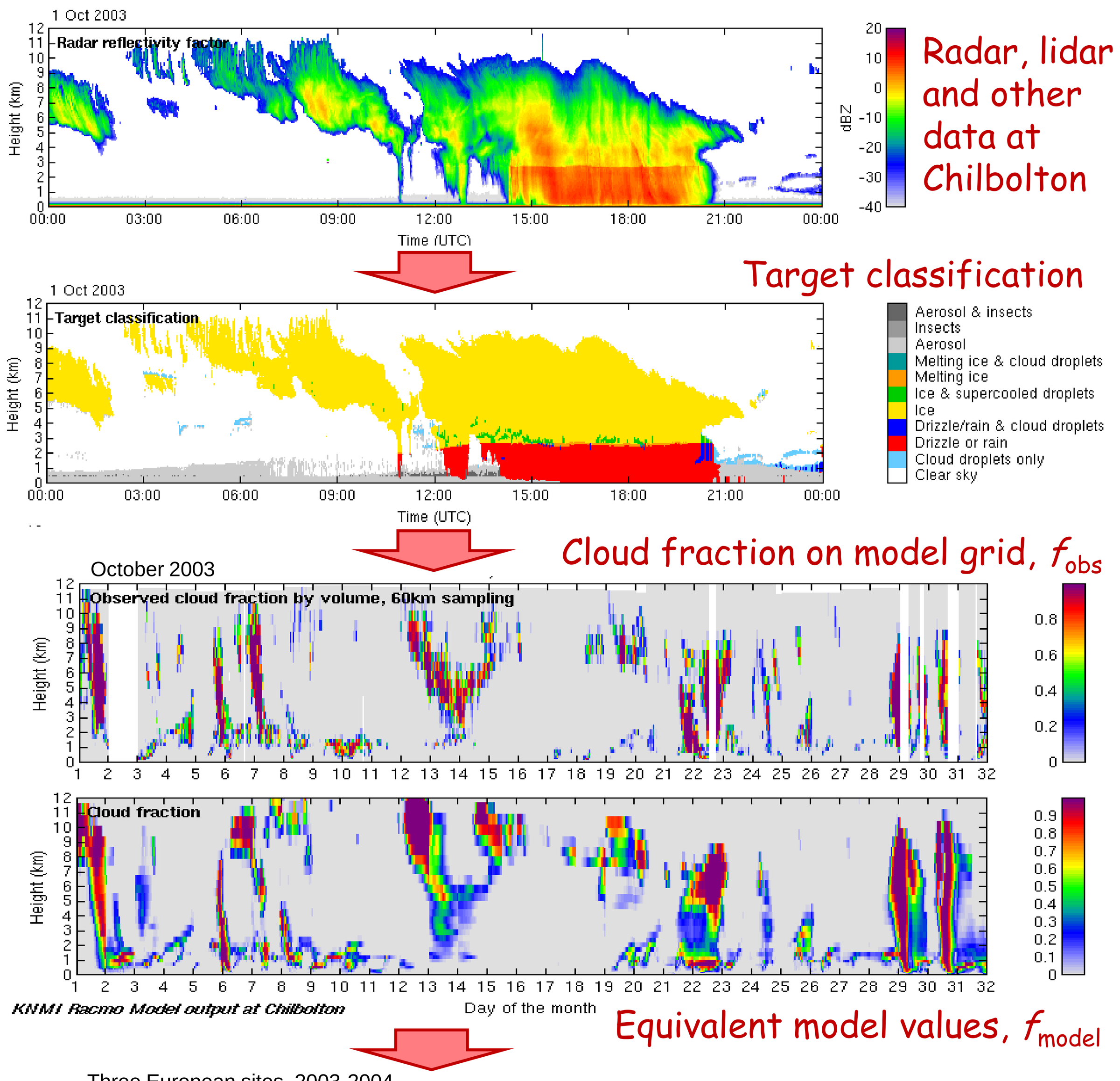
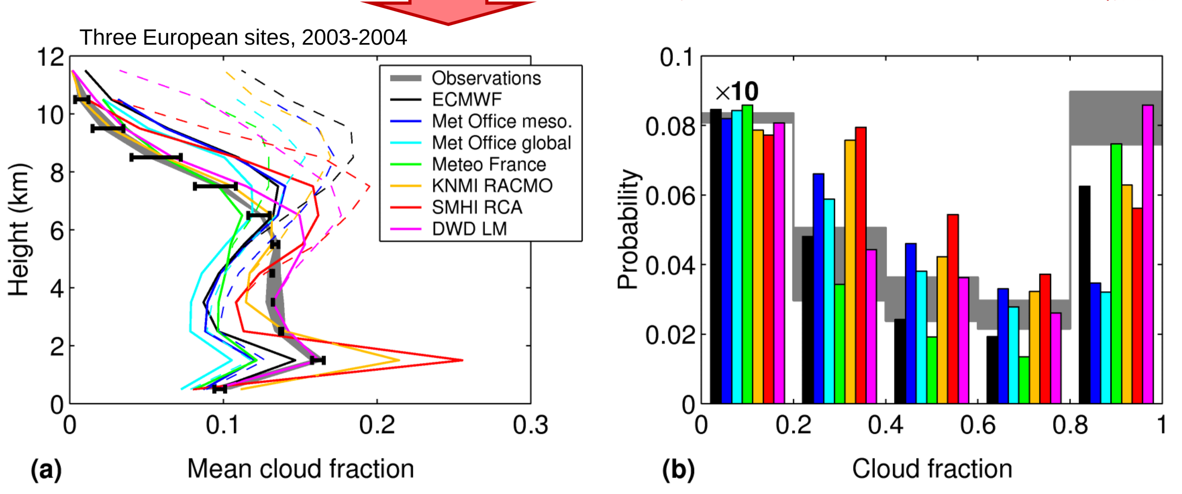
Evaluating model skill: what is the half-life of a cloud-fraction forecast?

Robin J. Hogan, Ewan J. O'Connor and Anthony J. Illingworth
University of Reading, UK *Email: r.j.hogan@reading.ac.uk*

1. What is Cloudnet?



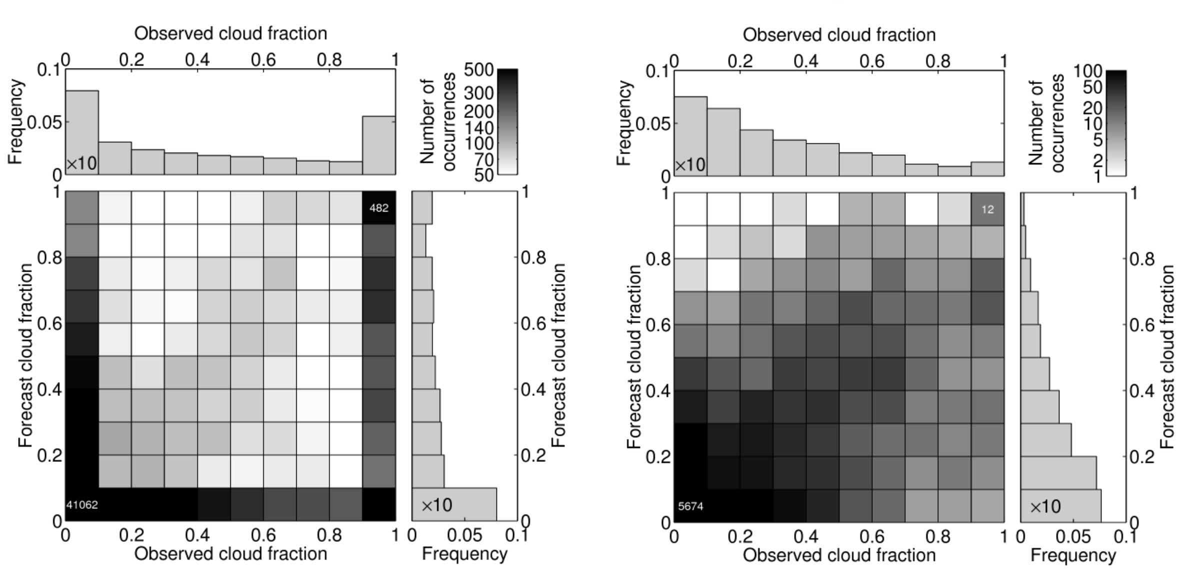


Model evaluation statistics reported by Illingworth et al. (BAMS 2007); water content statistics also available

But this only tests model climatology, how can we test whether clouds were forecast in the right place?

2. Joint probability distributions

Consider "DWD-EU" 7-km model over Murgtal in 2007:

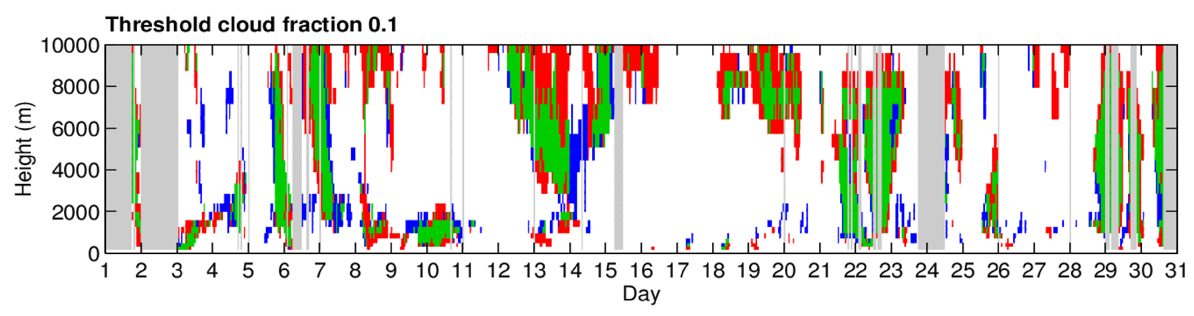


Degree of association appears poor for raw cloud fraction

Better association apparent if average to 6 hours

Simplify information by defining a contingency table:

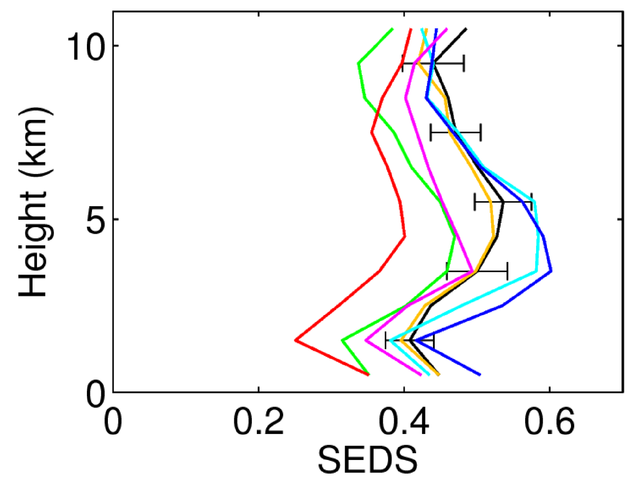
	Observed cloud ($f_{obs}>f_{thresh}$)	Observed clear-sky ($f_{obs}\le f_{thresh}$)
Forecast cloud ($f_{model}>f_{thresh}$)	*a*: Number of cloud hits	*b*: Number of false alarms
Forecast clear-sky ($f_{model}\le f_{thresh}$)	*c*: Number of misses	*d*: Number of clear-sky hits



5. Skill versus height

Mid-level clouds most skilfully forecast

- Surprising: physics of mixed-phase clouds not represented well
- Large-scale ascent has largest amplitude in mid-troposphere so cloud response most strong here?



- Met Office performs best: arguably the most sophisticated microphysics with separate liquid and ice

Boundary layer clouds least skilfully forecast

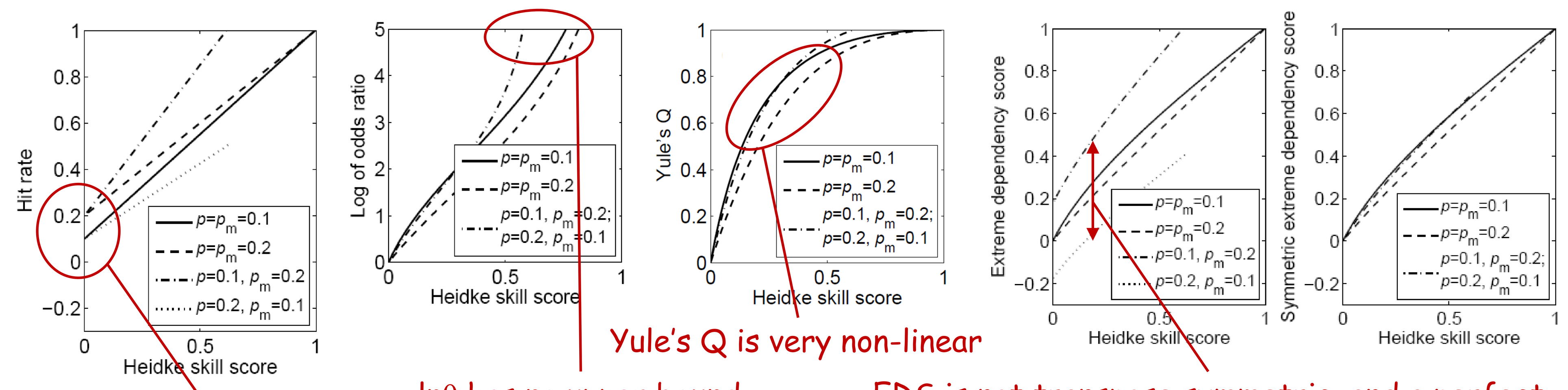
- Not a surprise: well-known forecasting problem
- Occurrence a subtle function of subsidence, surface fluxes, entrainment, stability, drizzle formation...

3. Desirable properties of a skill score

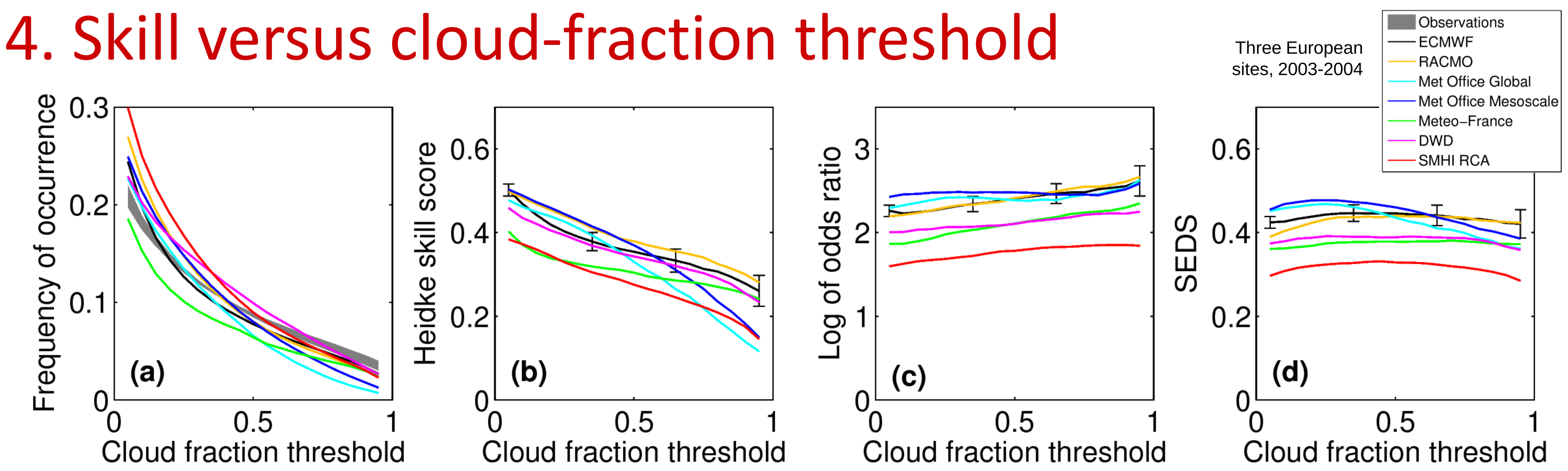
- ***Equitable:*** random forecasts have expected score of zero. This is essential!
- ***Transpose symmetric:*** no change if swap model and observations. Asymmetric scores tend to be *improper*; they can be *hedged* by over- or underestimating the frequency of occurrence.
- ***Uses full range of cloud fraction:*** better than assessing just when fraction exceeds a threshold.
- ***Useful for rare events:*** Most scores tend to meaningless limit as frequency of occurrence → 0.
- ***Linear:*** To calculate a half-life, score must depend on the inputs in a reasonably linear fashion.

Score	Equitable as $n \to 0$	Transpose symmetric	Uses full range	Useful for rare events	Linear	Definition and notes
Hit rate, *H* False alarm rate, FAR	N	N	N	N	Y	$H = a/(a+c)$ FAR $= b/(a+b)$ These and other non-equitable scores used by Mace et al. (1998) for cloud evaluation
Heidke Skill Score, HSS	Y	Y	N	N	Y	Define number of correct forecasts $x = a+d$, then define HSS to vary linearly between 0 for a random forecast and 1 for a perfect forecast using $\text{HSS} = (x-x_{random})/(x_{perfect}-x_{random})$
Log of Odds Ratio, lnθ	Y	Y	N	~	~	$\ln\theta = \ln(ad/bc)$ Analyzed by Stephenson (2000); property that a perfect forecast scores infinity
Yule's Q (also known as Odds Ratio Skill Score)	Y	Y	N	N	N	$Q = (ad-bc)/(ad+bc) = (\theta-1)/(\theta+1)$ Equivalent to lnθ, but bounded to 0-1 at the expense of being very non-linear
Mean Absolute Error Skill Score, MAESS	Y	Y	Y	N	Y	As HSS but with $x = \Sigma\lvert f_{model}-f_{obs}\rvert/n$
Extreme Dependency Score, EDS	N	N	N	Y	~	$\text{EDS} = 2\ln[(a+c)/n]/\ln(a/n) - 1$, where $n = a+b+c+d$ Shown by Stephenson et al. (2008) to tend to a meaningful limit for rare events
Symmetric Extreme Dependency Score, SEDS	Y	Y	N	Y	~	$\text{SEDS} = \{\ln[(a+b)/n]+\ln[(a+c)/n]\}/\ln(a/n)-1 = \ln(a_r/a)/\ln(a/n)$, where a_r is expected a for random forecast. Desirable properties of EDS plus transpose symmetry & equitability

Plot scores against HSS for particular frequencies of occurrence in observations, p, and model, p_m:



4. Skill versus cloud-fraction threshold



Frequency of $f > f_{thresh}$ (also known as base rate) decreases as f_{thresh} is increased

HSS indicates *less* skill for high cloud fractions, but HSS tends to zero for rare events

lnθ indicates *more* skill for high cloud fractions, but lnθ tends to infinity for rare events!

Symmetric Extreme Dependency Score reliable at low base rate and shows no significant trend

6. Estimating forecast "half life"

SEDS over first 1.5 days fitted by an inverse exponential

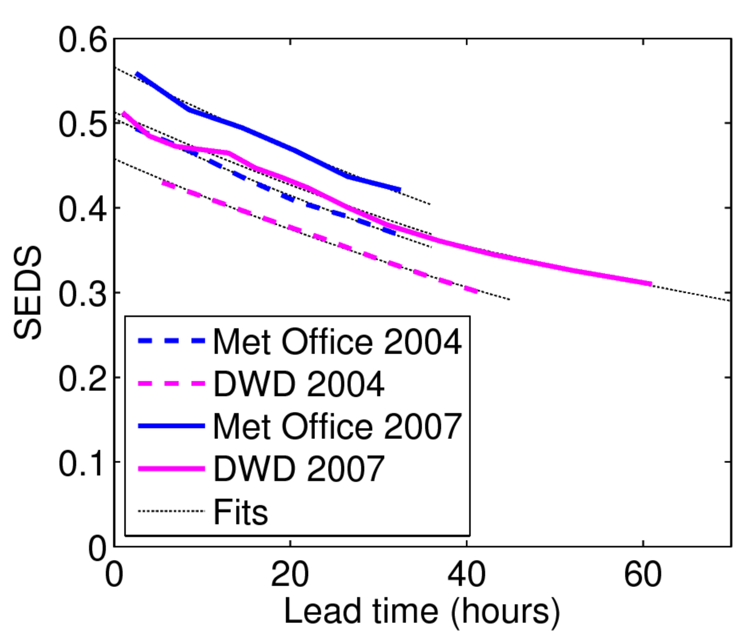
- DWD half-life **2.87 days** in 2004 and **3.15 days** in 2007
- Met Office half-life **2.91 days** in 2004 and **3.07** in 2007

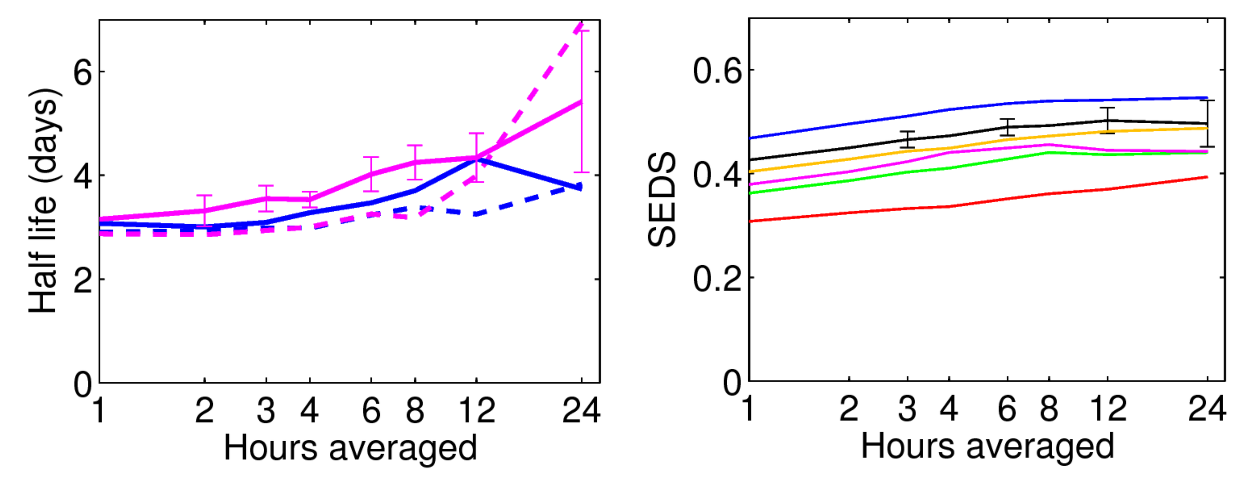
DWD forecasts available out to 3 days

- DWD half-life **4.31 days** for 1.5-3 day forecasts
- Forecast skill at short range dominated by convective timescales, at long range by large-scale weather systems

*Half-life for ECMWF 500-mb geopotential height is **9 days***

- Clouds less predictable than pressure field





Temporal averaging

- Absolute skill and half-life increase with temporal averaging
- Larger-scale features more predictable

Full results presented by Hogan et al. (2009)

References

Hogan, R. J., E. J. O'Connor and A. J. Illingworth, 2009: Verification of cloud-fraction forecasts. *Q. J. R. Meteorol. Soc.,* in revision.
Illingworth, A. J., R. J. Hogan, E. J. O'Connor, D. Bouniol, M. E. Brooks, J. Delanoe, D. P. Donovan, J. D. Eastment, N. Gaussiat, J. W. F. Goddard, M. Haeffelin, H. Klein Baltink, O. A. Krasnov, J. Pelon, J.-M. Piriou, A. Protat, H. W. J. Russchenberg, A. Seifert, A. M. Tompkins, G.-J. van Zadelhoff, F. Vinit, U. Willen, D. R. Wilson and C. L. Wrench, 2007: Cloudnet—continuous evaluation of cloud profiles in seven operational models using ground-based observations. *Bull. Am. Meteorol. Soc.,* ***88,*** 883–898.
Mace, G. G., C. Jakob, and K. P. Moran, 1998: Validation of hydrometeor occurrence predicted by the ECMWF model using millimeter wave radar data. *Geophys. Res. Lett.,* ***25,*** 1645–1648.
Stephenson, D. B., 2000: Use of the "Odds Ratio" for diagnosing forecast skill. *Weath. Forecasting,* ***15,*** 221–232.
Stephenson, D. B., B. Casati, C. A. T. Ferro and C. A.Wilson, 2008: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorol. Apps.,* ***15,*** 41–50.