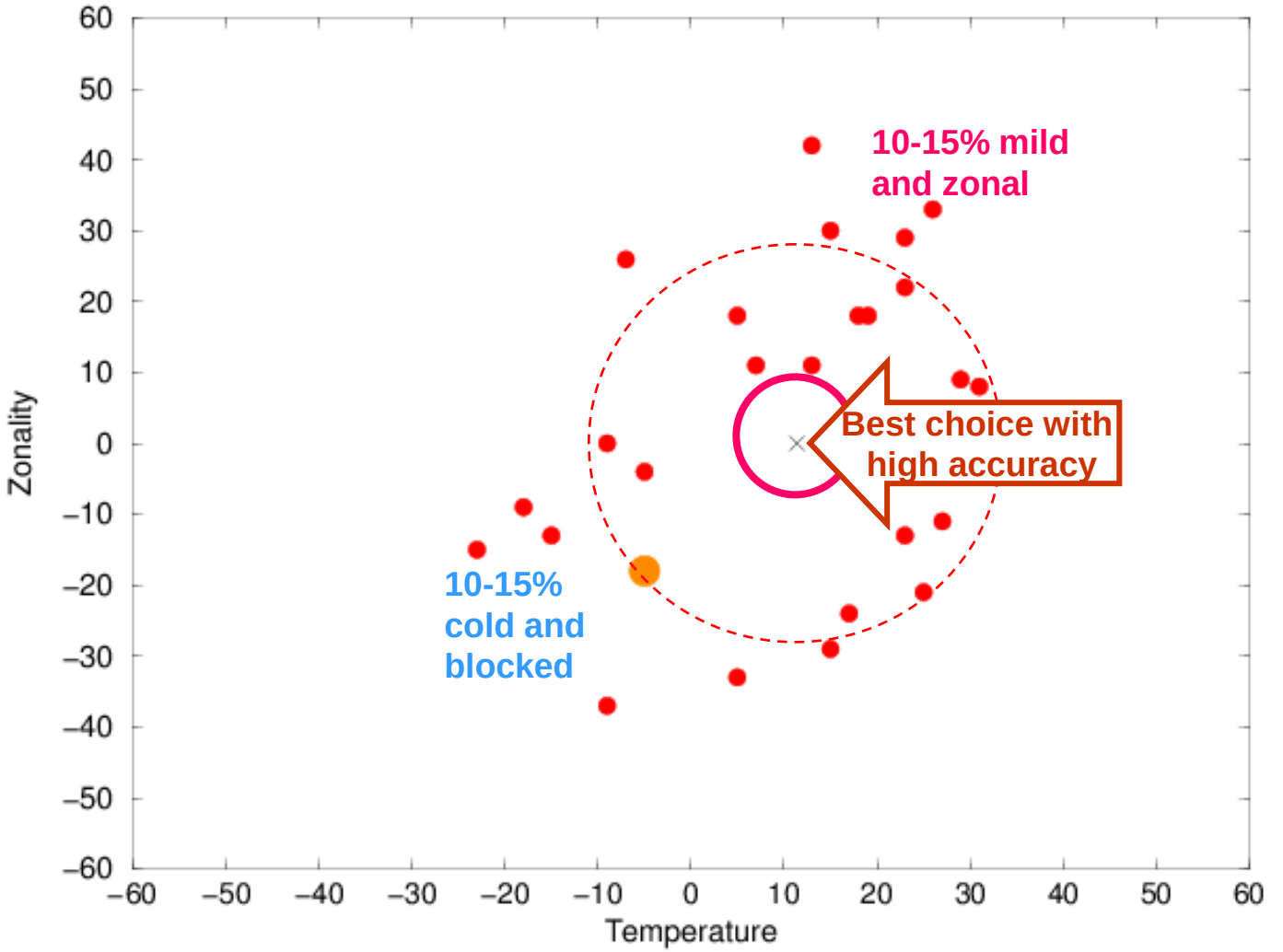
# The tubing system



# A third strategy is to use the ensemble to estimate the accuracy of the average



# The variance provides risk information

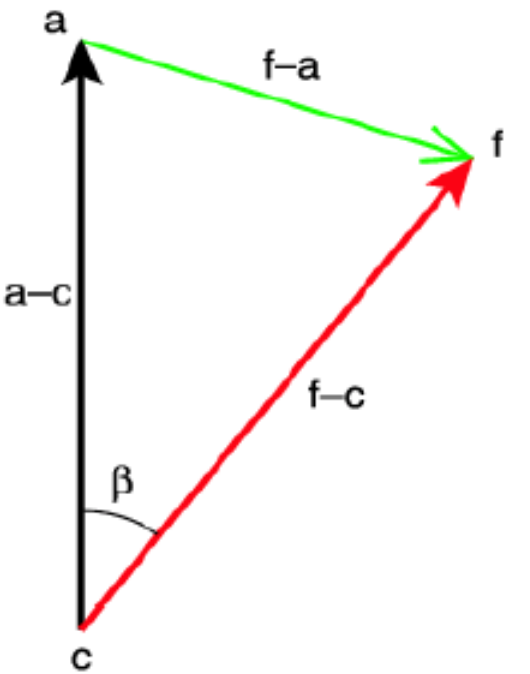


# A simple vector-geometrical alternative

**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**

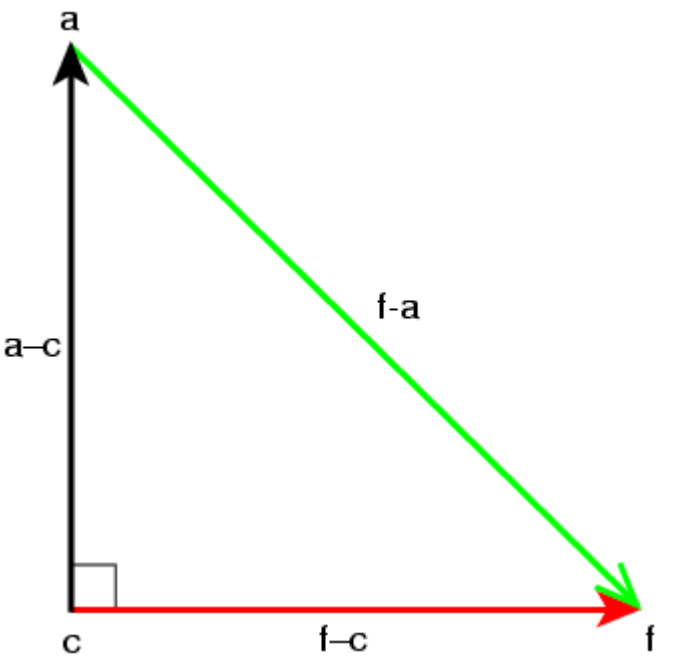
**With an underactive model f-c will become somewhat shorter. Also f-a will decrease and thus the RMSE**



# Uncorrelated forecasts

**When the forecast is lacking skill the f-c vector is perpendicular to the verifying vector a-c**

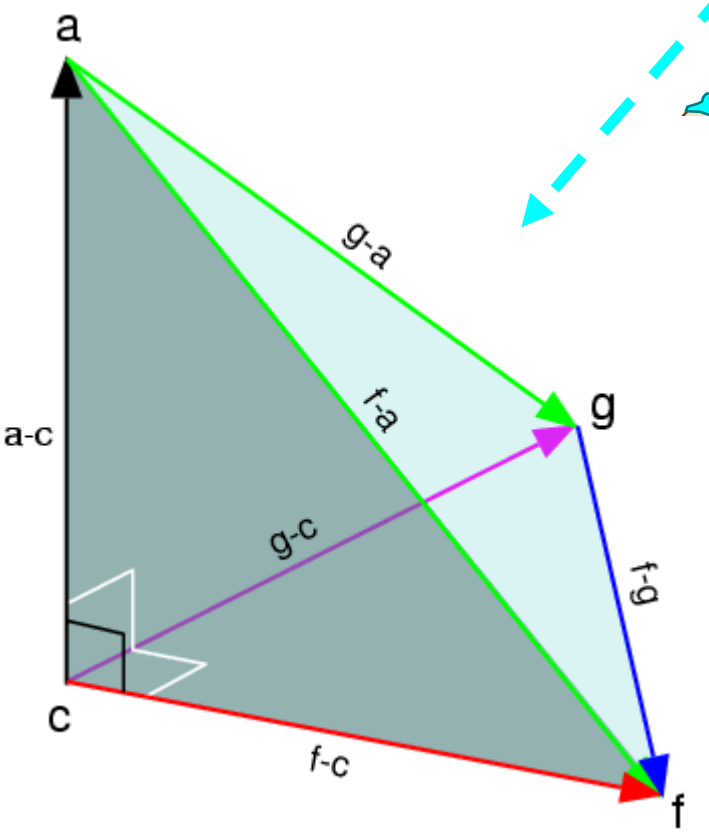
**Cosine of the angle 90° is zero which is also the value of the ACC (Anomaly Correlation Coefficient)**



# The spurious consistency-skill correlation

**Two forecasts systems (f) and (g) lack predictive skill an mutually uncorrelated.**

**This implies that all three vectors are perpendicular (90°)**

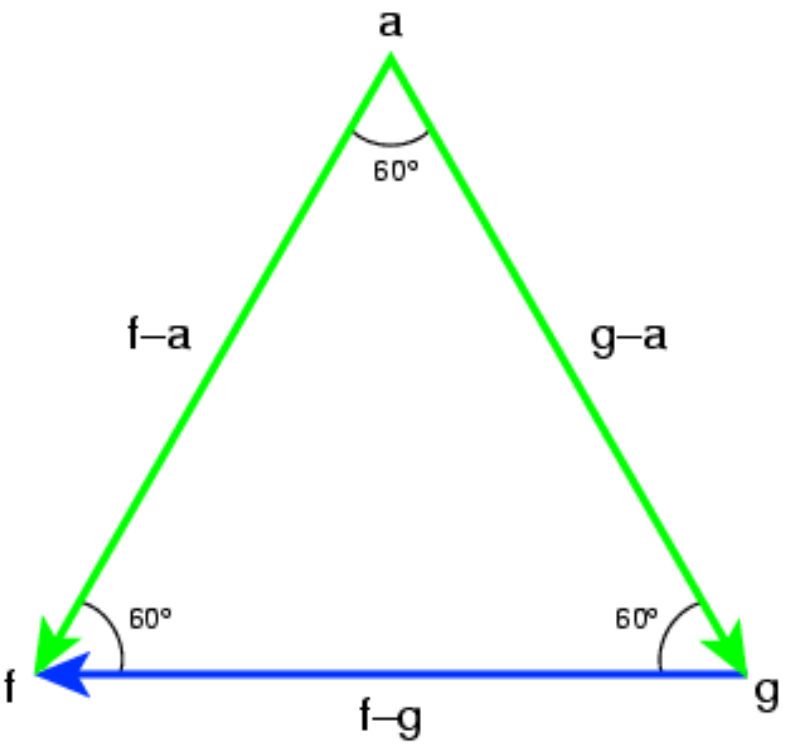



**Let us now watch this 3D figure from upper right...**

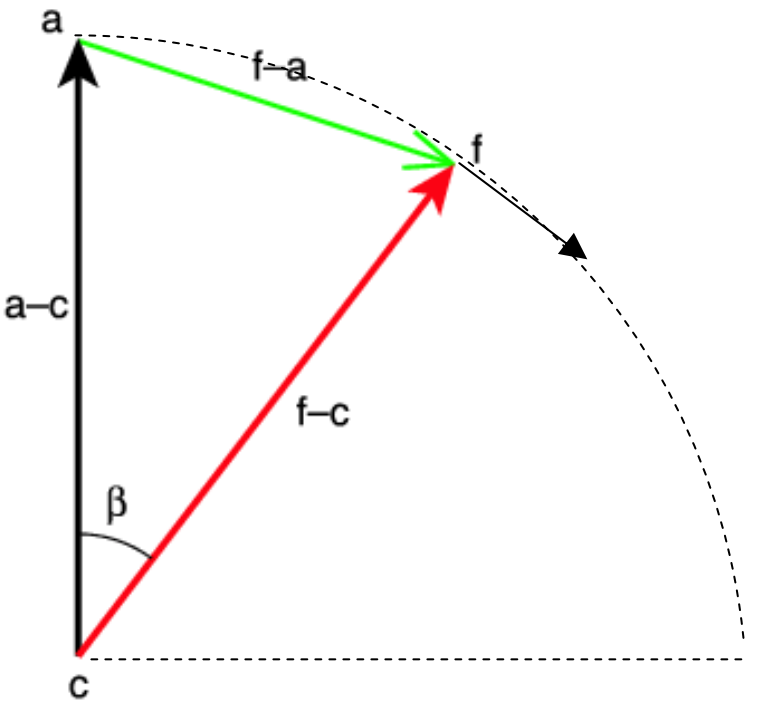
# The spurious consistency-skill correlation

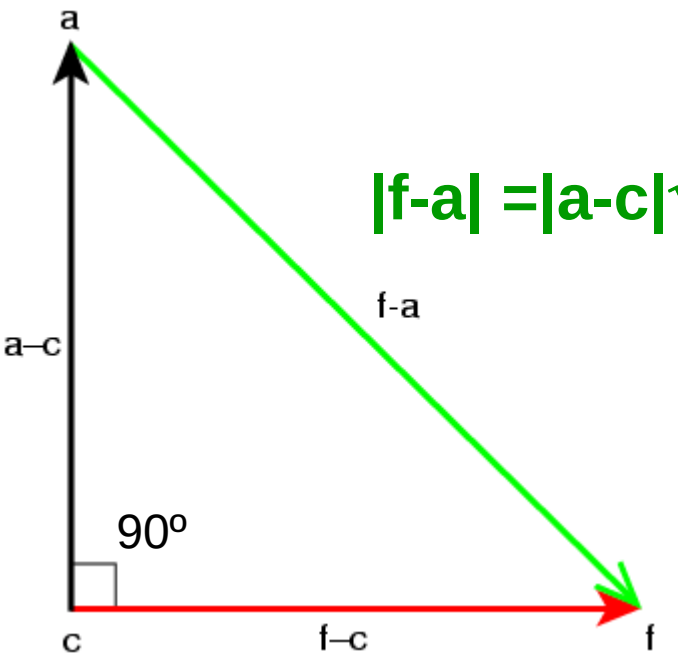
**Whereas the analysis vector (a) and the forecast vectors (f and g) are perpendicular, *their differences are not!* Their mutual angles are 60° which implies correlations of 50%.**

**It is when the forecasts start to display skill and mutual correlation that the 50% correlation starts to decrease to the 30% level sometimes reported at a D+5 or D+6 range**



# The RMS error saturation level



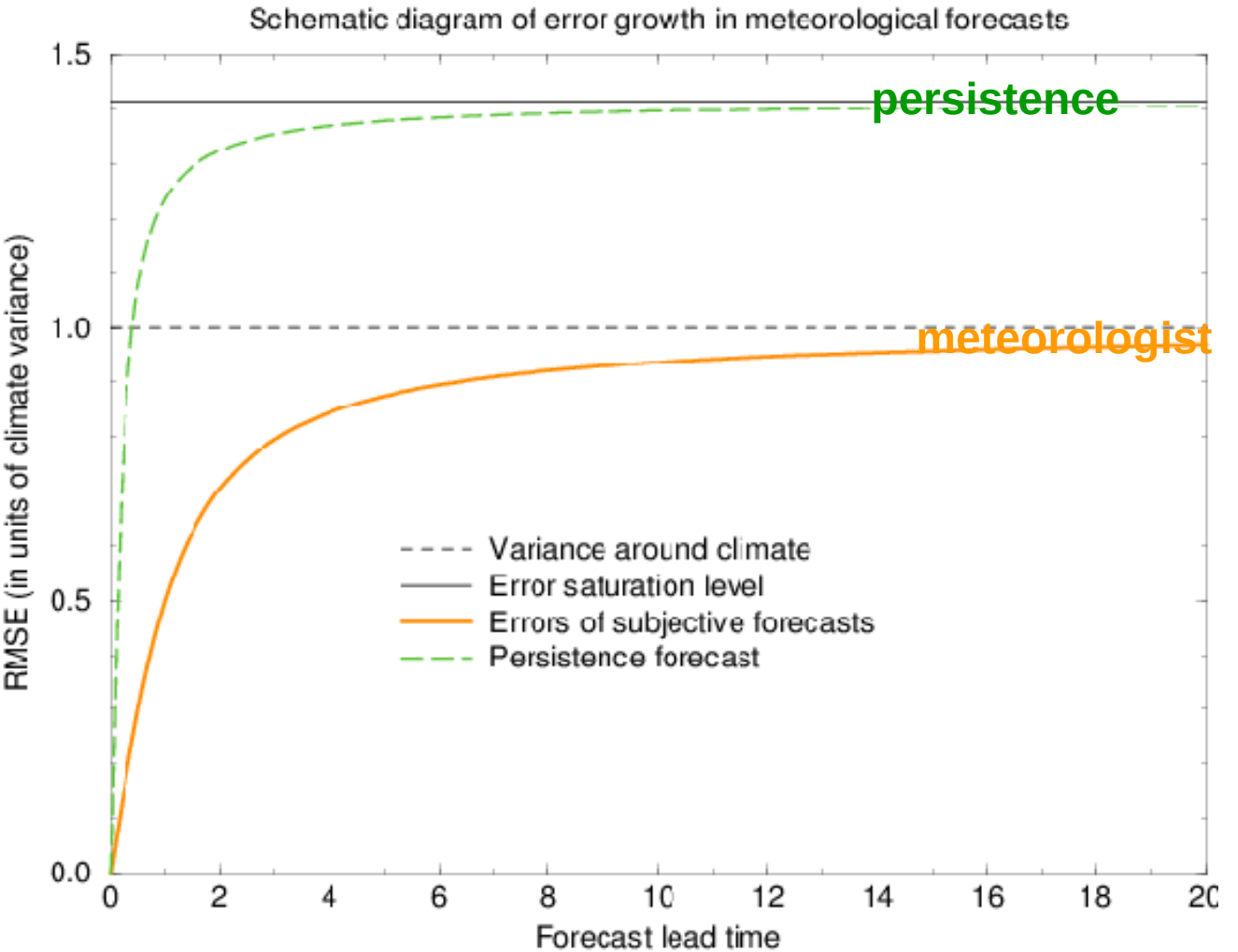

$|f-a| = |a-c|\sqrt{2}$

**With decreasing forecast accuracy the angle β will increase.**
**The maximum RMSE equals the variability times √2**

# The pre-NWP forecast accuracy

**A schematic illustration of the increase of RMSE with forecast time**

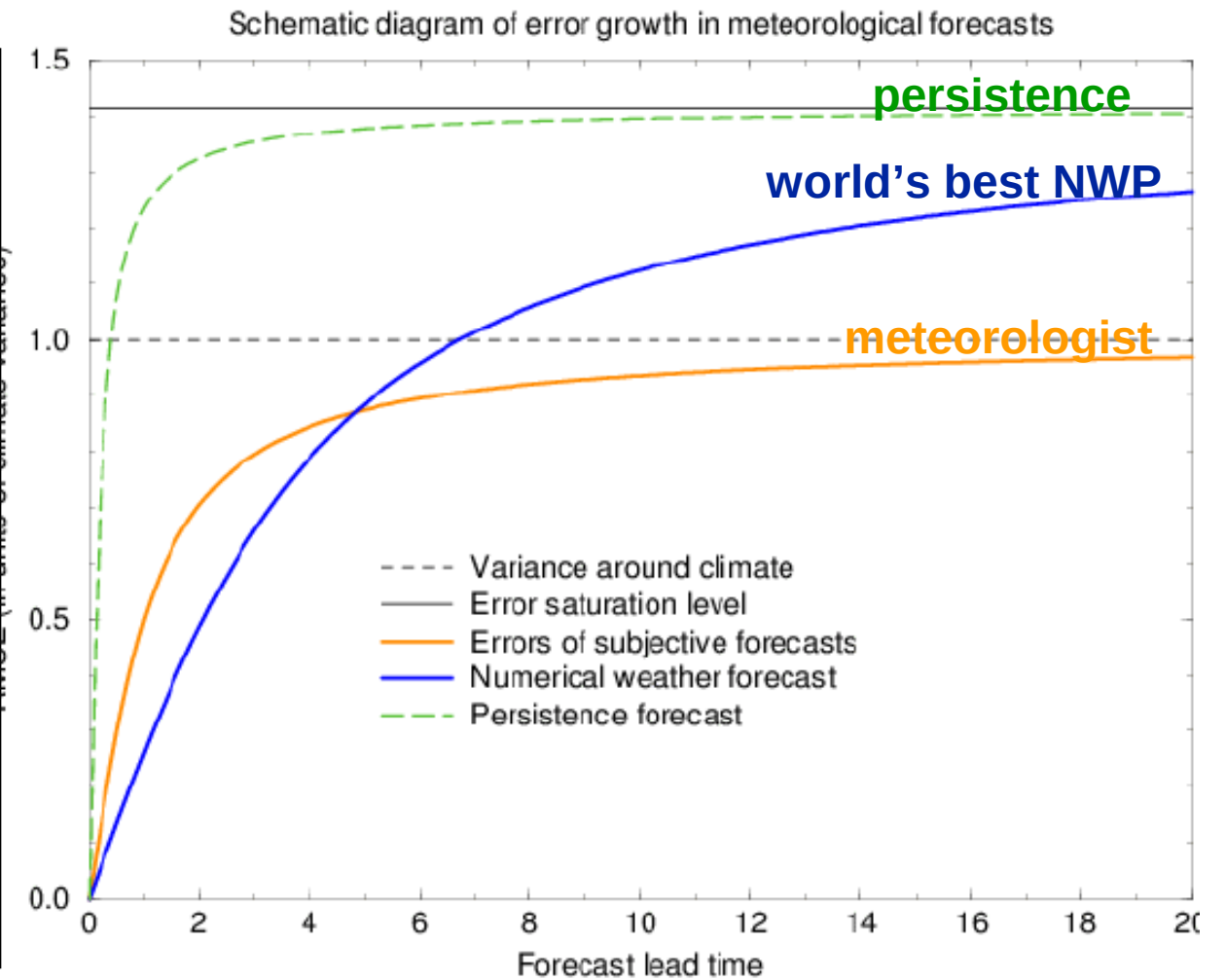
**The pre-NWP forecaster started from a persistence forecast which he skillfully extrapolated into the future, converging towards climate for longer ranges**




**The time unit can be anything from hours to days depending on the parameter (hours for clouds, days for temperature)**

# NWP more accurate - but also less

**A good NWP model is able to simulate all atmospheric scales throughout the forecast. It has the same variance as the observations and the persistence forecasts, which yields an error saturation level 41% above the climate**

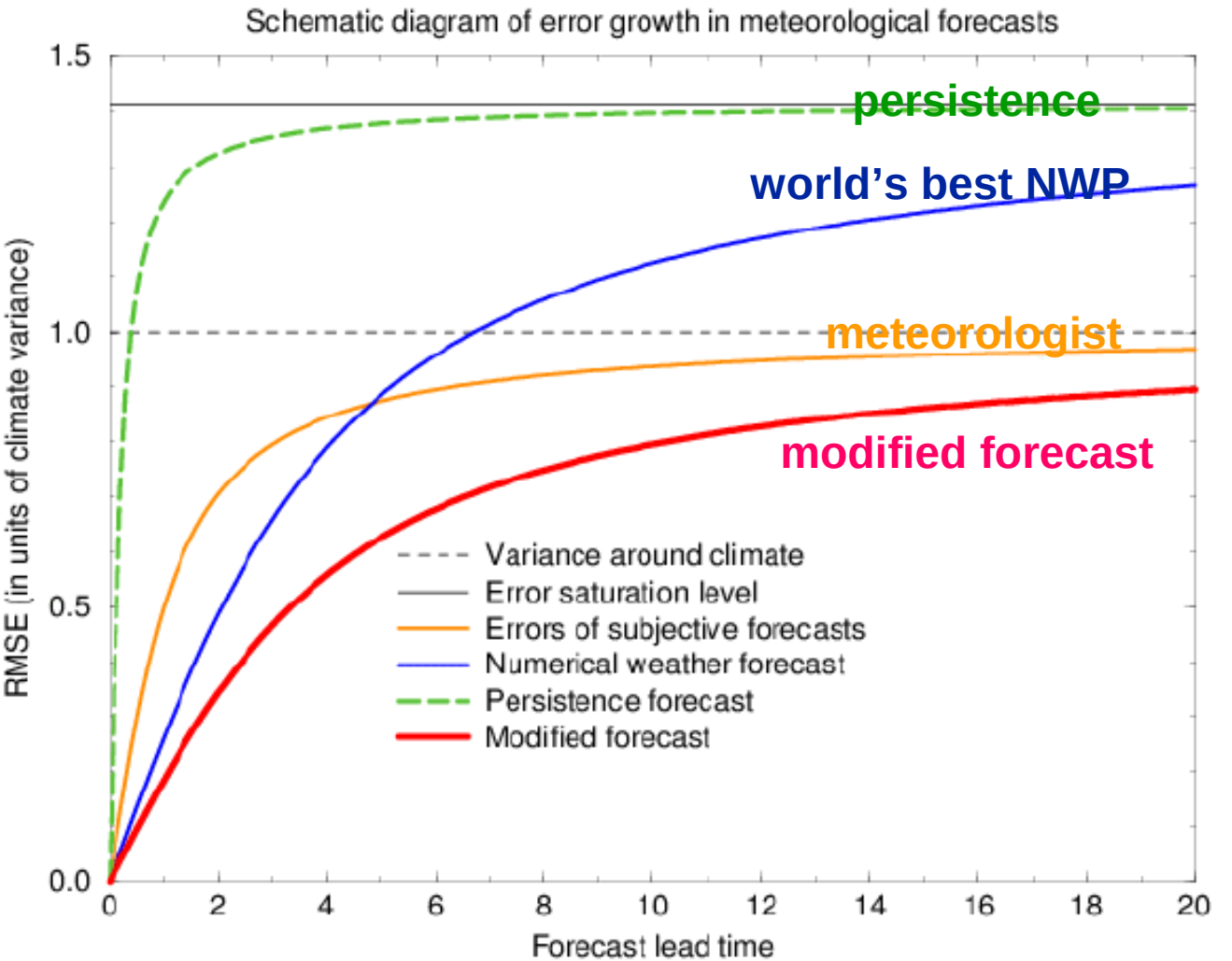


# The art of good forecasting

**The way out of the dilemma:**

**Combine the high accuracy of NWP in the short range with a filtering of the non-predictable scales for longer ranges**

**This can be done both with and without the EPS**



# What defines the Red Line (a good customer orientated system)

**A team of well educated, dedicated forecasters with a lot of time to make judgements**

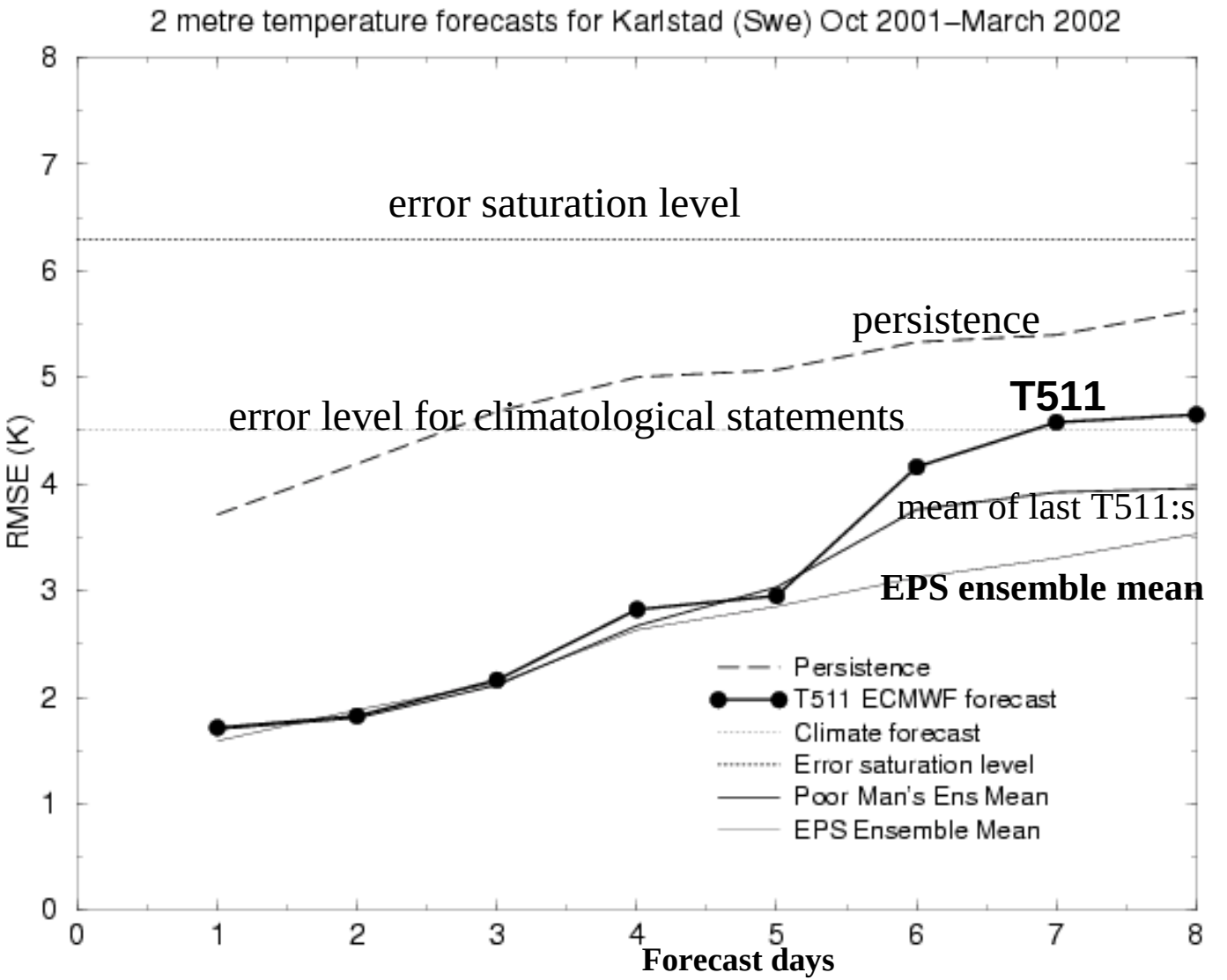
**Statistical post-processing where past experience is introduced in a optimal way**

**The mean (or median) of the forecasts in the ensemble system**

**A combination of two or three of the above**

# Observed 2 m T error growth (Karlstad)

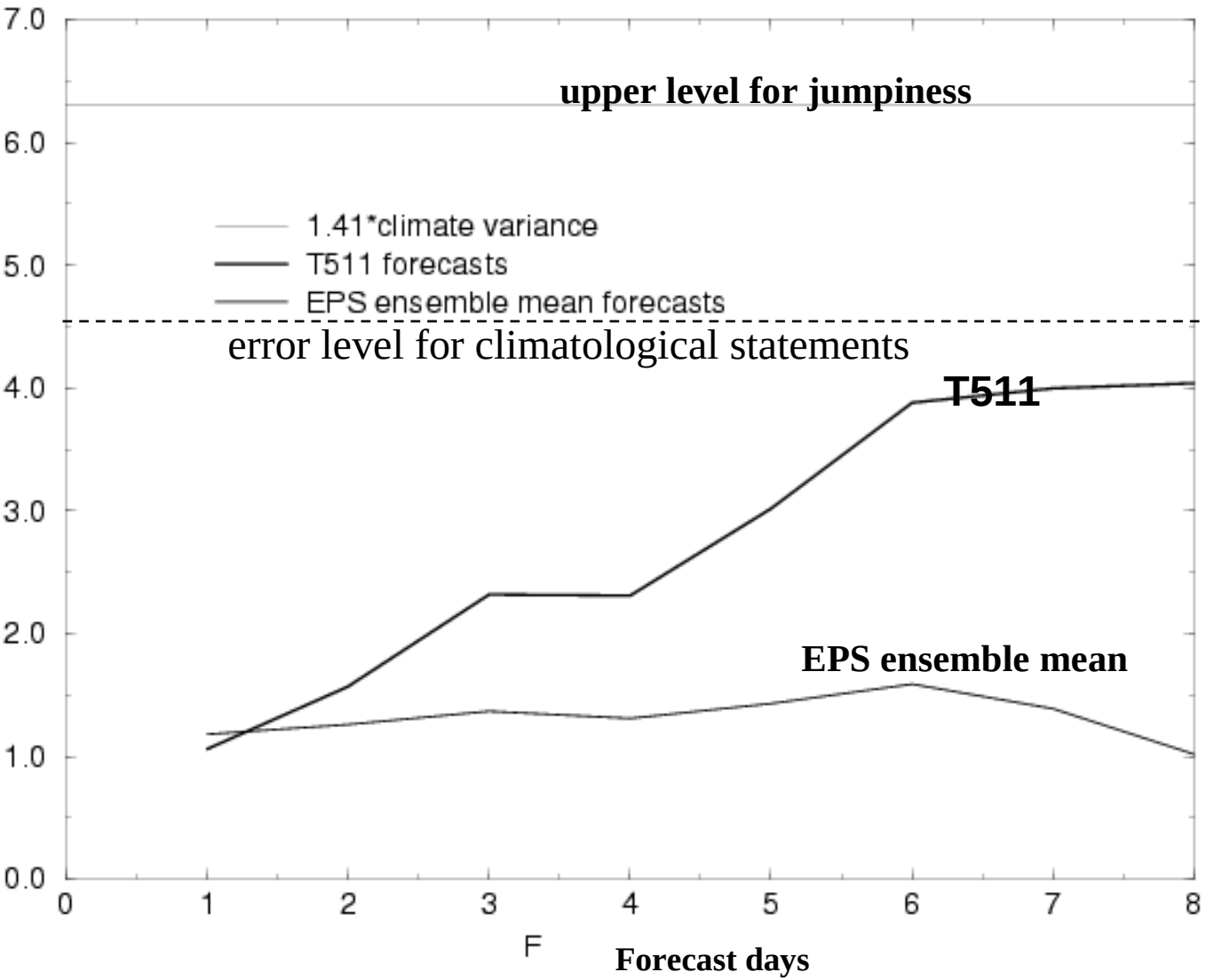
**The accuracy of different forecast methods for a city in central Sweden winter 2001-2002: persistence, climate, T511,** an average of the last three days T511 ("poor man's ensemble mean") **and the T255 EPS ensemble mean**



# Jumpiness in 2 m T forecasts (Karlstad)

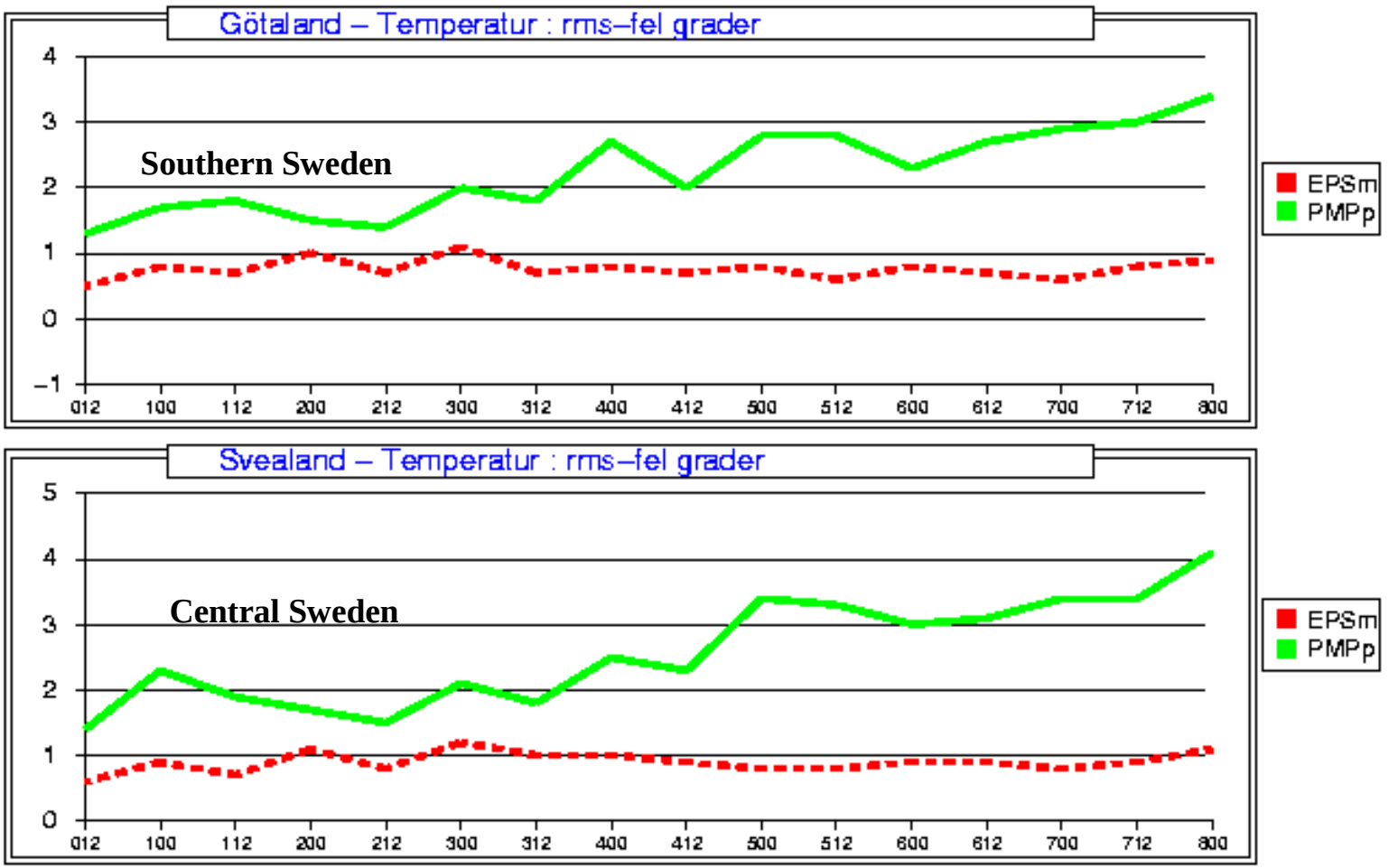
**The ensemble mean is not only the most accurate, it is by far also the least "jumpy"**

**While the T511 makes an average 24 hour "jump" of 2 - 4°K, the T255 ensemble mean is stable at around 1° K**



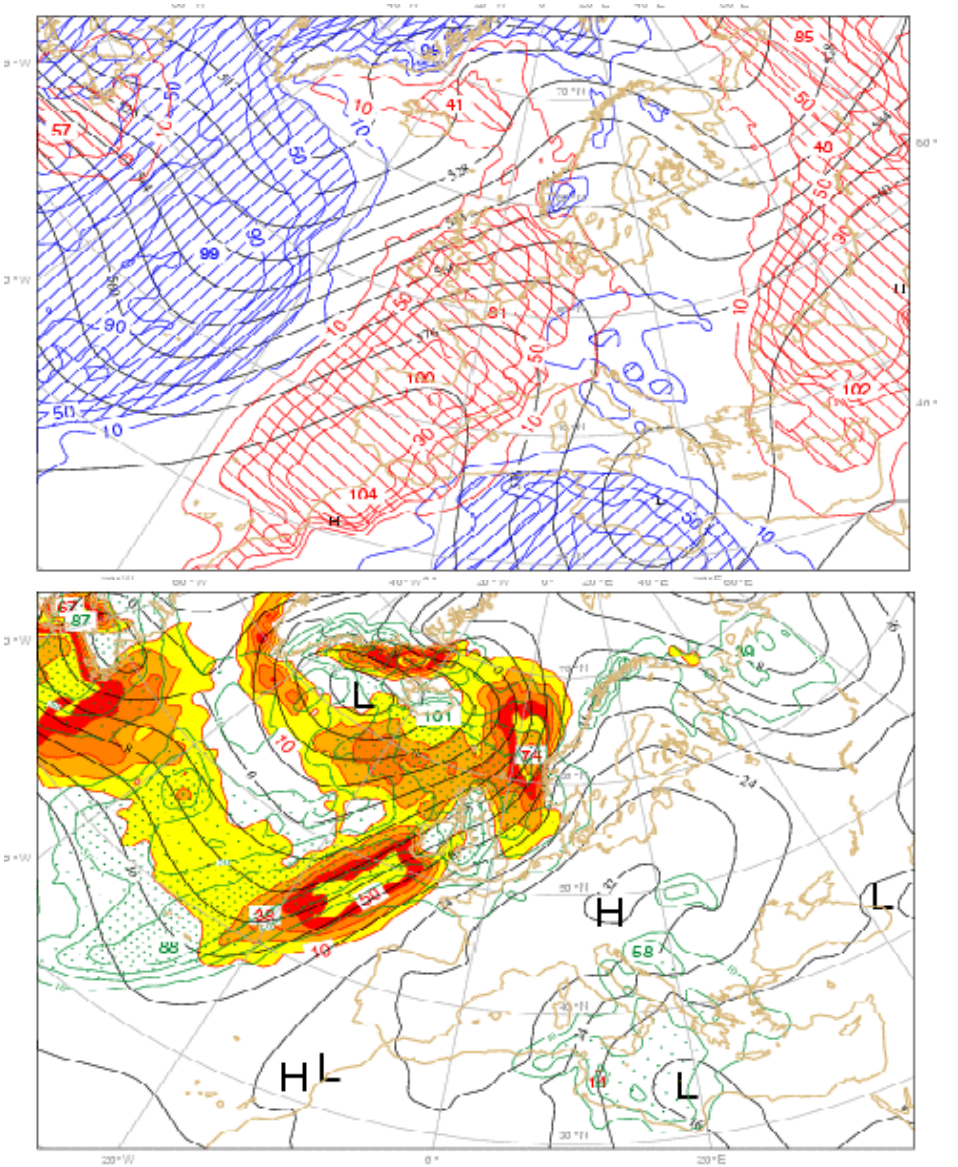
# “Jumpy” 2 m T forecasts winter 2002-2003

**The 24 h jumpiness for T511 and the T255 ensemble mean**

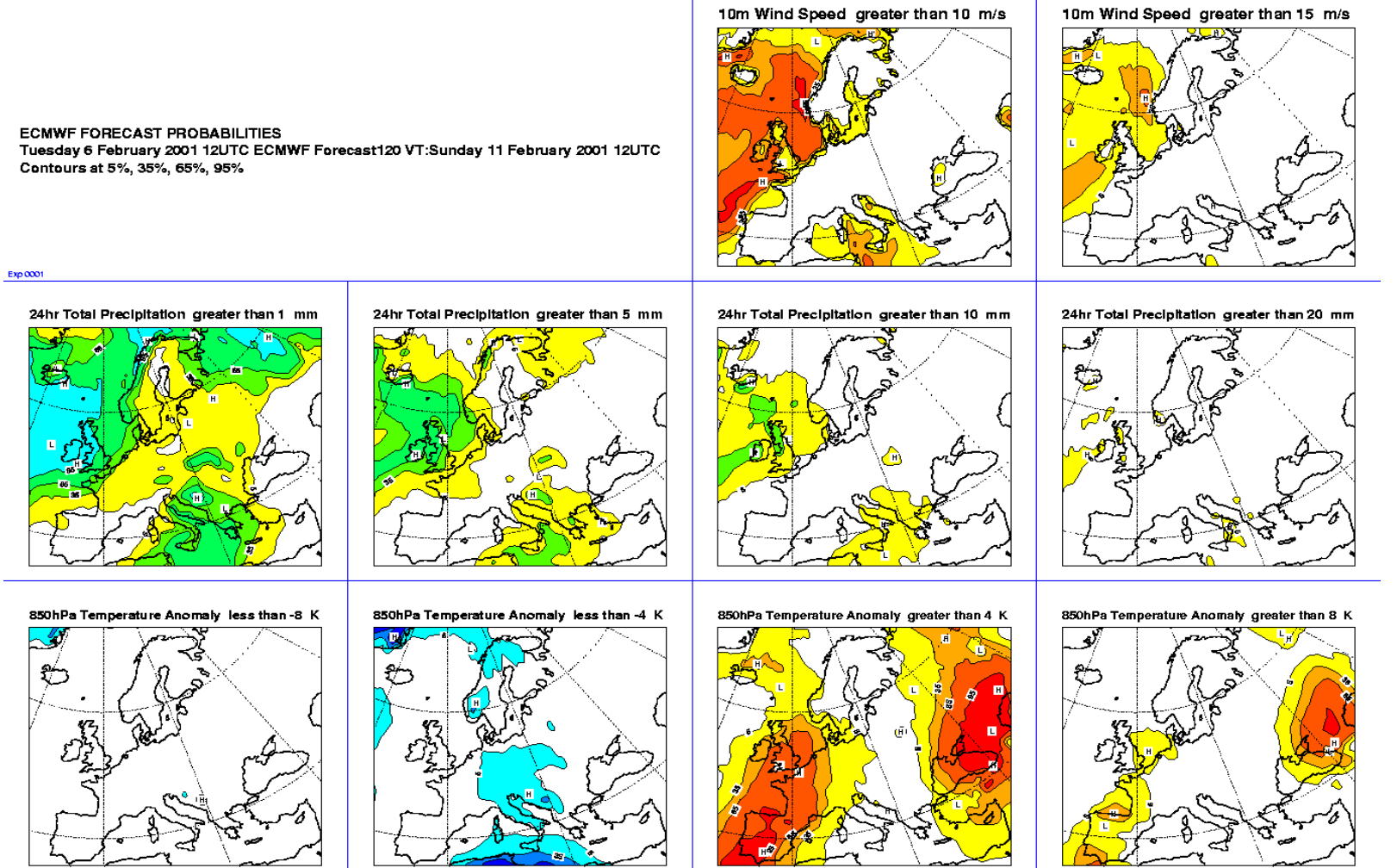


# The EPS contains a lot of information, spread out in several locations

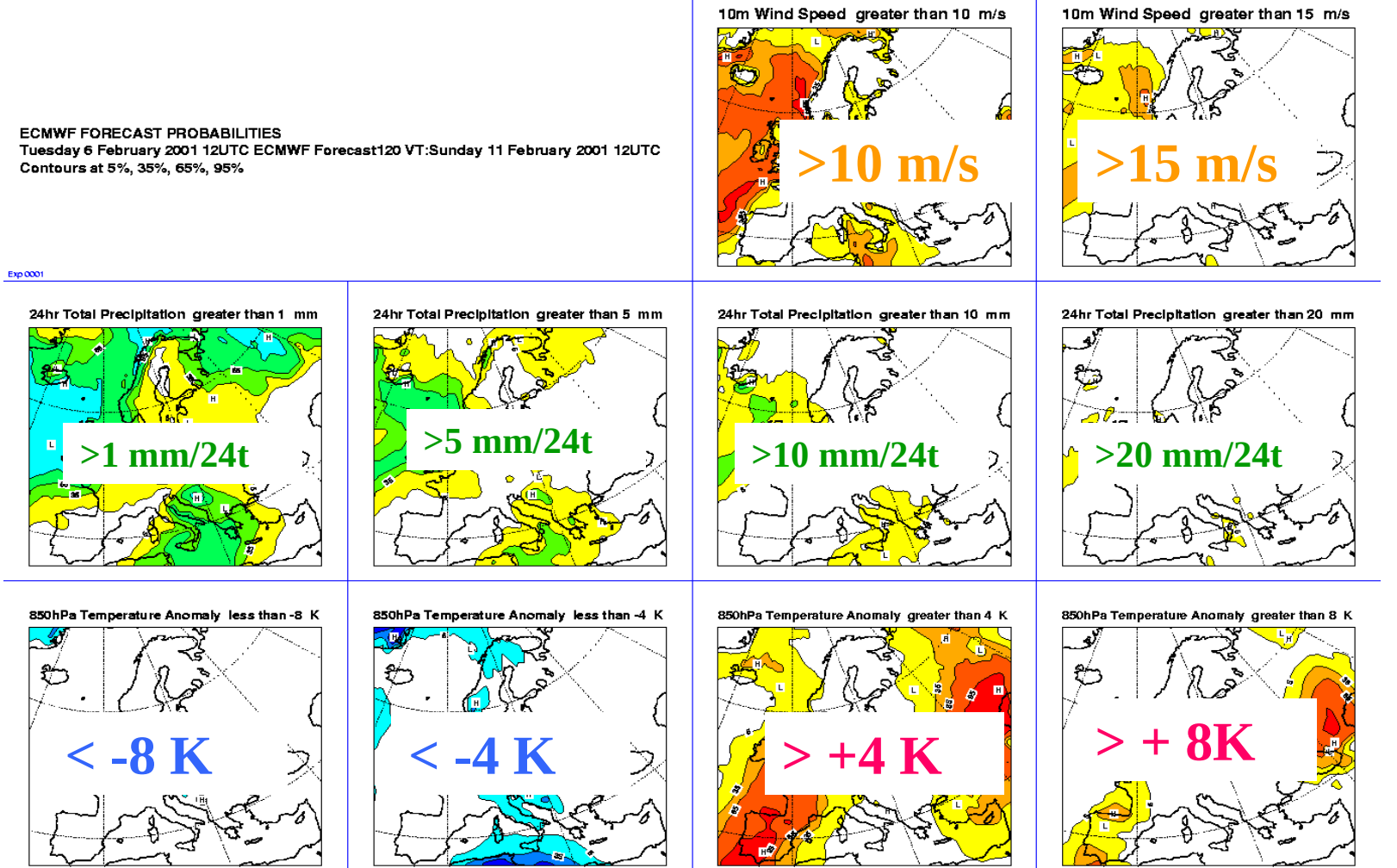
One solution is to create ”synoptic” EPS maps in the old Bergen School tradition: *combine several parameters in the same map!*



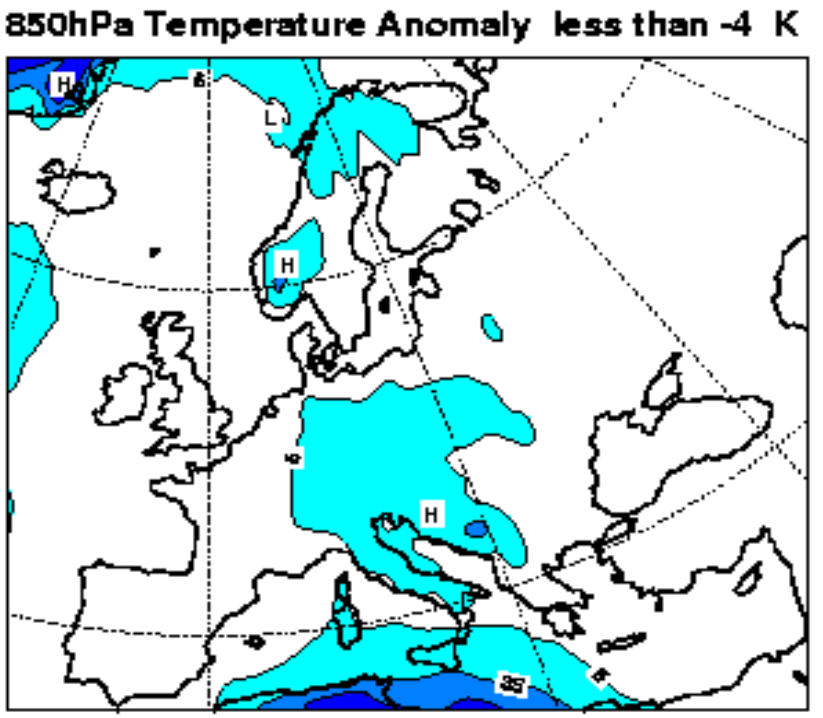
# Probability of strong winds, precipitation and anomalous (850 hPa) temperatures 6 February 2001 D+5

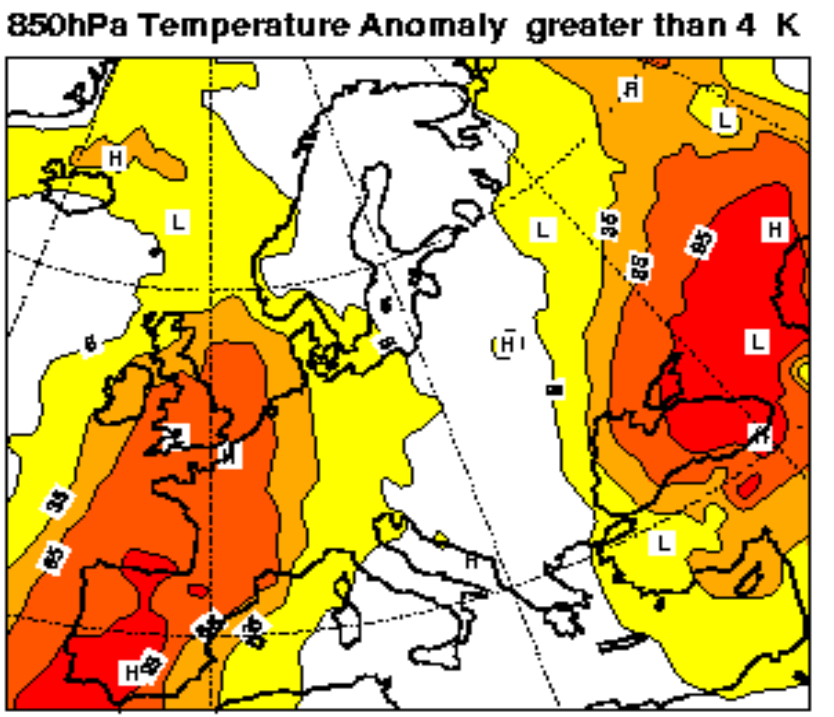


# Probability of strong winds, precipitation and anomalous (850 hPa) temperatures 6 February 2001 D+5



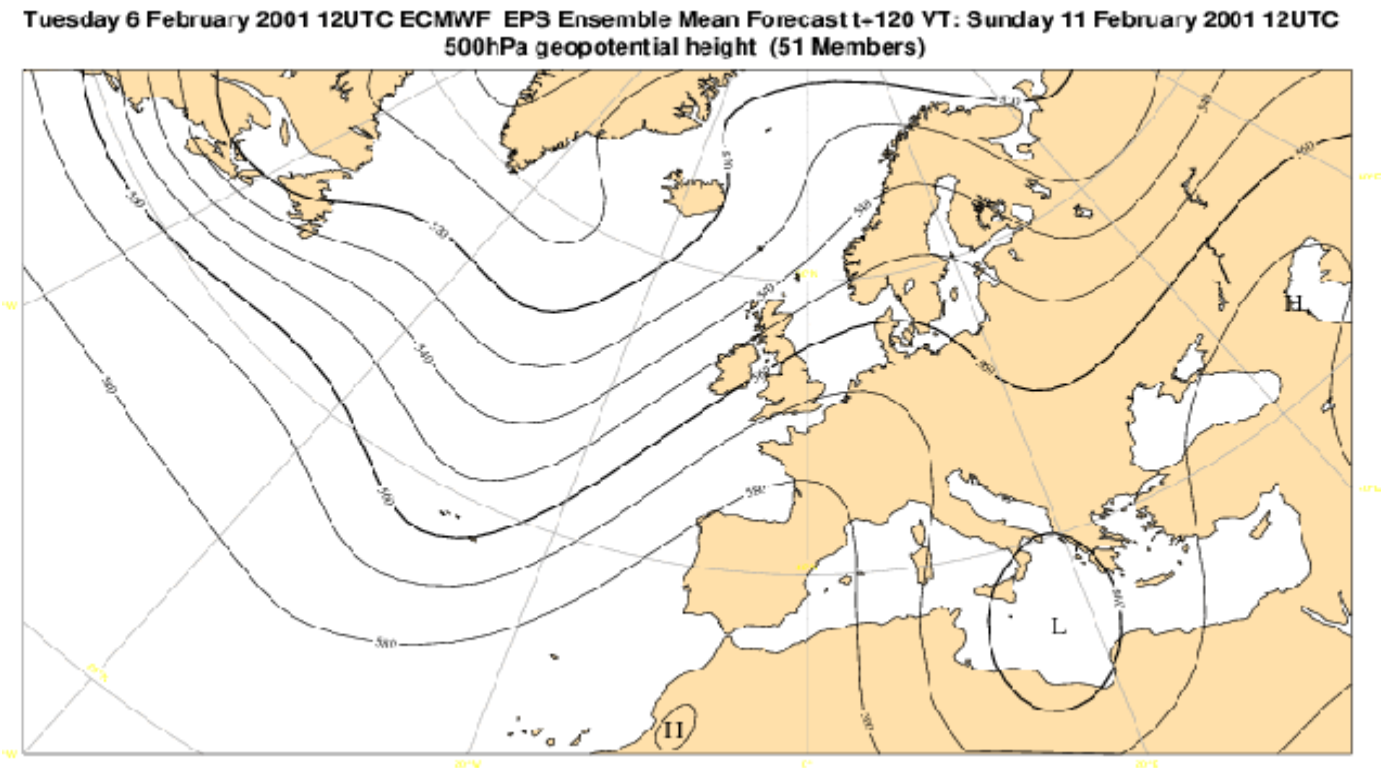
Probabilities that the 850 hPa temperature anomalies are >- 4 K or > + 4 K



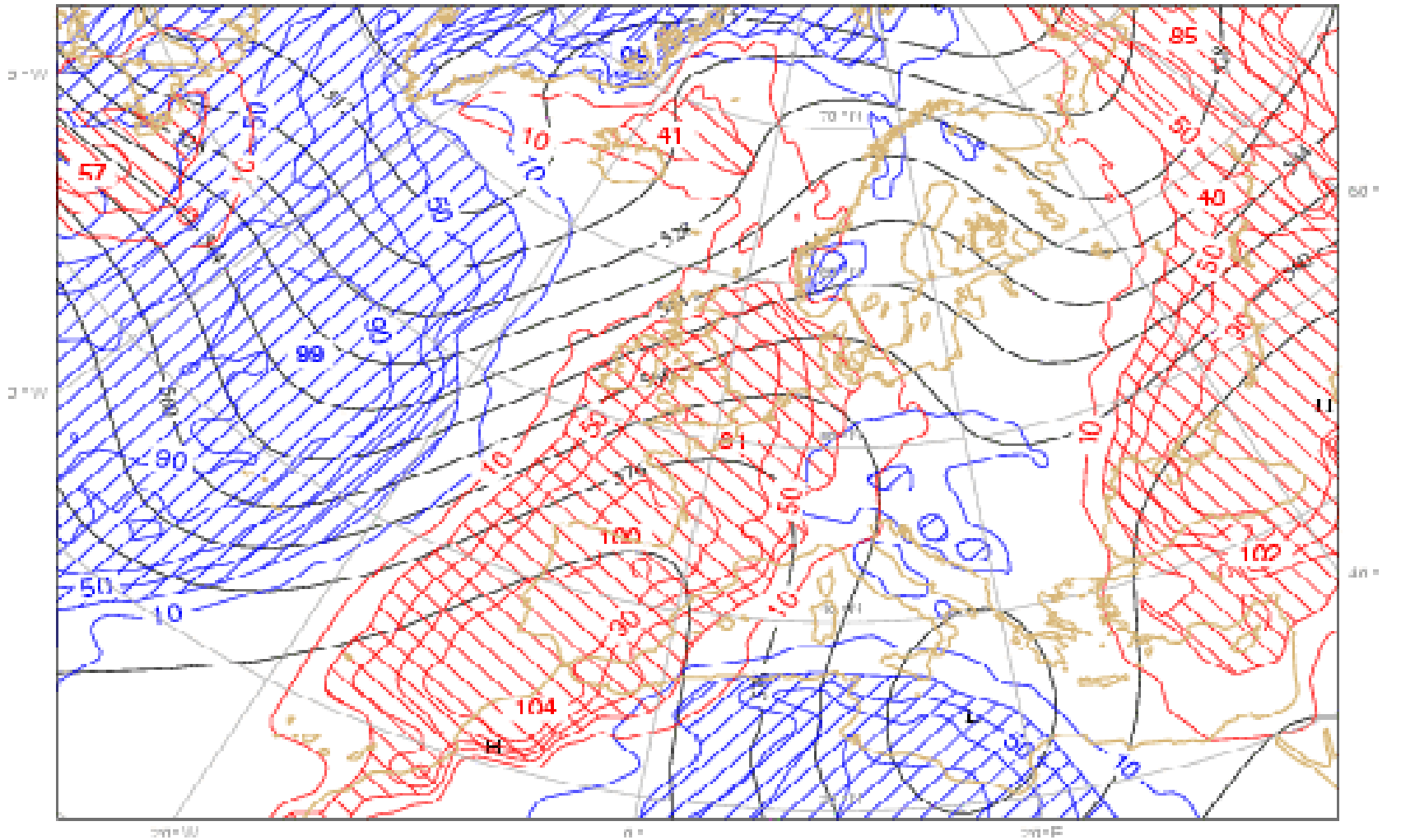




**…and combine them with the ensemble mean map...**

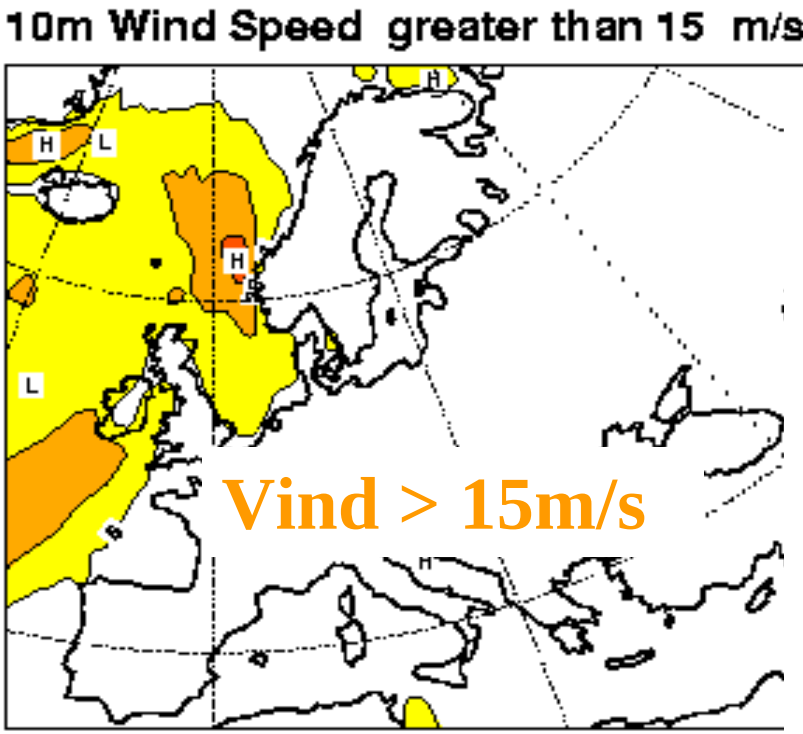
# …500 hPa 6 February 2001 +120 h

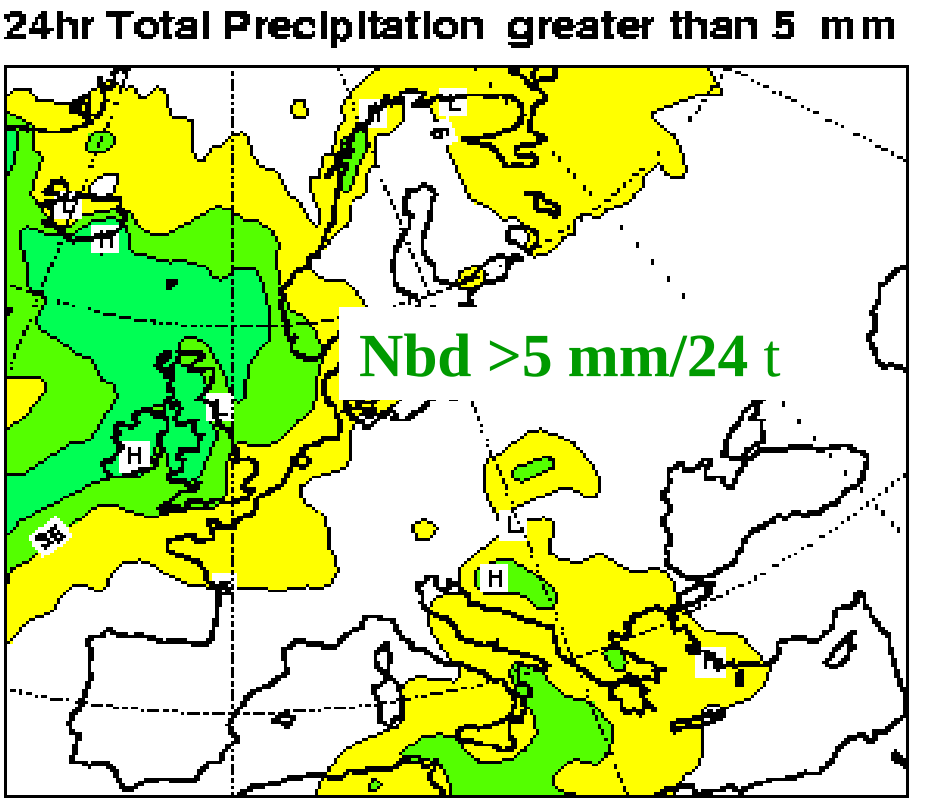


# …850 hPa temperature anomalies overlaid the ensemble mean 500 hPa 6 February 2001 + 120 h

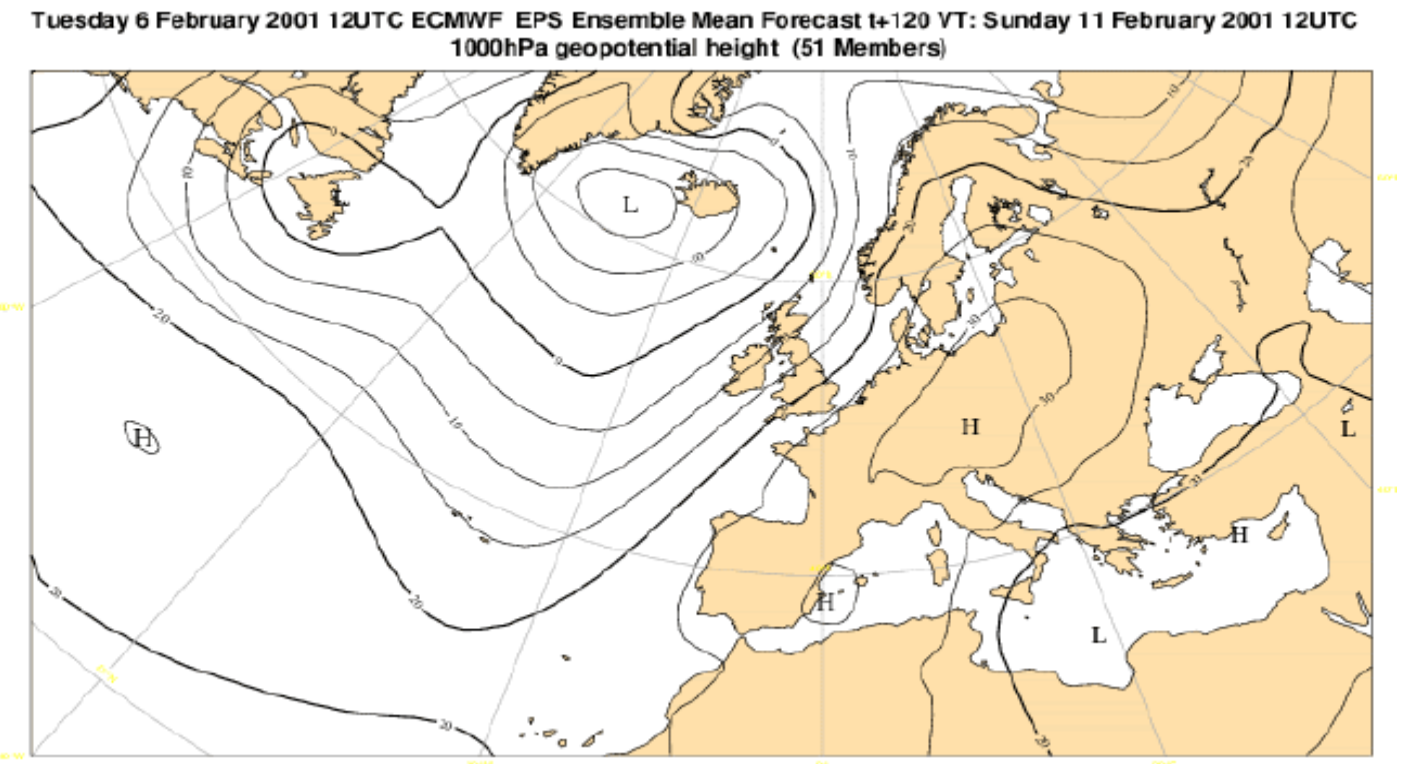


# The same for gale force winds and precipitation according to EPS from 6 February 2001 + 120 h

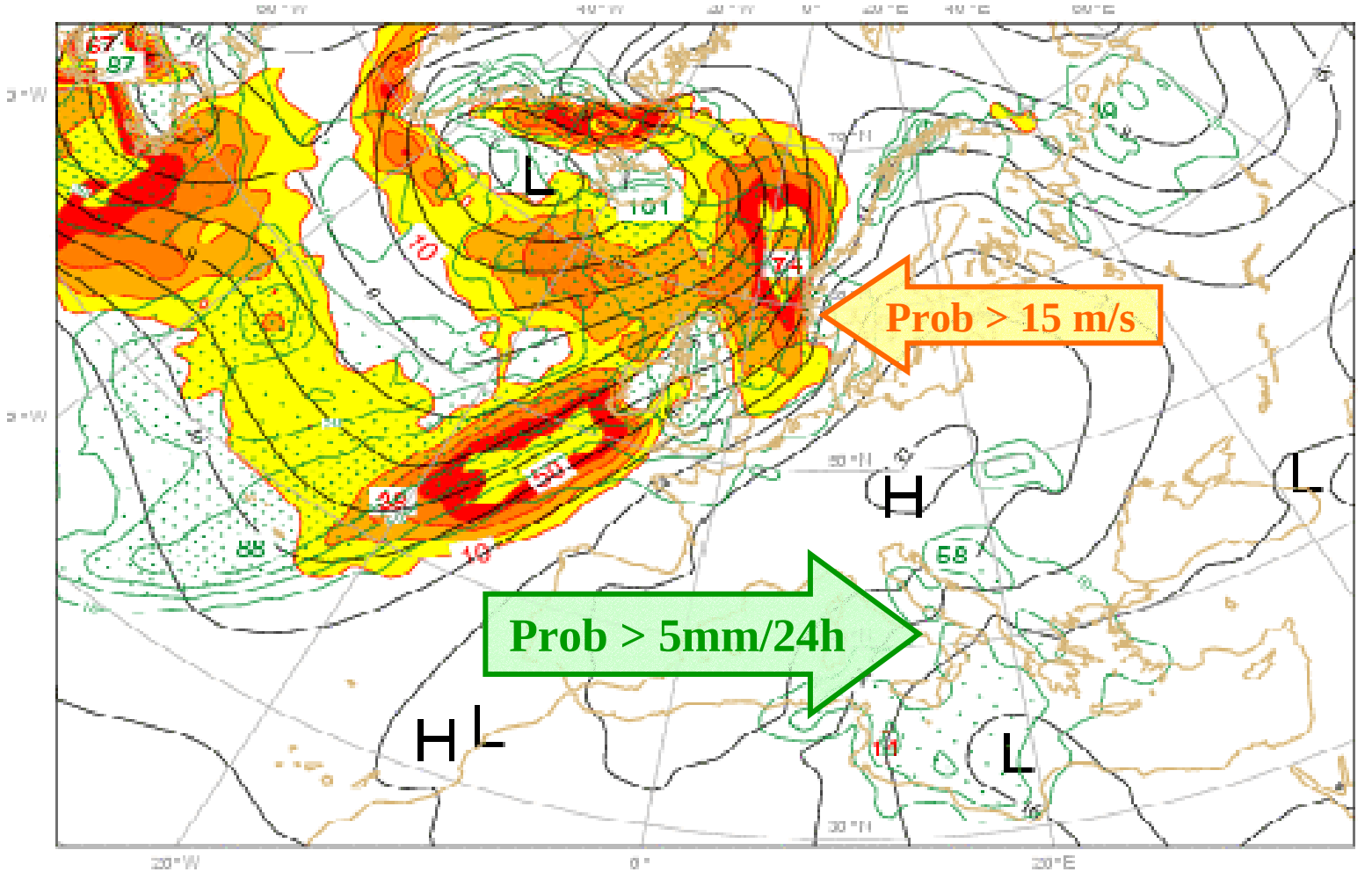


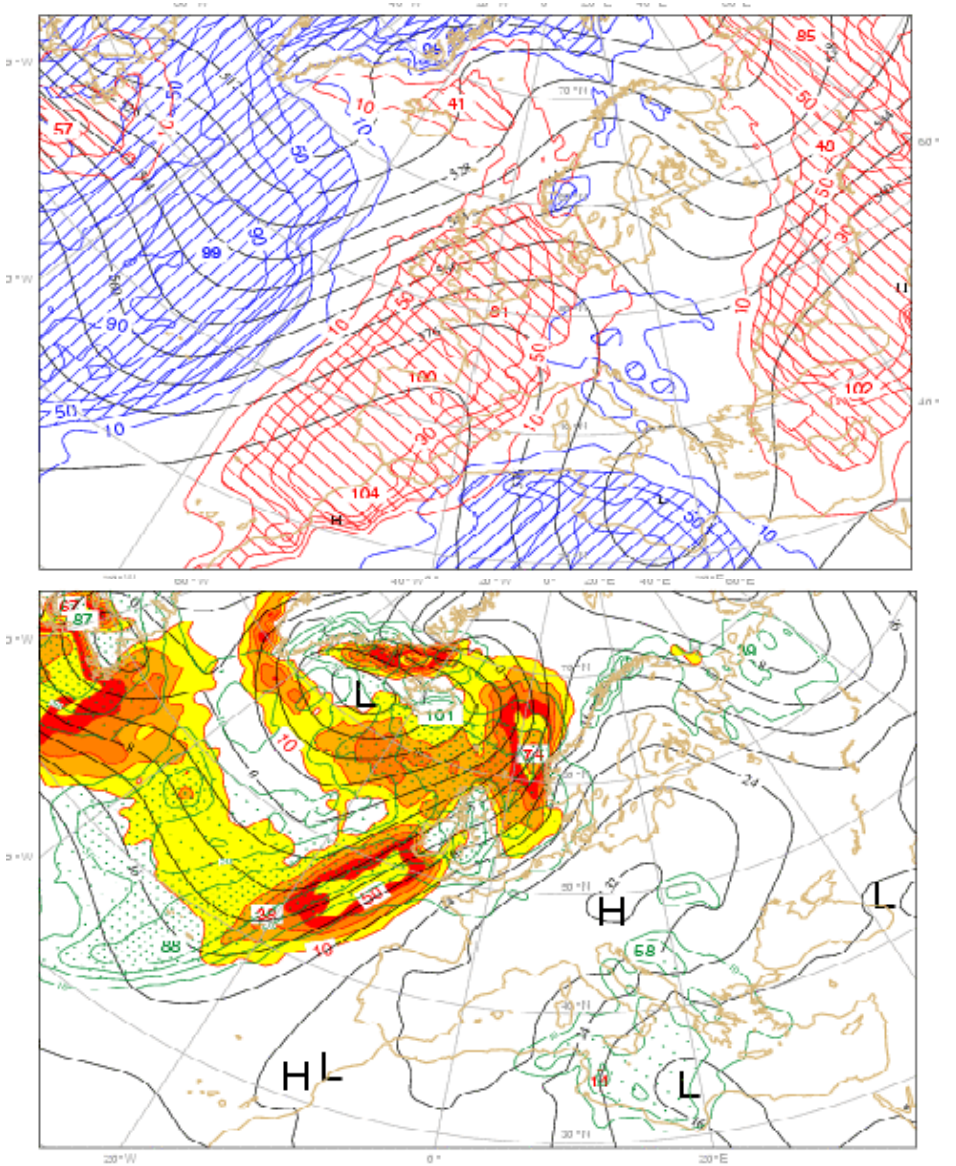



# … in combination with the ensemble mean forecast



# … 6 February 2001 12 UTC + 120 h

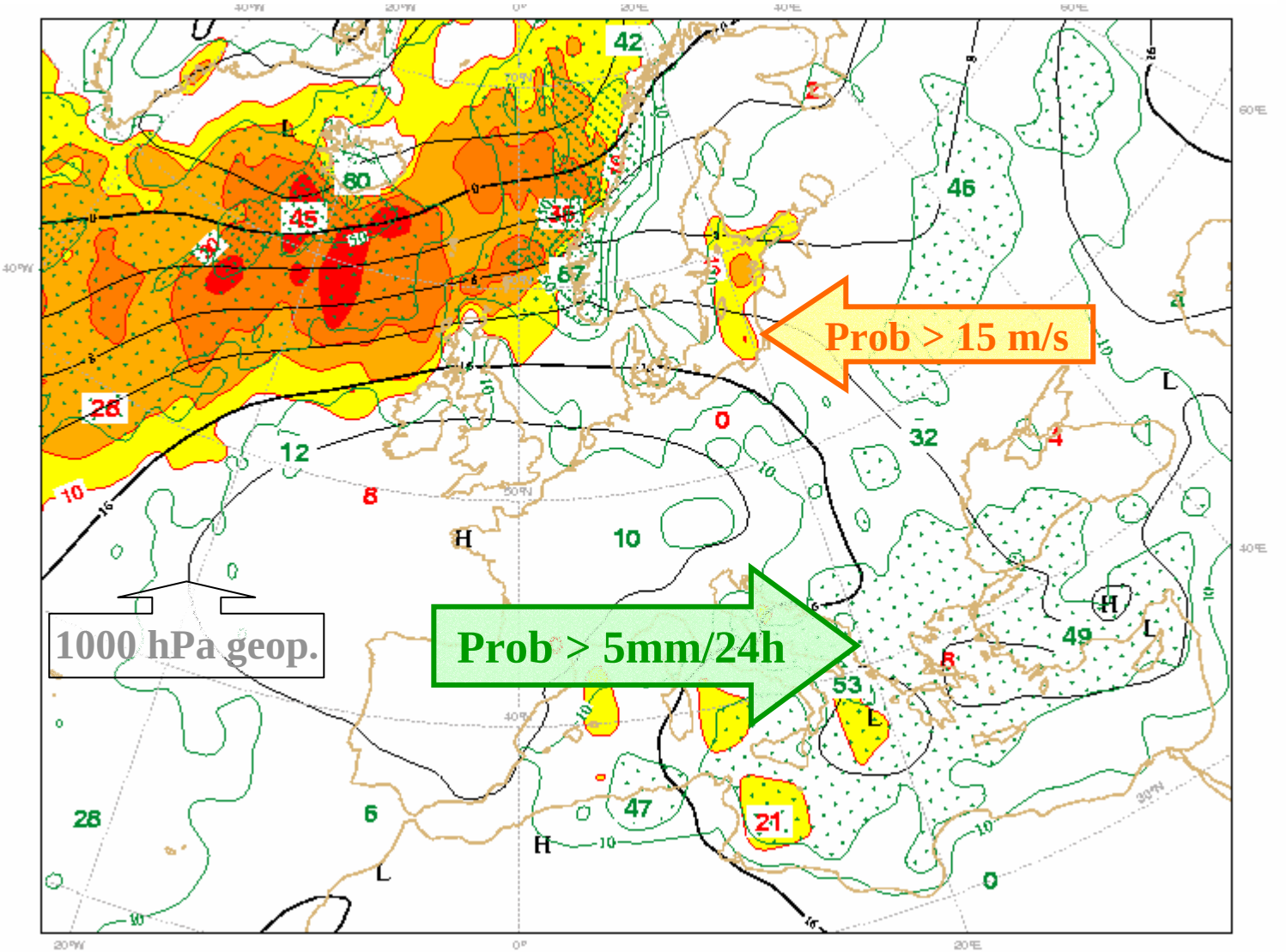




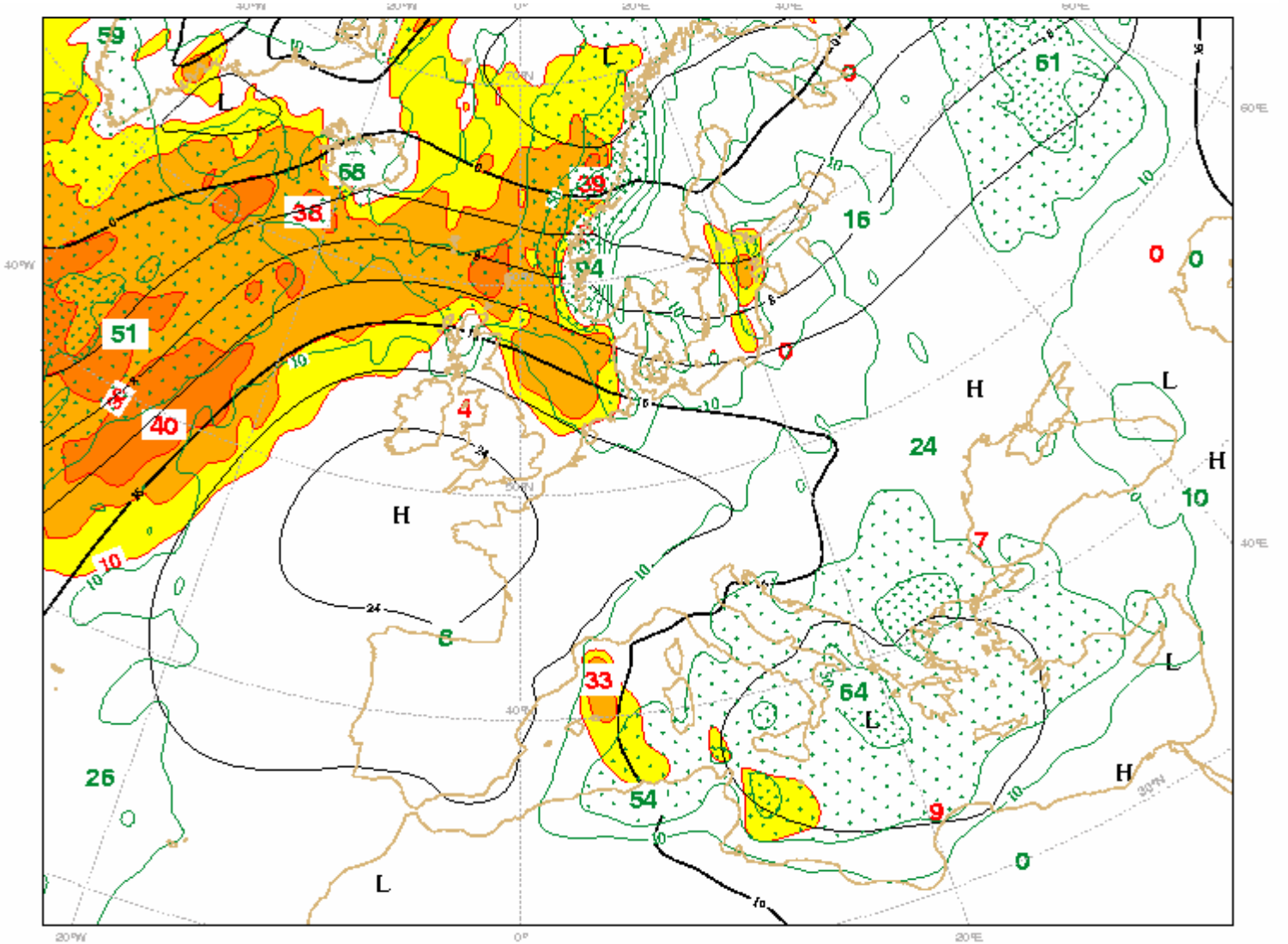
**The Synoptic EPS Map combines the mean and the spread**

**The following maps show a situation from 11 November 2001, which contained disturbances both in the north and south**

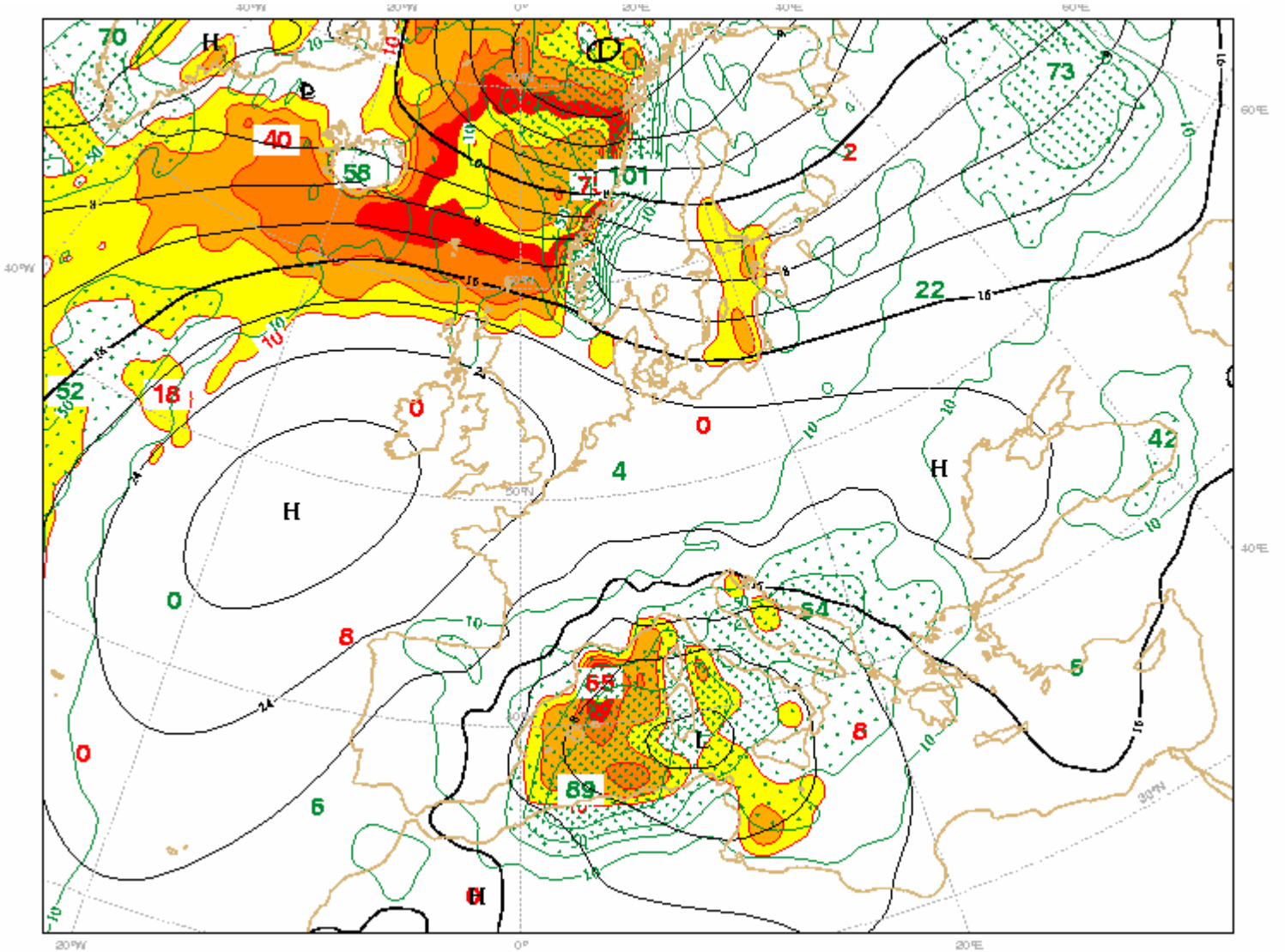
# EPS forecast 4 Nov 2001 D+7 verifying 11 November



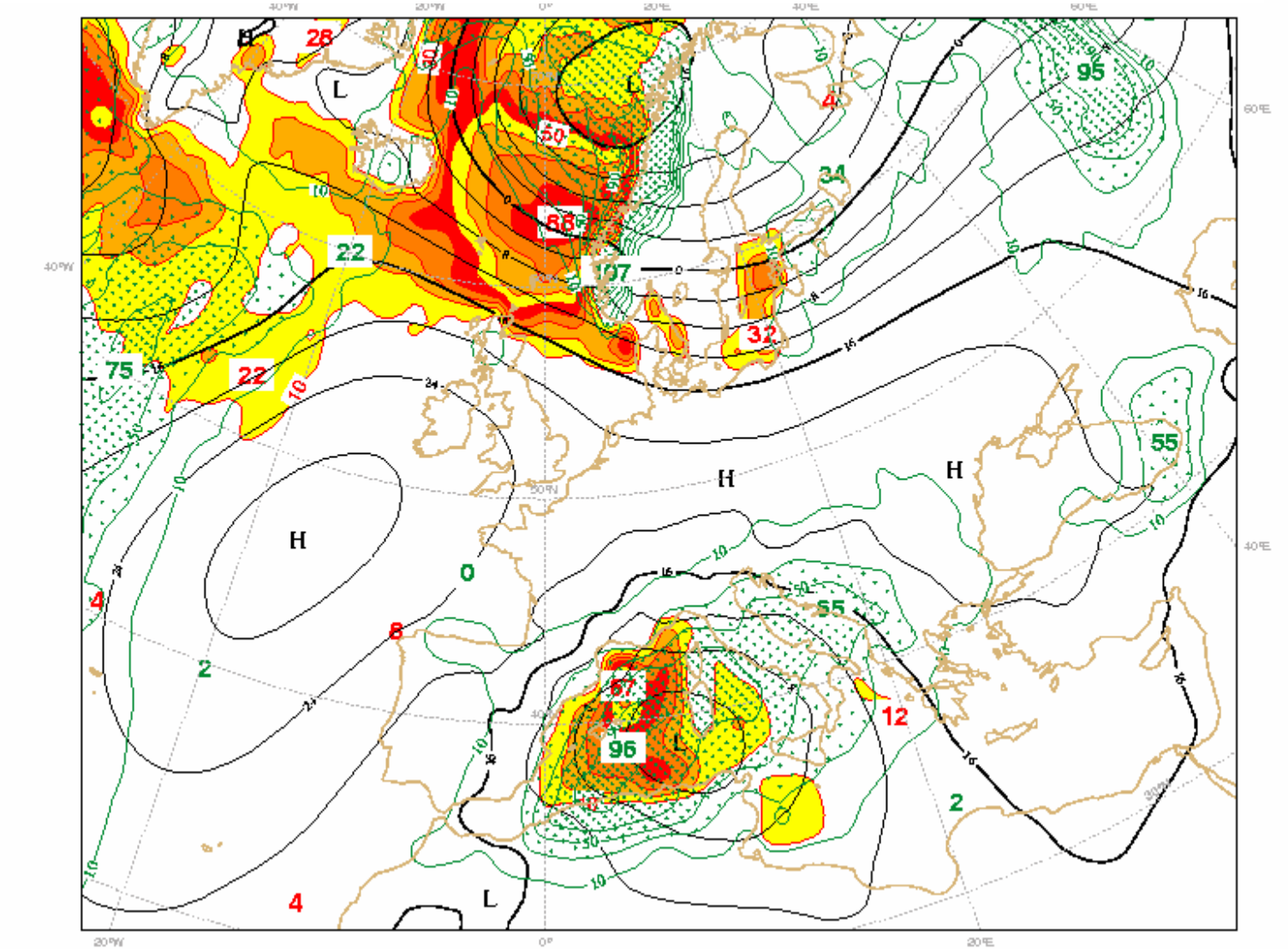
# EPS forecast 5 Nov 2001 D+6 verifying 11 November



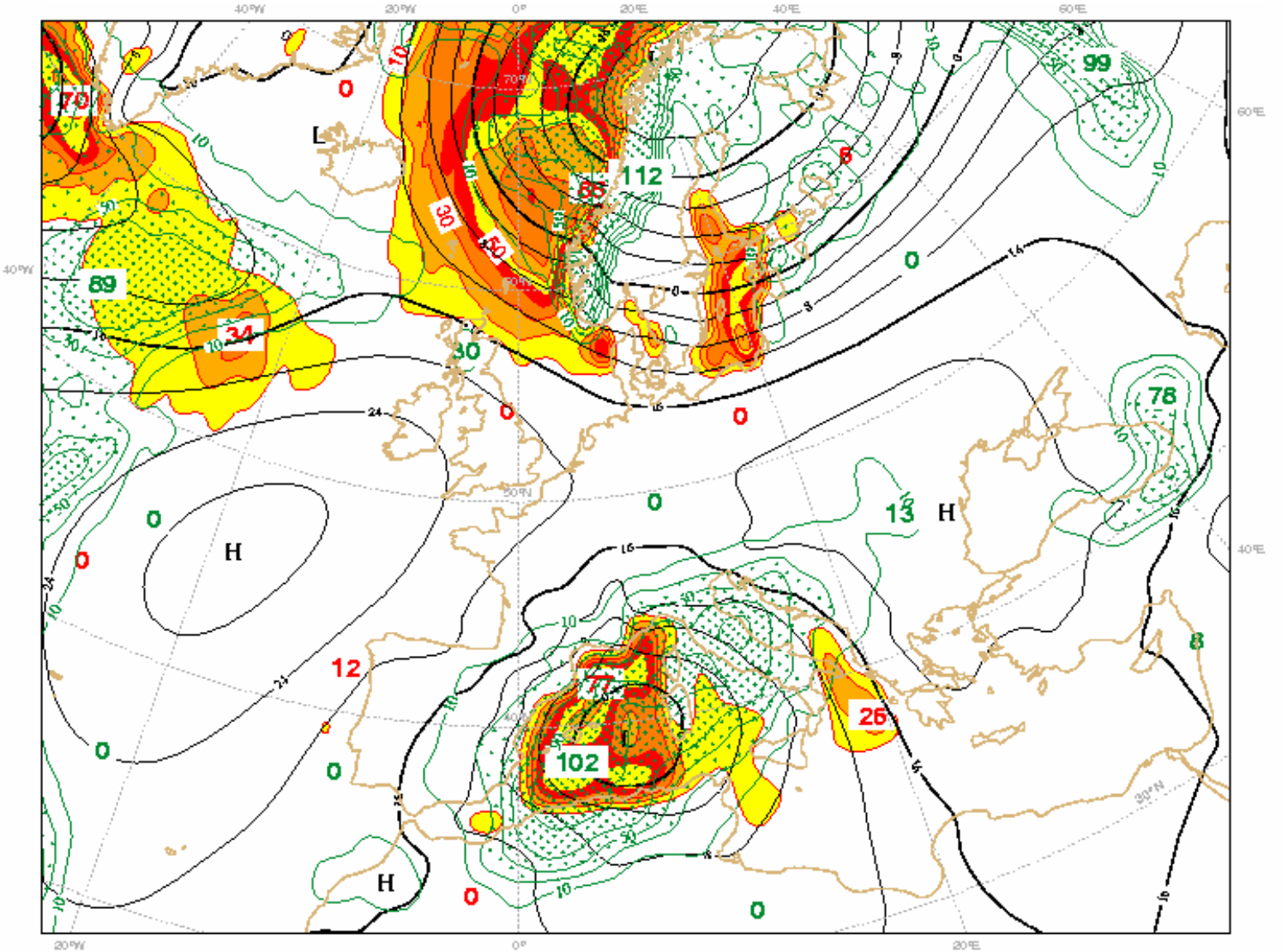
# EPS forecast 6 Nov 2001 D+5 verifying 11 November



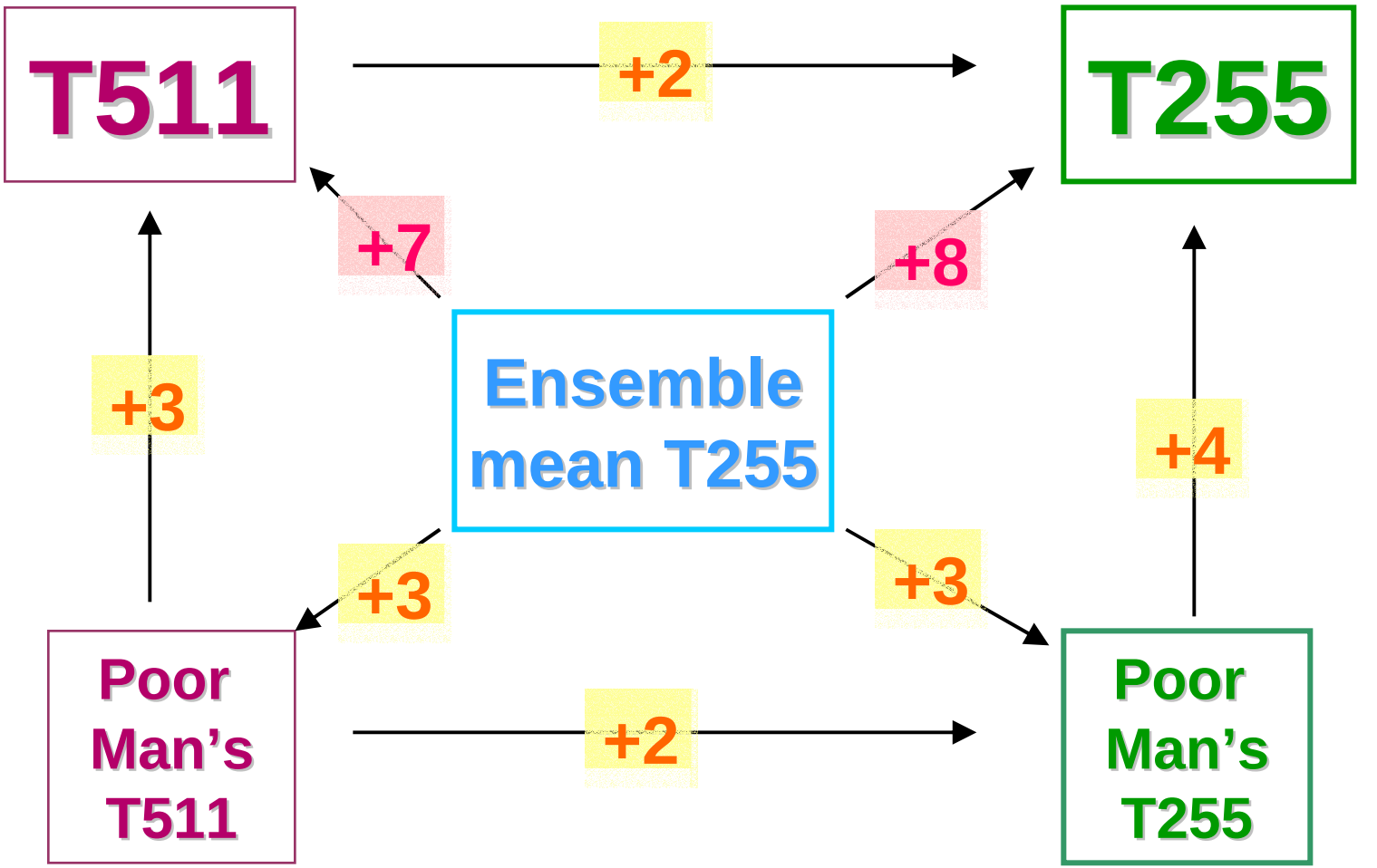
# EPS forecast 7 Nov 2001 D+4 verifying 11 November



# EPS forecast 8 Nov 2001 D+3 verifying 11 November



# The increase in anomaly correlation coefficient of the EM of 2 m-temperature D+7 forecasts compared with different other deterministic methods (9 European stations)



# T511 forecasts compared with T255 ensemble mean values

## Advantages with T511

**The best simulation of meteorological features**

**The best forecasts 1-2 days ahead, but then worse**

**Jumpy forecasts after 3-4 days**

**Good geographical resolution**

## The same for T255 ensemble mean values

**Occasionally a bit unrealistic flow patterns**

**The best categorical forecast values beyond 2 days**

**Drastically reduced jumpiness**

**The ensemble mean can be supplemented with probabilities, which is not possible with the T511**

# New types of products

**Tailor suited probabilities** (customer suited thresholds)

**Combinations of different elements** (wind+temp, rain+cloud)

**Individual self-learning corrections** (kalman filtering on own data)

**Averages in time and space** (forecasts will improve if considered for longer intervals or/and larger areas)

**Poor Man's EPS on HIRLAM** (old forecasts still hold some value)

**Day-to-day probability thresholds** (today <20mm/6h, next day <10 mm/24h after the situation)