QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **135**: 1494–1511 (2009)
Published online 10 August 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.481

Verification of cloud-fraction forecasts

Robin J. Hogan* Ewan J. O'Connor and Anthony J. Illingworth
Department of Meteorology, University of Reading, UK

ABSTRACT: Cloud radar and lidar can be used to evaluate the skill of numerical weather prediction models in forecasting the timing and placement of clouds, but care must be taken in choosing the appropriate metric of skill to use due to the non-Gaussian nature of cloud-fraction distributions. We compare the properties of a number of different verification measures and conclude that of existing measures the *Log of Odds Ratio* is the most suitable for cloud fraction. We also propose a new measure, the *Symmetric Extreme Dependency Score,* which has very attractive properties, being equitable (for large samples), difficult to hedge and independent of the frequency of occurrence of the quantity being verified. We then use data from five European ground-based sites and seven forecast models, processed using the 'Cloudnet' analysis system, to investigate the dependence of forecast skill on cloud fraction threshold (for binary skill scores), height, horizontal scale and (for the Met Office and German Weather Service models) forecast lead time. The models are found to be least skillful at predicting the timing and placement of boundary-layer clouds and most skilful at predicting mid-level clouds, although in the latter case they tend to underestimate mean cloud fraction when cloud is present. It is found that skill decreases approximately inverse-exponentially with forecast lead time, enabling a forecast 'half-life' to be estimated. When considering the skill of instantaneous model snapshots, we find typical values ranging between 2.5 and 4.5 days.

KEY WORDS skill scores; cloud radar; Cloudnet

Received 22 January 2009; Revised 22 June 2009; Accepted 25 June 2009

1. Introduction

A combination of better representation of physical processes, improved data assimilation and higher resolution has led to a notable increase in the skill of weather forecasts over the last few decades. A compelling demonstration of the improvement of the model of the European Centre for Medium Range Weather Forecasts (ECMWF) was presented by Simmons and Hollingsworth (2002), who examined the correlation of anomalies from climatology between analysed and forecast 500 hPa geopotential height for a number of forecast lead times. From their Northern Hemisphere 7-day anomaly correlations of 0.45 in 1980 and 0.58 in 2000, an assumed inverse-exponential decay results in a forecast 'half-life' (the lead time at which the anomaly correlation falls to 0.5) of around 6 days in 1980 and around 9 days in 2000. However, the predictability of quantities that are important to the general public (rainfall, surface temperature and cloud cover) is likely to be less, given the faster error growth of the small scales at which these phenomena exhibit structure (Lorenz, 1969; Mass *et al.*, 2002; Roberts, 2008).

In this paper we investigate objective ways in which the skill of vertically resolved cloud forecasts can be assessed using cloud radar and lidar observations, and ultimately how the forecast half-life can be estimated. We focus specifically on cloud fraction, defined as the volume-fraction of a forecast model grid-box that contains cloud, which is prognostic in some models (Tiedtke, 1993; Wilson *et al.*, 2008). Radar and lidar have received growing attention for evaluating clouds fraction in forecast models. Hogan *et al.* (2001) evaluated the cloud-fraction climatology of the ECMWF model using ground-based observations from a single site, which was extended to seven models and three sites by Illingworth *et al.* (2007).

In terms of verification of specific cloud forecasts, rather than evaluating just the model cloud climatology, cloud fraction presents an interesting challenge because its distribution is U-shaped rather than Gaussian (e.g. Hogan *et al.*, 2001). This means that traditional measures of error, such as the root-mean-squared difference from observations, can be misleading. Mace *et al.* (1998) were the first to use *skill scores* to evaluate the skill of the model in predicting cloud to occur at the right time. This methodology has recently been applied to spaceborne lidar data to evaluate the skill of global cloud forecasts (Miller *et al.*, 1999; Palm *et al.*, 2005; Wilkinson *et al.*, 2008). The approach has been essentially to build a contingency table, as shown in Table I, and then to define a skill score as a function of the four elements a–d.

However, there are a plethora of verification scores in the literature, most of which quantify the skill of a set of forecasts by a single number, which is difficult to interpret for a single model. Usually one would only use them in a relative sense, i.e. to compare the skill of two or more models (or the same model but with different forecast lead

*Correspondence to: Dr Robin J. Hogan, Department of Meteorology, The University of Reading, Earley Gate, PO Box 243, Reading RG6 6BB, UK. E-mail: r.j.hogan@reading.ac.uk

Table I. Simple 2×2 contingency table expressing the joint occurrence of cloud fraction f greater than or less than some threshold value f_t in the observations and in a model forecast. The variables a–d represent the number of 'hits', 'false alarms', 'misses' and 'correct negatives', respectively.

	Observed $f_o > f_t$	Observed $f_o \leq f_t$
Forecast $f_m > f_t$	a	b
Forecast $f_m \leq f_t$	c	d

times, time of year, height, etc.), but a number of them have undesirable properties that make them unsuitable or misleading for use with cloud-fraction forecasts.

An alternative approach proposed by Jakob *et al.* (2004) is to treat cloud fraction in a model as a probabilistic forecast of cloud occurring at a particular instant in a model gridbox, which is then compared to the instantaneous cloud occurrence from radar and lidar. This has the advantage of not making the uncertain step (the 'ergodic assumption') of inferring cloud fraction in a three-dimensional volume from the time series of cloud occurrence derived from radar and lidar above a single site. However, it is necessary for different verification metrics to be used that are appropriate for probabilistic rather than categorical forecasts, which increases the complexity of the interpretation of the results for model development. In this paper we assume that cloud fraction can be estimated from observations with sufficient reliability, and therefore that retrieved cloud fraction may be used to evaluate the corresponding modelled values directly via the use of skill scores; this is supported by the results of Henderson and Pincus (2009).

The remainder of this paper is organized as follows. Section 2 describes the data used and the processing that has been applied to derive joint histograms of cloud fraction between the observations and each model. In section 3, we discuss the desirable attributes of a good verification score, with emphasis on the issues relevant to cloud fraction. Section 4 then uses these criteria to present an analysis of the merits and weaknesses of various skill scores that can be used for cloud fraction. Several new scores are also introduced. Section 5 then uses the best skill scores to examine the skill of cloud-fraction forecasts from a number of forecast models, including an estimate of the half-life of a cloud forecast.

2. Method

2.1. Cloudnet processing

We use ground-based radar and lidar data from a number of sites in Europe. The three sites used in the 'Cloudnet' project by Illingworth *et al.* (2007) (Chilbolton in the UK, Cabauw in the Netherlands and Palaiseau in France), are supplemented by data from Lindenberg in Germany. We also use data from the ARM (Atmospheric Radiation Measurement) Mobile Facility (AMF; Miller and Slingo, 2007), during its deployment at Murgtal in the Black Forest region of Germany during the Convective Orographic Precipitation Study (COPS) in 2007 (Wulfmeyer *et al.*, 2008).

The model data consist of hourly snapshots of cloud fraction in each model level above each site. For the 2003–2004 period, the seven models are as described by Illingworth *et al.* (2007). These are

(1) the global ECMWF model, which had a horizontal resolution of 39 km in this period,
(2) the Regional Atmospheric Climate Model (RACMO) of the Dutch Meteorological Institute (KNMI) which used the ECMWF physics package but ran with a horizontal resolution of 18 km,
(3) the global version of the Met Office Unified Model with a 60 km resolution,
(4) the mesoscale version of the same model with 12 km resolution,
(5) the global Météo-France 'ARPEGE' model with 24 km resolution over Europe,
(6) the German Weather Service (DWD) 'Lokal' model with 7 km resolution, and
(7) the Swedish Meteorological and Hydrological Institute (SMHI) Rossby-Centre Regional Atmospheric Model (RCA) with 44 km resolution.

For the 2007 data, the models used are the 12 km resolution North-Atlantic/European (NAE) version of the Met Office Unified Model, and the 7 km DWD 'COSMO-EU' model (the European-domain version of the model of the Consortium for Small-Scale Modelling). The Met Office and DWD models are of particular interest as they reported cloud fraction for different forecast lead times, and therefore the degradation of skill with lead time can be quantified.

The observational data were processed to obtain cloud fraction on the grid of each of the models using the Cloudnet processing system, exactly as described by Illingworth *et al.* (2007). Firstly, a 'target categorization' was performed, in which the radar and lidar data were used to classify the targets in each radar-lidar pixel (typically 30 s in time and 60 m in height) into a number of different categories (liquid cloud, ice cloud and snow, melting ice, rain, insects, aerosols and combinations thereof). Cloud was deemed to occur if the pixel contained liquid cloud, ice cloud or snow, noting that observationally there is a continuum between ice cloud and snow (e.g. Hogan *et al.*, 2001). Then the grid of each model was superimposed on the observed time–height sections of cloud occurrence, and observed cloud fraction was defined simply as the fraction of each gridbox containing cloud. The dimensions of the gridboxes in height were determined by the vertical model levels, while we used the model wind speed to calculate a height-dependent sampling time that would correspond to the horizontal resolution of the model. It is acknowledged that this provides an imperfect estimate of the true cloud fraction of the three-dimensional volume, but in practice other approaches (e.g. sampling for a fixed time) are found to produce very similar skill scores.

DOI: 10.1002/qj

When rain is present at the ground, significant radar attenuation can occur, particularly at higher radar frequencies (Hogan *et al.*, 2003). Therefore, when a rain rate is measured that is greater than 8 mm h^{-1} for a 35 GHz radar, and greater than 2 mm h^{-1} for a 94 GHz radar, that period is excluded from the comparison. During the periods studied in this paper, Cabauw and Lindenberg had 35 GHz radars, Chilbolton had a 94 GHz radar in 2004 and a 35 GHz radar in 2007, and all other all other sites used a 94 GHz radar.

For each model and site, the known sensitivity of the radar was then used to remove ice clouds in the model too tenuous to be detected. This was done as described by Illingworth *et al.* (2007) in a way that recognized the horizontal variability of ice water content within a gridbox. An example of one month of cloud fraction from the observations and two of the models is shown in Figure 1.

2.2. Joint histograms and contingency tables

Both modelled and observed cloud fraction were then averaged to a uniform 1 km vertical grid. The reason for this is that, in comparing the skill of one model to another, there is the danger that a model with a higher vertical resolution will perform worse simply because it is being tested more stringently than a lower-resolution model. The same argument applies to models with different horizontal resolutions, but this is addressed in the analysis when we calculate skill for different horizontal scales by averaging the cloud fraction in time.

For each model and site, we have more than a year of co-located observed and modelled cloud fraction versus height. To ease the subsequent processing, joint histograms were computed at each height, with a resolution of 0.05 in cloud fraction. Figure 2 shows an example of a joint histogram for the DWD model at Murgtal in 2007 (at a lower resolution of 0.1 for clarity). Note that the joint histograms from the lowest 11 km of the atmosphere have been summed, the cloud fraction above 11 km being insignificant in the observations. Also shown are the histograms from the observations and model separately, which (at least in the observations) exhibit the characteristic U-shaped distribution. The model shows a tendency to underestimate the occurrence of completely overcast skies. The joint histogram shows most of the data lying around the edge, suggesting a rather poor association between the two datasets. However, the lower frequency events in the centre of the panel (in light grey) appear to show a better correlation.

In addition to calculating joint histograms for the 1-hourly model snapshots, the modelled and observed cloud fractions have been averaged in time to 2, 3, 4, 6, 8, 12 and 24 hours, in order to investigate the improvement in forecasts that results from the consideration of larger horizontal scales. The joint histogram for 6-hour averaging is shown in Figure 3. It is striking how much better the visual degree of association is between the two variables, reflecting the fact that larger-scale cloud structures are easier to forecast than individual clouds.

Joint histograms can be used to calculate almost any skill score. For those that depend on the value of the difference between the observed and modelled cloud

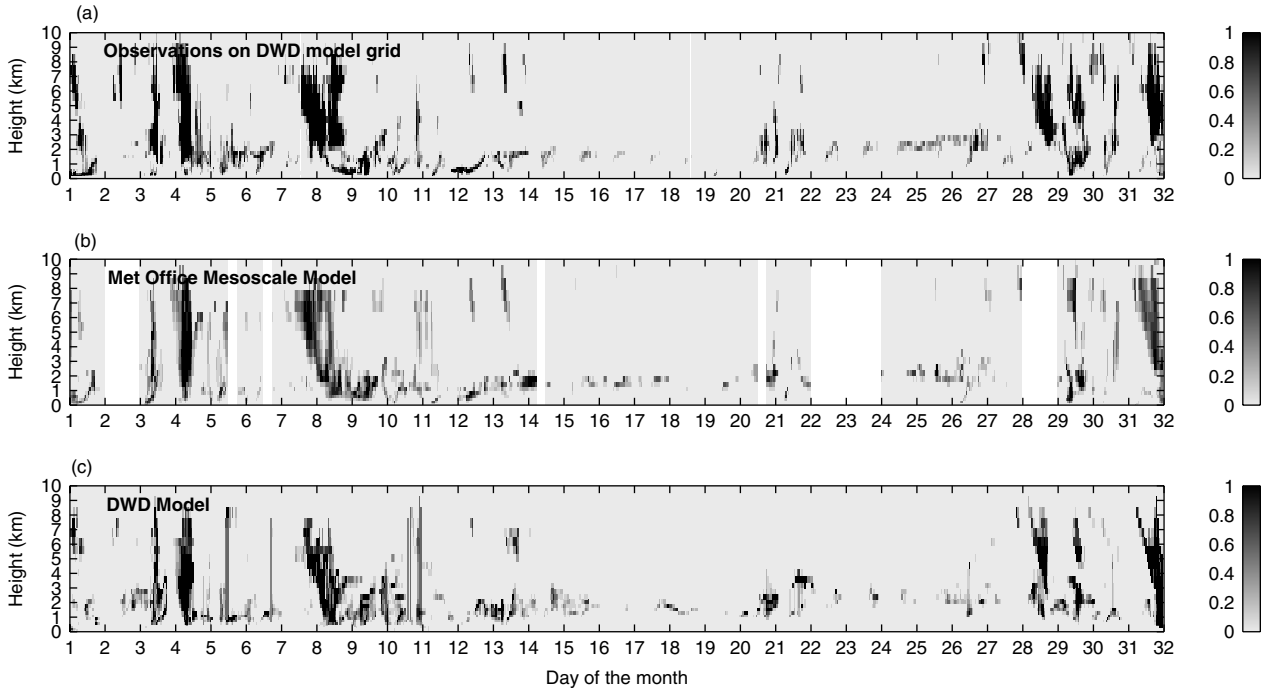

Figure 1. Comparison of (a) cloud fraction derived from the observations at Chilbolton during May 2004 on the grid of the DWD model, (b) the corresponding 6–11-hour forecasts of the Met Office mesoscale model, and (c) the corresponding 6–17-hour forecasts of the DWD model. Note that high clouds that would be undetectable in the observations have been removed from the models. The white regions in (b) indicate missing data not used in the analysis.

DOI: 10.1002/qj

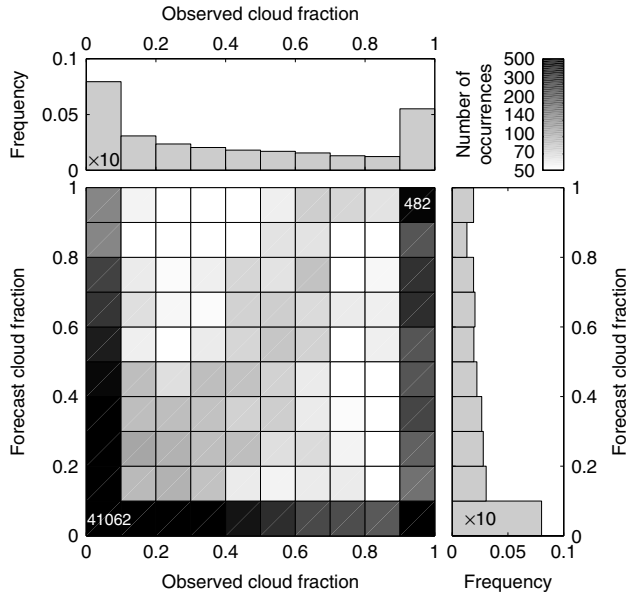

Figure 2. The main (lower left) panel shows the joint histogram of cloud fraction from the radar and lidar observations at Murgtal in 2007, and the cloud fraction modelled by the DWD model (with a 0–2-hour lead time) for the same period and location. Cloud fraction at 1 km intervals between 0 and 11 km above the ground have been included. The white numbers in the top-right and bottom-left intervals show the number of events in these bins. The corresponding (logarithmic) scale is shown at the top right. The panels to the top and the right show the probability distribution of observed and modelled cloud fraction, respectively; note that the magnitude of the lowest bar is shown at a tenth of its true value.

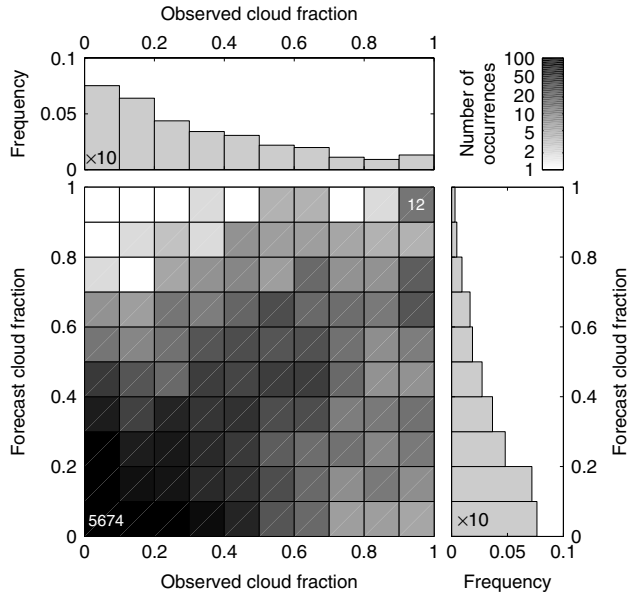

Figure 3. As Figure 2, but after averaging the observed and modelled cloud fraction into 6-hour periods.

fraction, $f_o - f_m$ (where f_o and f_m are individual values of cloud fraction from the observations and model, respectively), this information is available from the histogram with a precision of 0.05. For binary skill scores, we convert the joint histogram into a contingency table simply by dividing it into four quadrants for a particular value of the cloud-fraction threshold, f_t, and summing the values in each quadrant to yield the elements a–d. For the case shown in Figure 2, the contingency table for $f_t = 0.1$ is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 7194 & 4098 \\ 4502 & 41062 \end{pmatrix}. \quad (1)$$

This will be used to demonstrate the calculation of skill scores in section 4.

As will be described in detail in section 3, equitable skill scores are defined such that a random forecast would yield an expected score of zero (or some other constant value). It is therefore necessary to define the elements of the joint histogram that would be expected for a forecast that had the same probability distribution as the actual forecast, but was perfectly random. It is convenient to define two column vectors, $\mathbf{p}_o$ and $\mathbf{p}_m$, which contain the number of occurrences of cloud fraction in each 0.05 interval between 0 and 1, for the observations and the model separately. If the random joint histogram is represented as a matrix, $\mathbf{P}_r$, then

$$\mathbf{P}_r = \frac{1}{n}\mathbf{p}_m\mathbf{p}_o^T, \quad (2)$$

where n is the total number of elements in either of the vectors, and T represents the transpose.

The random binary contingency table can then be constructed either by dividing $\mathbf{P}_r$ into quadrants as before, or using the elements of the actual contingency table a–d. In the latter case, the random contingency table would have the following elements

$$a_r = (a+b)(a+c)/n; \quad (3)$$
$$b_r = (a+b)(b+d)/n; \quad (4)$$
$$c_r = (c+d)(a+c)/n; \quad (5)$$
$$d_r = (c+d)(b+d)/n, \quad (6)$$

where $n = a + b + c + d$ is the total number of elements. An important point to note is that the quantities a_r to d_r are the *expected values* in the contingency table, for a random forecast with the same probability of forecasting occurrence as the actual forecast system. (Likewise, the elements of $\mathbf{P}_r$ are the expected values for a random forecast.) A particular sample of n random forecasts may yield a different distribution of values of a–d, particularly for small n, and so a skill score calculated from this sample may imply the random forecasts to have positive or even negative skill (i.e. worse than random) for that sample. It is only if the skill score is some kind of linear function of a–d that the *expected value* of the skill score will be the baseline value assigned to a random forecast (Gandin and Murphy, 1992). For the remainder of this paper we use the term 'random forecast' to mean one for which the elements of the contingency table or the joint histogram are the expected values for a random forecast, given by (2)–(6). This will be true as $n \to \infty$, and hence is a good approximation for the large samples ($n \sim 10^5$) considered here.

DOI: 10.1002/qj

If a single skill score is to be calculated for clouds at any height, then an important aspect to generating both the random joint histogram and the random contingency table is that they should be calculated separately at each height, and then summed. In this way the skill score will be referenced to a random forecast that has some knowledge of the climatological probability of cloud fraction as a function of height (such as from a free-running climate model, or a weather forecast from a different year). In the case of the joint histogram shown in Figure 2 (which corresponds to heights between 0 and 11 km), the elements of the corresponding random contingency table are (noting that they will not in general be integers)

$$\begin{pmatrix} a_\mathrm{r} & b_\mathrm{r} \\ c_\mathrm{r} & d_\mathrm{r} \end{pmatrix} = \begin{pmatrix} 2581.0 & 8711.0 \\ 9115.0 & 36449.0 \end{pmatrix}. \tag{7}$$

This issue is discussed further in section 4.2. Other climatological variations in cloud fraction, such as the seasonal or diurnal cycle, could be accounted for in the same way, although for the midlatitude sites considered here this is a lesser effect than the variation with height and so is not included.

3. Desirable properties of verification scores

We now review the properties on which the various scores will be judged in section 4. Although the focus is on cloud fraction, almost all of these properties are desirable in the verification of any variable.

1. *Equitability.* An 'equitable' skill score awards all random forecasts an equal expected score (zero in the case of all such scores in this paper), even if they have the correct probability distribution (Gandin and Murphy, 1992). Also, a forecast system that always predicts the same value would be awarded zero. In section 4.1, some of the non-equitable scores that have been used for verifying clouds in the literature are listed. A complication is that the cloud-fraction climatology varies significantly with height, and a model with no skill at predicting clouds at the right *time,* but which nonetheless predicts clouds with about the right frequency versus *height* (e.g. randomly selected forecasts from a different year) could achieve a positive score by the normal definition of many equitable skill scores. An aim of this paper is to characterize the skill versus forecast lead time, assuming an approximately inverse-exponential decay towards a no-skill baseline (corresponding to randomly selected forecasts from the same model), and so it is necessary for this baseline to correspond to a score of zero. In section 4 we show how this can be achieved for a number of skill scores, by making use of the random joint histogram and the random contingency table presented in section 2. A further subtlety is that it turns out many scores are only strictly equitable in the limit of large n (a property which may be termed 'asymptotic equitability'; I. T. Jolliffe, personal communication). Full consideration of the consequences of this property will be given in a future paper, but in this paper we refer to such scores as 'equitable (for large samples)'.
2. *Difficulty to 'hedge' and transpose symmetry.* A score can be hedged if it encourages a forecaster to 'play the score' by issuing a forecast that differs from his or her true belief in order to yield a higher value (e.g. Jolliffe, 2008). It practice it is difficult to design a score that is completely impossible to hedge. Scores most easy to hedge are those that reward over-prediction of occurrence while penalizing under-prediction (or vice versa), since forecasts that predict cloud more often will tend to receive a higher score. We therefore favour scores that are *transpose symmetric,* which means that swapping the observations and the forecast does not change the score (Stephenson, 2000). For the scores considered in this paper, transpose asymmetry is a reliable indicator of the ones that are the easiest to hedge. It should be noted that transpose asymmetry can be justified if the economic or human cost of a 'miss' is much worse than a 'false alarm' (e.g. for a tornado forecast), but for clouds this argument does not apply.
3. *Independence of the frequency of occurrence.* Often a binary skill score is used to evaluate a continuous variable (such as rain rate or cloud fraction) by applying a threshold f_t as shown in Table I. For example, Illingworth *et al.* (2007) calculated skill scores using a low threshold of $f_\mathrm{t} = 0.05$. A higher f_t leads to a lower 'observed frequency of occurrence' $p = (a + c)/n$, which is termed the 'base rate' in the general verification literature. Likewise, the 'modelled frequency of occurrence' $p_\mathrm{m} = (a + b)/n$ decreases with increasing f_t. The variation of the score with f_t should then indicate whether the model is better or worse at predicting more intense events. However, it was shown by Stephenson *et al.* (2008) that this is usually not possible because virtually all scores have an intrinsic dependence on p, and tend to a meaningless limit (usually zero) for vanishingly rare events. In section 4.5, we present a modified version of the score introduced by Stephenson *et al.* (2008) that does not have this dependence, yet is also equitable (for large samples) and transpose symmetric.
4. *Dependence on exact value of the prediction.* For radiative transfer the difference between a cloud fraction of 0.5 and 1 is as important as the difference between 0 and 0.5, but for a binary skill score using a single value of f_t, only one of these differences can be distinguished. One can calculate the skill as a function of f_t, but it is more convenient if a single score can be reported that is dependent on the full range of cloud-fraction values forecast. Such a score is presented in section 4.6.

DOI: 10.1002/qj

5. *Linearity.* One of the aims of this paper is to determine the 'half-life' of a cloud forecast, which assumes we can define a score that has an inverse-exponential decrease with forecast lead time. However, the relationship between one score and another is often nonlinear, which means that the calculated half-life will depend on which score is chosen. Therefore the *linearity* of the score is important, something that has not been considered in the literature in this context before. The other advantage of linearity is that it can ensure that a score is truly equitable (discussion in section 2.2 of Gandin and Murphy, 1992). Linearity can be defined with respect to several things, for example to a change in one or more elements of the contingency table or to some measure of the probability of making the forecast by chance. This property is examined further in section 4.3.

There are two further properties of a score that ought to be considered in their design, but for which the optimum property depends on the application:

6. *Complement symmetry.* A skill score calculated from a binary contingency table is *complement symmetric* if swapping occurrence for non-occurrence does not change the score (Stephenson *et al.*, 2008). For clouds there is no compelling reason to regard this as an important property.
7. *Dependence on forecast bias.* Suppose a particular forecast has a bias such that it under-predicts the occurrence of cloud, but whenever it does predict a cloud, a cloud is always observed, i.e. $b = 0$ and $c > 0$. Such a forecast is the best it can be, given its bias. The same is true of a forecast that over-predicts the occurrence of cloud, but has no 'misses'. Some skill scores (e.g. the Odds Ratio) would give such forecasts top marks, while others would penalize them to different degrees due to the non-zero value of either b or c. This is an important factor to consider when trying to assess whether one forecast is more skilful than another, since some scores are more tolerant of bias than others. In the verification of a completely continuous variable (e.g. temperature), it is a simple matter to remove the bias by recalibrating the forecast distribution to match the observed distribution before applying the threshold. This is not generally possible in the case of cloud fraction since it is only partially continuous, having 'mass points' at zero and one. In section 4.3, we present a new score that is the same as the Heidke Skill Score, except that it is tolerant of biases.

4. Skill scores for cloud fraction

4.1. Non-equitable scores

We first briefly mention some of the non-equitable scores that have been used for verifying clouds in the literature, although do not pursue their use in this paper. Mace *et al.* (1998) and Miller *et al.* (1999) quantified the skill of cloud occurrence in the ECMWF model using ground-based radar and spaceborne lidar, respectively. The scores they used were not equitable and in some cases were relatively easy to hedge: they were Proportion Correct $(a + d)/n$, Hit Rate $H = a/(a + c)$, False Alarm Ratio $b/(a + b)$ and Threat Score $a/(a + b + c)$ (although note that they used the term 'Hit Rate' for what we call 'Proportion Correct', and 'Probability of Detection' for what we call 'Hit Rate'). The properties of H are shown in Table II, and are the same for False Alarm Ratio. Most of the other scores discussed in this paper are also shown in Table II.

Another strictly non-equitable score that has been used for rainfall verification is the 'Fractions Skill Score' of Roberts (2008). This was designed not to go to zero for a random forecast, but rather to go to zero for the worst possible forecast.

4.2. Generalized skill scores and the Heidke skill score

Several of the equitable scores considered in this paper fall into the category of a 'generalized' skill score S, defined by

$$S = \frac{x - x_\mathrm{r}}{x_\mathrm{p} - x_\mathrm{r}}, \tag{8}$$

where x is some function of the elements of the contingency table or the joint histogram, x_r is the value of x that would be obtained by a random forecast while x_p is the value of x that would be obtained by a perfect forecast. Thus it can be seen that S will vary between 0 for a random forecast and 1 for a perfect forecast, therefore being equitable (for large samples) and bounded. The other desirable properties described in section 3 depend on what is chosen for x.

The simplest example of such a score is the Heidke Skill Score (HSS; Heidke, 1926), which is obtained by setting $x = a + d$. In this case a perfect forecast would have $x_\mathrm{p} = n$ and a random forecast would have

$$x_\mathrm{r} = a_\mathrm{r} + d_\mathrm{r}. \tag{9}$$

It turns out that the same score is obtained by setting $x = b + c$, and indeed many other linear combinations of the elements in the contingency table. HSS is therefore both transpose- and complement-symmetric. The 'traditional' definition of HSS involves substitution of (3) and (6) into (9), followed by rearrangement of (8) to obtain

$$\mathrm{HSS}_\mathrm{trad} = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}. \tag{10}$$

In the appendix it is shown how the standard error on HSS may be estimated. Using the values in (1) yields $\mathrm{HSS}_\mathrm{trad} = 0.531 \pm 0.005$. However, this does not account for the fact that a forecast may have no skill in terms of simulating weather systems at the right time, but may

DOI: 10.1002/qj

Table II. Summary of the properties of the various skill scores considered in section 4, where the numbers of the properties correspond to those discussed in section 3.

Property	H	HSS	ETS	OSS	$\ln\theta$	Q	EDS	SEDS	MAESS
Desirable properties for cloud fraction verification									
1. Equitable (for large samples)	No	Yes	Yes	Yes	Yes	Yes	No	Yes	N/A
2. Transpose symmetric	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
3. Uses full range of cloud fraction	No	No	No	No	No	No	No	No	Yes
4. Meaningful as $p \to 0$	No	No	No	No	Nearly	No	Yes	Yes	No
5. Linear	Yes	Yes	No	Yes	Nearly	No	Nearly	Nearly	Yes
Other properties									
6. Complement symmetric	No	Yes	No	Yes	Yes	Yes	No	No	Yes
7. Biased forecast can get perfect score	Yes	No	No	Yes	Yes	Yes	Yes	No	No

Note that (i) the two transpose-asymmetric scores (H and EDS) can be easily hedged by predicting cloud all the time, and (ii) the 'equitability' property is not applicable to MAESS since it is not a categorical score of the type considered by Gandin and Murphy (1992) in the original definition of equitability.

still have the correct climatology with regard to the distribution of clouds as a function of height. Indeed, substitution of the values in (7) into (10) yields $\mathrm{HSS_r} = 0.028$. This is overcome simply by ensuring that a_r and d_r in (9) are calculated by summing the values at all heights, as described in the paragraph preceeding (7). The resulting value for HSS in this case is then 0.518.

The conceptual simplicity of HSS makes it suitable as a reference against which other categorical skill scores can be compared. For example, Figure 4(a) shows the theoretical variation of Hit Rate with HSS for four configurations of the overall frequency that the event is observed and modelled (p, p_m), which are (0.1, 0.1), (0.2, 0.2), (0.1, 0.2) and (0.2, 0.1). These have been calculated by considering a population of $n = 1000$ individual forecasts, and simulating every possible combination of a–d that is consistent with the values of p and p_m. Note that HSS can take on negative values corresponding to forecasts that are worse than random, but this is not shown.

The non-equitability of the Hit Rate is evident by the fact that it does not have a constant value for a random forecast indicated by HSS = 0. Its transpose asymmetry is evident from the difference between the dotted and dot-dashed lines, which means that forecasts that over-predict the occurrence of cloud will typically perform better than forecasts that under-predict cloud. It is this property that makes Hit Rate easy to hedge: simply by changing a random selection of forecasts of clear-sky to forecasts of cloud is guaranteed to improve the score awarded. The limit of this behaviour is to predict cloud all the time (resulting in $c = d = 0$), leading to a perfect score of $H = 1$. HSS also takes bias into account, but penalizes under- and over-prediction equally: the maximum score with a factor-of-two bias in its prediction of cloud is 0.615, regardless of the sign of this bias.

The Heidke skill score is uniquely related to the widely-used Equitable Threat Score (Gilbert, 1884; Doswell *et al.*, 1990), given by $\mathrm{ETS} = (a - a_\mathrm{r})/(a + b + c - a_\mathrm{r})$. This relationship is shown by the solid line in Figure 4(b). ETS can be seen to be simply a nonlinear

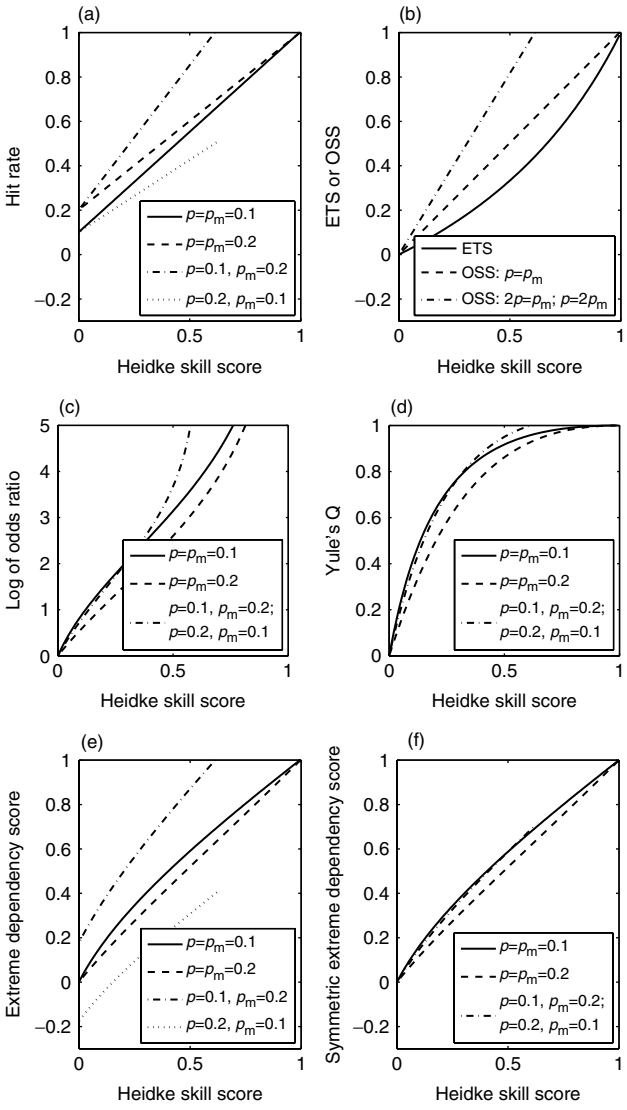

Figure 4. Comparison of various skill scores against the Heidke Skill Score (HSS), calculated numerically for various values of the frequency of occurrence in the observations (p) and the model (p_m): (a) Hit Rate, (b) Equitable Threat Score and Overlap Skill Score, (c) Log of Odds Ratio, (d) Yule's Q, (e) Extreme Dependency Score, and (f) Symmetric Extreme Dependency Score. Note that when p is twice or half p_m, the maximum HSS attained is 0.615.

DOI: 10.1002/qj

version of HSS, and since linearity is desirable for calculating half-life, ETS is not considered further in this paper. Another related score is the Peirce Skill Score (e.g. Stephenson *et al.*, 2008); this is identical to HSS for unbiased forecasts, but for biased forecasts it is transpose asymmetric so is also not considered further in this paper.

4.3. Linearity and the Overlap Skill Score

To examine the concept of linearity in more detail, consider the verification of a model that predicts cloud $p_\mathrm{m} = 1/3$ of the time, but cloud is observed $p = 1/5$ of the time. If we reorder each forecast into 'misses', 'hits', 'false alarms' and 'correct negatives' then the result can be shown schematically in Figure 5. The skill of the model can be changed simply by sliding the forecast cloud events left and right. Figure 5(a) depicts the scenario for a perfectly random forecast and Figure 5(c) the scenario for the best forecast that is possible given the bias of the model. The scenario in Figure 5(b) can be seen to lie half way between the other two, and we may define a *linear* score as one that would award the scenario in Figure 5(b) a value half-way between the scores it would award for the other two scenarios. More generally, for p and p_m held constant, we define a linear score as one for which changing a 'miss' to a 'hit' (i.e. adding one to a and d and subtracting one from b and c) increases the score by the same amount no matter the current values of a–d. The values of several scores are shown beneath each scenario in Figure 5, and it can be seen that HSS is linear by this definition, while ETS is not.

The dependence of HSS on bias is evident from the fact that the maximum score of 1 is not awarded for the scenario in Figure 5(c), yet it is for the scores Q and EDS discussed in the following sections. For some applications, we may wish to define a linear score that does award 1 for the scenario in Figure 5(c). This may be achieved by considering the analogy with the cloud 'overlap parameter' defined by Hogan and Illingworth (2000), which quantifies the degree of vertical overlap of clouds in different layers in the atmosphere (which could be illustrated by the grey regions in Figure 5). Within the framework of the generalized skill score defined by (8), the metric we consider is the number of times that cloud is forecast or observed (or both), such that $x = a + b + c$. Clearly a random forecast would result in $x_\mathrm{r} = a_\mathrm{r} + b_\mathrm{r} + c_\mathrm{r}$, but the best possible forecast given that the forecast system may be biased is $x_\mathrm{p} = \max(a + b, a + c)$. The result we refer to as the *Overlap Skill Score,* OSS. Figures 4(b) and 5 show that OSS varies linearly from 0 (random) to 1 (best possible) even when the model is biased; for unbiased forecasts it is equal to HSS. In general, the OSS for cloud fraction is not as satisfactory as for some of the other skill scores, so its use is not pursued further in this paper.

4.4. Log of Odds Ratio and Yule's Q

Stephenson (2000) advocated the use of the Odds Ratio, defined simply as $\theta_\mathrm{trad} = ad/bc$ (where the subscript refers to the 'traditional' definition to contrast with our slightly modified definition below). It can vary over many orders of magnitude, so the 'Log of Odds Ratio', $\ln\theta_\mathrm{trad}$, is more convenient. $\ln\theta_\mathrm{trad}$ is equitable (for large samples) with a random forecast scoring zero, although it is unbounded and so a perfect forecast would score infinity. The standard error of the Log of Odds Ratio, $\sigma_{\ln\theta}$, is given by

$$\sigma^2_{\ln\theta} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

(Agresti, 1996).

In the case of the contingency table represented by (1), the Log of Odds Ratio is $\ln\theta_\mathrm{trad} = 2.77 \pm 0.03$. However, the corresponding random contingency table given in (7) yields a non-zero score of $\ln\theta_\mathrm{r} = 0.17$, representing the residual skill of a random forecast that has some representation of the vertical cloud fraction climatology. We are therefore motivated to redefine the score to ensure that such a forecast yields a score of zero, as follows

$$\ln\theta = \ln\left(\frac{ad}{bc}\frac{b_\mathrm{r}c_\mathrm{r}}{a_\mathrm{r}d_\mathrm{r}}\right) = \ln\theta_\mathrm{trad} - \ln\theta_\mathrm{r}. \qquad (11)$$

Thus the Log of Odds Ratio would be be reduced to 2.60.

Figure 4(c) compares $\ln\theta$ to HSS for four different combinations of p and p_m. Being unbounded, it cannot be perfectly linear, but for the typical range of 1–3 found in this paper, it is close enough to linear to be useful for calculating forecast half-life in section 5.5. A simple method to overcome the unboundedness of $\ln\theta$ is to use the related skill score, 'Yule's Q' (Yule, 1900), referred to as the 'Odds Ratio Skill Score' by Stephenson (2000). This is defined as

$$\mathrm{Q} = \frac{(\theta - 1)}{(\theta + 1)},$$

and is zero for a random forecast and unity for a perfect forecast. However, it is clear from Figure 4(d) that this is at the expense of strong nonlinearity, tending to 'saturate' at the high-skill end of its range, and hence it is unsuitable for estimating half-life. We therefore do not use it further in this paper.

4.5. Symmetric Extreme Dependency Score

As the threshold, f_t, is increased, the frequency of occurrence, p, naturally decreases, but a problem with all the scores considered so far is that they have an intrinsic dependence on p, and in the limit $p \to 0$ they asymptote to zero (except for $\ln\theta$, which tends to infinity, and Yule's Q, which tends to one). Thus it is often not possible to use them to determine whether a model is better or worse at forecasting more extreme events.

DOI: 10.1002/qj

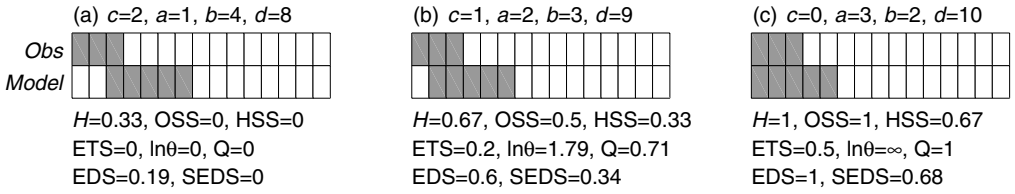

Figure 5. Schematic to illustrate the concept of a 'linear' skill score. The results of 15 forecasts are compared to observations, where a grey rectangle indicates that cloud occurred in either. (a) corresponds to a model that produces perfectly random forecasts, (c) to the best possible forecast given the bias of the forecasting system, and (b) to forecasts that lie half-way between the two by the definition of linearity given in section 4.3. The corresponding values of various skill scores are also provided for each scenario.

A potential solution was presented by Stephenson *et al.* (2008), who proposed the 'Extreme Dependency Score' (from a statistic introduced by Coles *et al.*, 1999), defined as

$$\mathrm{EDS} = \frac{2\ln\{(a+c)/n\}}{\ln(a/n)} - 1. \tag{12}$$

Stephenson *et al.* (2008) demonstrated that this score tends to a useful value for rare events, by first expressing it in terms of Hit Rate $H = a/(a+c)$ and frequency of occurrence $p = (a+c)/n$:

$$\mathrm{EDS} = \frac{2\ln p}{\ln(Hp)} - 1, \tag{13}$$

and then making the assumption that H has a power-law dependence on p for small p, i.e. $H \simeq \kappa p^{\delta}$. Here, $\delta = 1$ corresponds to H converging to zero at the same rate as a random forecast, while $\delta = 0$ corresponds to H not converging at all, i.e. a perfect forecast. From this, it may be easily shown that in the limit $p \to 0$, EDS $\simeq (1-\delta)/(1+\delta)$. This therefore indicates a measure of skill, typically lying between 0 for a random forecast and 1 for a perfect forecast.

Figure 4(e) shows that for $p = p_{\mathrm{m}}$, EDS is equitable and is quite similar to HSS. However, for $p \neq p_{\mathrm{m}}$, it is not equitable as it does not have a constant value for random forecasts. This may be demonstrated mathematically by setting a in the denominator of (12) to the random value given by (3) such that the denominator becomes $\ln(Bp^2)$, where $B = (a+b)/(a+c)$ is the frequency bias of the forecast. Further rearrangement yields

$$\mathrm{EDS} = \frac{-\ln B}{2\ln p + \ln B}. \tag{14}$$

Thus it can be seen that when $B \neq 1$, EDS is not zero and depends on the bias.

It is also clear that EDS is not transpose symmetric, since c appears in its definition but not b. It is this property that makes it easy to hedge: predicting cloud all the time would result in $c = d = 0$, and hence a perfect score of EDS = 1 (the same problem is true of Hit Rate, which also does not depend on b). Stephenson *et al.* (2008) avoided this problem in their analysis of 6-hour rainfall accumulations by recalibrating the forecast before evaluating it, thereby forcing $B = 1$. It would clearly be desirable not to have to do this, particularly for cloud fraction which is continuous over only a small part of probability space (around 80% of the time, cloud fraction is zero in the troposphere).

We therefore propose a transpose-symmetric modification of the EDS that we shall refer to as the *Symmetric Extreme Dependency Score:*

$$\mathrm{SEDS} = \frac{\ln\{(a+b)/n\} + \ln\{(a+c)/n\}}{\ln(a/n)} - 1 \tag{15}$$

$$= \frac{\ln(a_{\mathrm{r}}/a)}{\ln(a/n)}. \tag{16}$$

This may be expressed in terms of B, p and H as

$$\mathrm{SEDS} = \frac{\ln(Bp^2)}{\ln(Hp)} - 1. \tag{17}$$

If it is assumed that the frequency bias B remains constant as p is decreased, then it can be easily shown that SEDS has the same behaviour for rare events as EDS, tending to SEDS $\simeq (1-\delta)/(1+\delta)$ in the limit $p \to 0$. However, this score is equitable (for large samples), as may be demonstrated by replacing a in (16) by a_{r}, resulting in SEDS = 0.

The meaning of this score may be explained in a different way, by considering the fraction of the time that an event is correctly forecast, a/n. For random forecasts, it can be seen from (3) that $a_{\mathrm{r}}/n = pp_{\mathrm{m}}$. For perfect forecasts, $a_{\mathrm{p}}/n = p$. Perfect forecasts are also unbiased ($p = p_{\mathrm{m}}$), and so this may be written as $a_{\mathrm{p}}/n = (pp_{\mathrm{m}})^{1/2}$. Therefore the power of (pp_{m}) is an index of skill, and for forecasts of arbitrary skill we may manipulate (16) to obtain $a/n = (pp_{\mathrm{m}})^{1/(\mathrm{SEDS}+1)}$.

Figure 4(f) depicts SEDS versus the Heidke Skill Score for four different combinations of p and p_{m}. It is identical to EDS for $p = p_{\mathrm{m}}$, but for $p \neq p_{\mathrm{m}}$ it remains close to HSS, and penalizes equally under- and over-prediction by a factor of two. It is therefore no longer easily hedged by simply over-predicting occurrence, as was the case for EDS. The slight curvature of the relationship in Figure 4(f) shows that it is not perfectly linear, yet it is precisely this nonlinearity that makes it useful as $p \to 0$. We conclude that SEDS is a very useful general purpose skill score, as well as providing robust verification for rare events. This will be demonstrated in section 5.2.

The standard error in SEDS may be estimated by following a similar approach to Stephenson *et al.* (2008)

DOI: 10.1002/qj

for the EDS: by assuming p and B are constant and assuming $\sigma_{\mathrm{SEDS}} \simeq \sigma_H \partial\ \mathrm{SEDS}/\partial H$ yields

$$\sigma_{\mathrm{SEDS}} = \sqrt{\frac{H(1-H)}{np}} \times \frac{-\ln(Bp^2)}{H(\ln Hp)^2} = \sqrt{\frac{1}{a} - \frac{1}{a+c}} \times \frac{-\ln(a_\mathrm{r}/n)}{\ln(a/n)^2}. \quad (18)$$

To account for the fact that a random forecast with the correct height-dependent climatology of cloud fraction would be awarded a positive score, we use the definition given by (16), but ensure that a_r is calculated by summing the values at all heights, as described in the paragraph preceeding (7).

4.6. Mean Absolute Error Skill Score

All the skill scores considered so far require the cloud fraction forecast varying continuously between 0 and 1 to be discretized into a binary 'cloud' or 'no-cloud' forecast before being evaluated. As discussed in property 4 of section 3, it is desirable if a score can take full account of the range of values forecast. The obvious candidate is the Mean Squared Error Skill Score, MSESS (Murphy, 1988), obtained by setting $x = \langle (f_\mathrm{o} - f_\mathrm{m})^2 \rangle$ in (8). This may be calculated straightforwardly from the joint histogram discussed in section 2.2. The equivalent random value, x_r, may be calculated using the same method but applied to the random joint histogram. The equivalent value for a perfect forecast is simply $x_\mathrm{p} = 0$.

As with the Heidke skill score, MSESS is transpose symmetric. In terms of equitability, note that this property is usually only defined for categorical skill scores, but it can be seen from its definition that MSESS is equitable to the extent that a random forecast *with the same probability distribution as the actual forecast* will (by design) score zero, although a constant forecast will not necessarily score zero. The dependence of MSESS on the square of the error deserves some consideration. In terms of the effect of clouds on the instantaneous radiation budget, we could argue that, all else being equal, a model forecasting cloud fraction incorrectly by 0.2 on one occasion is equivalent to it forecasting cloud fraction incorrectly by 0.1 on two occasions. However, the definition of MSESS would penalize the first case twice as much as the second. We are therefore motivated to introduce the *Mean Absolute Error Skill Score* (MAESS) using the generalized skill score definition as before, but setting $x = \langle |f_\mathrm{o} - f_\mathrm{m}| \rangle$ in (8). Changing the forecast cloud fraction on a particular event by 0.1 would have the same effect on MAESS, regardless of how close that particular forecast was to the observations. This behaviour is more in keeping with the preference for scores to behave linearly, discussed in sections 3 and 4.3.

To estimate error bounds for MAESS, we use a Monte Carlo technique similar to that proposed by Déqué (2003) for MSESS. Since MAESS is calculated by comparing Mean Absolute Error (MAE) to the equivalent value for a random forecast ($\mathrm{MAE_r}$), we assume the main source of error to be due to estimating $\mathrm{MAE_r}$ from a finite number of samples. We first generate over 100 sets of random forecasts, each set containing n values of cloud fraction drawn randomly from the actual distribution of forecast values. MAE is calculated for each set by comparing with the observed values of cloud fraction, and the standard deviation of the resulting values of MAE is deemed to be the standard error of $\mathrm{MAE_r}$. The standard error of MAESS is calculated from this.

5. Results

5.1. Climatology

Before examining the skill of the models in predicting cloud at the right time, we assess their ability to reproduce the observed climatology of cloud fraction f. The most complete way to do this is by comparing the full probability distribution, as was done by Hogan *et al.* (2001) in three height ranges. Here we attempt to present this information as a continuous function of height. Figure 6 compares the frequency of occurrence of $f > f_\mathrm{t}$ for four different thresholds f_t between seven models and the corresponding observations for 2003–2004 averaged over the three original Cloudnet sites (Illingworth *et al.*, 2007), which are Chilbolton, Palaiseau and Cabauw. This is equivalent to the complement of a cumulative probability distribution.

The strengths and weaknesses of each individual model are different. Figure 6(a) shows that the ECMWF model has about the right occurrence of $f > 0.05$, but tends to underestimate the cloud fraction between 1 and 5 km when $f > 0.05$. The RACMO model (Figure 6(b)) tends to predict larger cloud fraction than ECMWF, and overpredicts the occurrence of $f > 0.05$. The Met Office models (Figures 6(c) and (d)) both appear to overestimate the occurrence of high cloud, even though all model fields have been modified in an attempt to remove clouds too tenuous to be detected. They also both significantly underestimate the occurrence of completely cloudy gridboxes (those with $f > 0.95$). The Météo-France and DWD models (Figures 6(e) and (f)) both appear to have a rather good cloud fraction distribution at all heights, while the SMHI-RCA model (Figure 6(g)) overestimates cloud occurrence at all heights except between 2 and 4 km. Further figures for each of the models used in the Cloudnet project can be found at www.cloud-net.org.

5.2. Dependence of skill on cloud fraction threshold

We now consider the skill of the model cloud fraction forecasts for the same dataset, i.e. their ability to place clouds at the right time and height, rather than just to have the correct distribution, which was assessed in the previous section. Firstly, the sensitivity of the scores to cloud fraction threshold, f_t, is assessed. Naturally, this can only be applied to the categorical scores, not continuous scores such as MAESS. Figure 7 shows the

DOI: 10.1002/qj

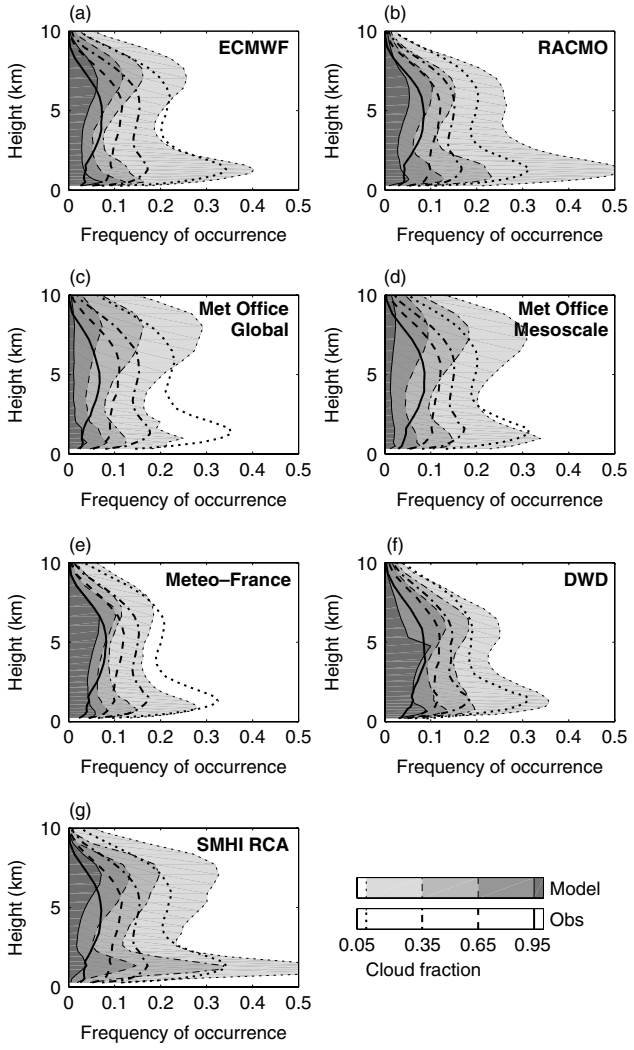

Figure 6. Frequency of occurrence of cloud fraction f greater than four thresholds ($f_t = 0.05$, 0.35, 0.65 and 0.95) for seven models (thin lines and shading), versus height, over Chilbolton, Cabauw and Palaiseau in 2003 and 2004. The corresponding frequencies of occurrence for the observations on the grid of each of these models are shown by the thick lines (the observations differ slightly due to the different grids of each model).

dependence of HSS, $\ln\theta$, EDS and SEDS on f_t for the seven models used during Cloudnet (for forecast lead times shown in Table I of Illingworth *et al.*, 2007), together with the frequency of occurrence.

In section 4, we described how the standard error on each score may be calculated, but this assumed that each forecast was independent, yet in reality forecast errors are correlated in both time and height. To approximately account for this, we have calculated 'transition probabilities' from the observed cloud occurrence distribution (using a threshold cloud fraction of $f_t = 0.1$), i.e. the probability of 'cloud' transitioning to 'clear', or vice versa, as a function of temporal and vertical separation. It is found that the e-folding decorrelation height is around 2 km and the e-folding decorrelation time is around 6 hours. Under the assumption that the autocorrelation of the forecast errors is the same as the autocorrelation of the cloud occurrence, and given that the resolution of the comparison is 1 km in height and 1 hour in time, we estimate that only 1 in 12 of the comparison points are independent. This may be approximately represented by multiplying the standard errors assuming independence by a factor of $\sqrt{12}$. A further factor of 1.96 is then applied to yield approximate 95% confidence intervals, which are shown in Figure 7.

Before examining the relative performance of each model, the information provided by each score must be reconciled. HSS exhibits an apparent decrease in skill with increased f_t, while $\ln\theta$ shows a slight but steady increase. Since the scores go in opposite directions, this must be due to the intrinsic dependence of these scores on frequency of occurrence p, examined in detail by Stephenson *et al.* (2008). By contrast, the SEDS remains close to constant with f_t for all but the Met Office models, confirming that it has little intrinsic dependence on p, in accordance with section 4.5. Therefore the underlying skill of most models appears to be approximately the same for low and high cloud fractions. SEDS could hence be extremely useful for other applications (particularly rainfall verification) where both the normal range of values and the extremes need to be assessed accurately. It should be noted that for $p > 0.05$, the Log of Odds Ratio is only weakly dependent on p, so previous evaluation of cloud-fraction forecasts using this metric (e.g. Wilkinson *et al.*, 2008) is not invalidated.

Figure 7(d) depicts the non-equitable EDS for these data and appears to show a very different picture; the ordering of the models is different with the SMHI-RCA model now performing much better, and the two Met Office models exhibiting an apparent strong decrease in skill with f_t. This can be explained by the fact that EDS rewards over-prediction and penalizes under-prediction of cloud occurrence: the Met Office models increasingly under-predict occurrence as f_t is increased, while the SMHI-RCA model over-predicts occurrence for $f_t < 0.5$. Therefore we conclude that EDS can only be used with confidence if it is possible to first 'calibrate' the observations to remove the bias, as was done by Stephenson *et al.* (2008).

The three other scores in Figure 7 are transpose symmetric, so treat over- and under-prediction equally, but they still differ in the extent to which they penalize the magnitude of any bias. From Table II, we see that $\ln\theta$ can award a perfect score to a biased forecast, whereas HSS and SEDS cannot. Figure 7(a) shows that the two versions of the Met Office model can significantly underestimate the occurrence of cloud for thresholds greater than 0.5, and the subsequent panels indeed show that HSS and SEDS award them a progressively lower score relative to the other models as f_t is increased, HSS doing this most strongly. By contrast, $\ln\theta$ appears to show the Met Office models remaining competitive with respect to the other models for large f_t.

Given these arguments and considering the size of the confidence intervals, we conclude that for $f_t < 0.5$, the ECMWF, RACMO and Met Office models perform similarly well in terms of forecast skill, with the DWD

DOI: 10.1002/qj

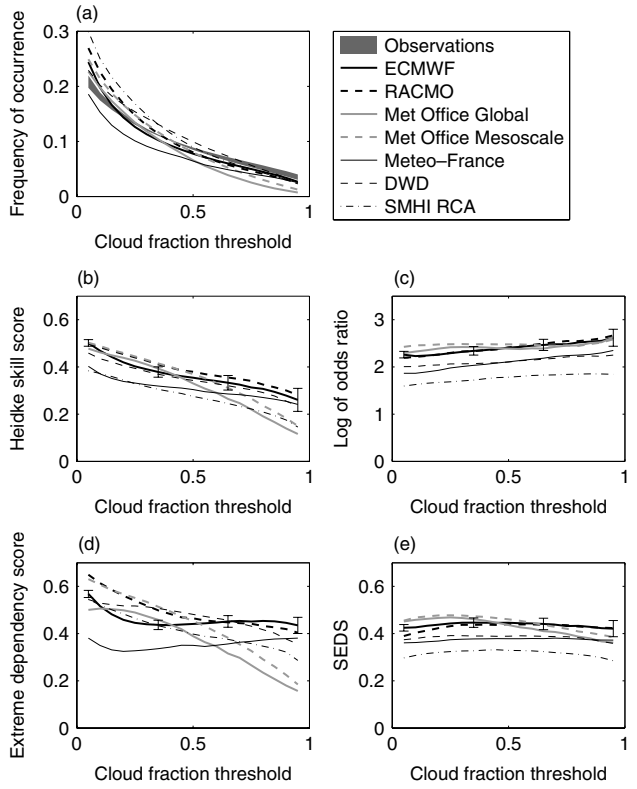

Figure 7. (a) The frequency of occurrence of cloud fraction greater than a threshold f_t in a gridbox, versus f_t, for observations (p) and seven models (p_m) at Chilbolton, Cabauw and Palaiseau in 2003 and 2004. Clouds above 11 km are discounted, and the spread of observed values (indicated by the width of the grey band) indicates the range of cloud fractions obtained when the observations are averaged to the grids of the various models. (b)–(e) The corresponding dependence of four categorical skill scores on f_t: HSS, $\ln\theta$, EDS and SEDS. The error bars on ECMWF scores indicate approximate 95% confidence intervals, the calculation of which is described in section 5.2; the intervals are approximately the same for the other models.

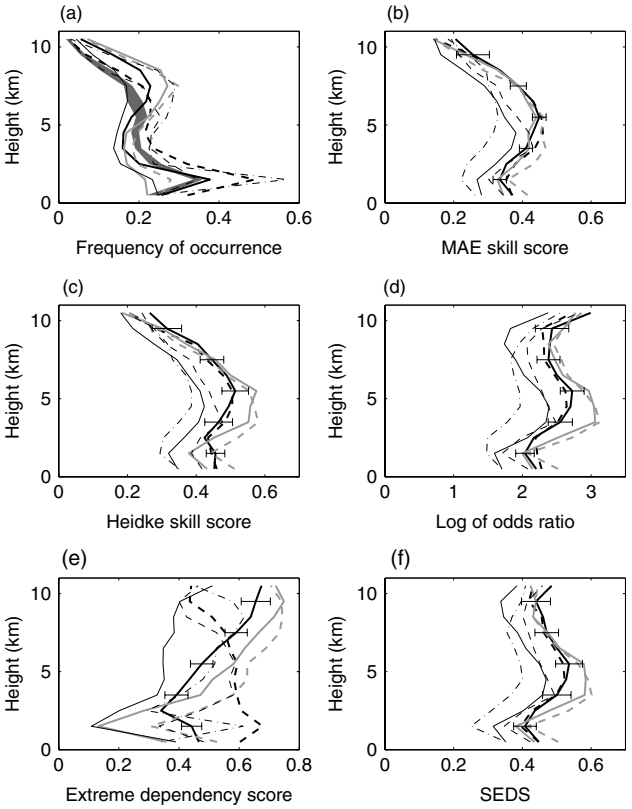

Figure 8. (a) Frequency of occurrence of cloud fraction greater than $f_t = 0.1$ versus height for the observations and the seven models shown in Figure 7 over the same period. Note that cirrus clouds too tenuous to be detected by the radar have been removed from the models, as described in section 2. (b)–(f) Mean Absolute Error Skill Score (MAESS), HSS, $\ln\theta$, EDS and SEDS as a function of height for the seven models, where all but MAESS were calculated using $f_t = 0.1$. Error bars indicate 95% confidence intervals.

and Météo-France models performing a little poorer and the SMHI-RCA model a little poorer still.

5.3. Dependence of skill on height

Figure 8 depicts the skill scores as a function of height in the atmosphere; in addition to the four categorical scores used in Figure 7 (now applied with a threshold of $f_t = 0.1$), we also show the MAESS, which is applied to the entire range of cloud fraction values at a precision of 0.05, as outlined in section 2. Confidence intervals are calculated as in section 5.2, but since scores are calculated as a function of height, we need only consider correlation of errors in time rather than height, so every sixth point is considered as independent rather than every twelfth.

As discussed in the previous section, we need to take account of the dependence of each score on p (shown in Figure 7(a)), particularly above 7 km where p decreases significantly with height. In this region, MAESS and HSS decrease significantly, $\ln\theta$ increases slightly and SEDS decreases slightly for most models. For the arguments given in sections 4.5 and 5.2, we are inclined to believe SEDS over the other scores, indicating that skill does indeed decrease slightly for high-altitude clouds.

A much less consistent picture is evident for EDS in Figure 7(e), with the Met Office models appearing to perform much better at high altitudes and the RACMO model better at low levels. This is again simply due to the differing biases of the various models with height and the dependence of EDS on the sign of the bias, as found in section 5.2. We therefore do not use EDS in the remaining parts of the paper.

All other scores in Figure 7 report a minimum skill between 1 and 2 km, presumably due to the well-known difficulty in forecasting stratocumulus and cumulus (e.g. Randall *et al.*, 1985; Martin *et al.*, 2000). Interestingly, all models show a maximum skill at mid-levels, between 4 and 6 km, where the frequency of cloud occurrence is at a minimum. This is possibly because such clouds are associated with large-scale synoptic systems that are easier to predict than smaller-scale clouds, and for which the amplitude of the vertical velocity variations is largest in the mid-troposphere. This appears to run counter to the prevailing perception that mid-level clouds are poorly simulated by current models (e.g. Ryan *et al.*, 2000; Bodas-Salcedo *et al.*, 2008). However, these studies concerned only the *bias* in models, whereas here we are evaluating the *skill* of the model in predicting clouds at the correct time, and indeed the Cloudnet data

DOI: 10.1002/qj

analyzed here also reveal this systematic underestimate in mid-level cloud fraction in all models except DWD (Illingworth *et al.*, 2007). This bias is not so evident in the frequency of occurrence of $f > 0.1$ in Figure 7(a); rather its origin is due to the tendency to underestimate f when $f > 0.1$. Thus we conclude that the models have good skill at predicting when mid-level clouds occur, but they almost all underestimate mean cloud fraction when some cloud is present.

According to HSS, $\ln\theta$ and SEDS, the two versions of the Met Office model have the highest skill at mid-levels. Two aspects to the microphysics of mid-level clouds in the Met Office model are unique or nearly unique amongst the models considered here. Firstly, no distinction is made between ice cloud and ice precipitation (i.e. snow) in the model, an assumption shared by the retrievals. Secondly, ice and liquid water mixing ratios are treated as separate prognostic variables (Wilson and Ballard, 1999), the DWD model being the only other to take this approach. Conceivably either of these properties could result in better behaviour at mid-levels. It should be noted from Figure 7(b), however, that the MAESS of the Met Office is no higher than that of ECMWF and RACMO at this altitude. This is likely to be because MAESS penalizes the Met Office model for its substantial underestimate of cloud fraction when some cloud is present, even if it does predict cloud occurrence at the right time.

5.4. Dependence of skill on spatial scale

As discussed in section 2, the raw model data are available as hourly snapshots in the model column closest to each site, while the observed cloud fractions are calculated from radar and lidar data in periods centred on each hour by sampling for a time equivalent to that necessary to advect a gridbox of cloud over the site. By averaging both observed and modelled cloud fraction temporally over 2, 3, 4, 6, 8, 12 and 24 hours before calculating skill scores, the models' ability to simulate increasingly larger scales is assessed. Note that these are not continuous averages, but averages of the hourly snapshots. Figure 9 depicts MAESS, $\ln\theta$ and SEDS as a function of the number of hours averaged. The first point on the abscissa corresponds to no averaging. In the case of $\ln\theta$ and SEDS, a cloud-fraction threshold of $f_t = 0.1$ has been used, as before. It can be seen that all scores tend to increase with temporal averaging, in agreement with the increase in association apparent visually between Figures 2 and 3.

5.5. Dependence of skill on lead time

Lastly in this paper, the performance of the regional UK Met Office and German DWD models versus forecast lead time is characterized; these are the only models that provided forecasts at different lead times for the same verification time. Figure 10 depicts the performance of the 12-km-resolution Met Office mesoscale and 7-km-resolution DWD 'Lokal' models in 2004. The domains of both models covered the three original 'Cloudnet' sites used by Illingworth *et al.* (2007), while the DWD model also included Lindenberg within its domain, which lay outside the domain of the Met Office mesoscale model. All scores indicate a decrease of skill with lead time for all models and all sites. At Chilbolton (UK) the Met Office model performed better than the DWD model, presumably due to more UK data (particularly from weather radar) being assimilated. At Cabauw (Netherlands), and Palaiseau (France), the Met Office model appears to perform better but by a narrower margin. The DWD model

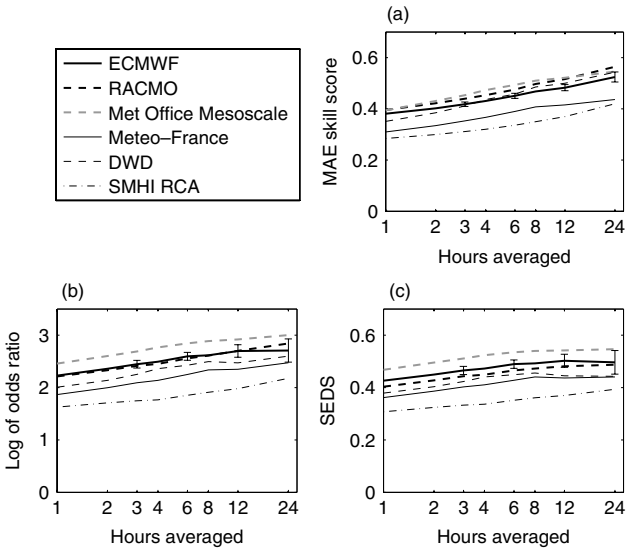

Figure 9. Three cloud-fraction skill scores for the same period as shown in Figure 7, as a function of the number of hours of temporal averaging that has been performed on the modelled and observed values. Note that the Met Office global model has been omitted because it provides only 3-hourly snapshots of cloud fraction, which is insufficiently frequent to show reliably in this figure. Error bars indicate 95% confidence intervals.

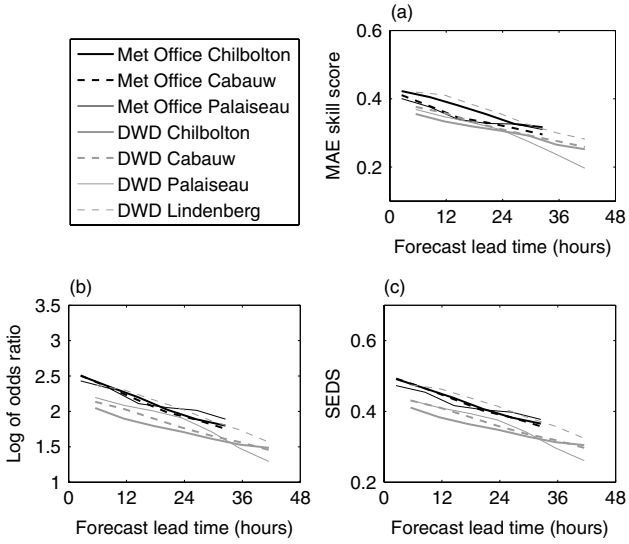

Figure 10. Three cloud fraction skill scores versus forecast lead time for the Met Office mesoscale model and DWD 'Lokal' models in 2004 over various European sites.

DOI: 10.1002/qj

performance is unsurprisingly best at Lindenberg (Germany), where it is as good as or slightly better than the Met Office model over Chilbolton.

The same analysis has been performed at three sites in 2007: Chilbolton and Lindenberg as before, but now including Murgtal in the Black Forest of southwest Germany. In the intervening time, the regional versions of both the Met Office and DWD models increased the size of their domains: the Met Office NAE model now extending west and east, and encompassing Lindenberg within its domain. The DWD 'COSMO-EU' model recorded forecasts out to a lead time of around 60 hours. The same scores are shown in Figure 11. The Met Office model again performs best over Chilbolton, while over Germany we see the DWD model performing better over Lindenberg but worse over Murgtal. It is interesting to note that the overall skill of both models appears to have improved notably in 2007 compared to 2004.

To estimate the 'half-life' of these forecasts requires fitting an inverse exponential to the curves, but for individual sites they are too noisy. We therefore combine the data for the sites shown in Figures 10 and 11 (excluding Lindenberg in 2004) for each of the two one-year periods. The resulting smoother curves for the three skill scores are shown by the thick lines in Figure 12. To fit the inverse-exponential, linear regressions have been performed to the natural logarithm of the score versus the lead time, yielding an equation for the best-fit of the score S as a function of time t:

$$S(t) = S_0 \exp(-t/\tau), \tag{19}$$

where S_0 is the score at $t = 0$ and τ is the e-folding decay time. These are then converted to the 1-day score $S_1 = S_0 \exp(-1/\tau)$ (where τ is in days) and the half-life $\tau_{1/2} = \tau \ln 2$, which satisfy

$$S(t) = S_1 \times 2^{(1-t)/\tau_{1/2}}. \tag{20}$$

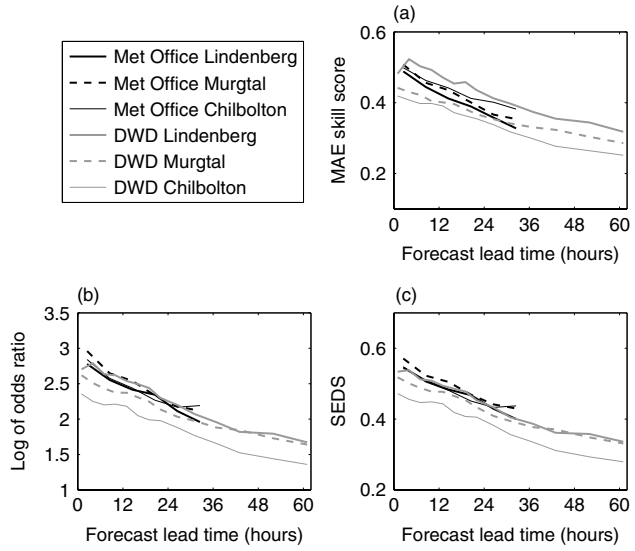

Figure 11. Three cloud fraction forecast skill scores versus forecast lead time for the Met Office 'North-Atlantic/European' and DWD 'COSMO-EU' models in 2007 over various European sites.

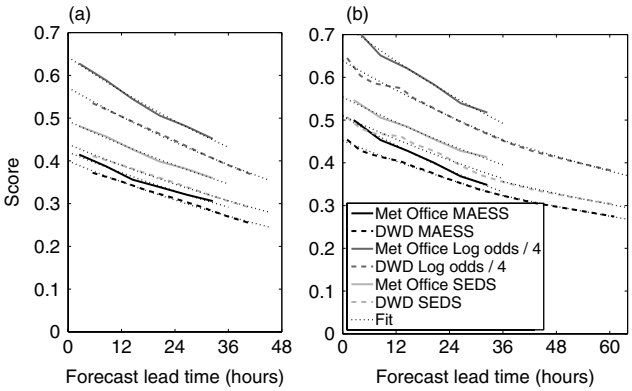

Figure 12. (a) As Figure 10, but combining the data from Chilbolton, Cabauw and Palaiseau in 2004 to obtain better statistics. (b) As Figure 11, but combining the data from Chilbolton, Lindenberg and Murgtal in 2007. In each case the dotted lines show inverse-exponential fits to the data with parameters given in Tables III and IV. Note that $\ln\theta$ is shown at a quarter of its true value.

The best-fit lines are shown by the dotted lines in Figure 12. In the case of the DWD model in 2007, all three skill scores indicate a slower rate of decline of skill after around 1.5 days. Therefore two fits are performed in this case, one for the 0–31-hour forecasts and the other for the 31–61-hour forecasts. In the other cases only one fit is performed. There is a slight shallowing of the slope evident between 18 and 24 hours for the Met Office model in Figure 12(a), although it is not different enough from the likely noise that two fits are justified.

The 1-day scores for the two models, two years and four of the skill scores, are shown in Table III. The standard errors were calculated by summing in quadrature the intrinsic error in the score (the calculation of which was discussed for each score in section 4, but multiplying by $\sqrt{12}$ to approximately account for the fact that only every sixth point in time and second point in height are independent), and the error arising from the errors in the coefficients of the least-squares fit, which is an indication of how well the points in Figure 12 are fitted by an inverse-exponential. Both models improve significantly (well in excess of the error estimates) between 2004 and 2007, with the Met Office maintaining its lead over DWD, despite two of the sites in 2007 being located in Germany.

Table IV shows the corresponding values of half-life. In comparing the three scores applied to binary contingency tables (HSS, $\ln\theta$ and SEDS), we typically find the shortest half-life estimated from $\ln\theta$ and the longest from SEDS or HSS. This difference can largely be attributed to the differences in the linearity between each score: Figure 4(c) shows a slightly concave relationship between $\ln\theta$ and HSS for the range of $\ln\theta$ considered here, while Figure 4(f) shows a slightly convex relationship between SEDS and HSS. The half-life calculated from Yule's Q is around 6 days in each case, demonstrating that this strongly convex score (illustrated in Figure 4(d)) yields a misleadingly long half-life, and verifying that linearity is an important property to consider.

DOI: 10.1002/qj

Table III. Values of the Mean Absolute Error Skill Score (MAESS), the Log of Odds Ratio ($\ln\theta$), the Heidke Skill Score (HSS) and the Symmetric Extreme Dependency Score (SEDS), for a lead time of 24 hours, i.e. the value S_1 in (20) for the best-fit lines shown in Figure 12 (except HSS).

Score	Met Office	DWD
2004: Chilbolton, Cabauw and Palaiseau		
MAESS	0.330 ± 0.041	0.308 ± 0.045
$\ln\theta$	1.97 ± 0.06	1.77 ± 0.05
HSS	0.404 ± 0.012	0.380 ± 0.011
SEDS	0.390 ± 0.012	0.346 ± 0.011
2007: Chilbolton, Lindenberg and Murgtal		
MAESS	0.383 ± 0.027	0.361 ± 0.022
$\ln\theta$	2.24 ± 0.07	2.04 ± 0.07
HSS	0.461 ± 0.016	0.425 ± 0.015
SEDS	0.442 ± 0.016	0.407 ± 0.016

Standard error calculation is discussed in the text.
Table IV shows the corresponding forecast half-lives.

Table IV. Forecast half-life $\tau_{1/2}$ (days) for the fits shown in Figure 12. The Met Office values were calculated from 0–31-hour forecasts, while the DWD values were calculated from the range of lead times shown at the top of the corresponding column.

2004: Chilbolton, Cabauw and Palaiseau			
Score	Met Office	DWD 0–42 h	
MAESS	2.85 ± 0.15	2.67 ± 0.08	
$\ln\theta$	2.61 ± 0.08	2.74 ± 0.05	
HSS	2.81 ± 0.06	2.97 ± 0.04	
SEDS	2.94 ± 0.09	2.90 ± 0.07	
2007: Chilbolton, Lindenberg and Murgtal			
Score	Met Office	DWD 0–31 h	DWD 31–61 h
MAESS	2.43 ± 0.09	3.09 ± 0.11	4.31 ± 0.16
$\ln\theta$	2.67 ± 0.14	3.05 ± 0.15	4.00 ± 0.22
HSS	2.88 ± 0.15	3.73 ± 0.24	4.51 ± 0.20
SEDS	3.09 ± 0.15	3.05 ± 0.21	4.30 ± 0.22

The standard errors are calculated from the error in the coefficients of the least-squares fit.
Table III shows the corresponding 1-day scores.

In terms of actual values of forecast half-life, the values for the Met Office in 2004 range between 2.61 and 2.94 days, and do not change significantly compared to the error estimates by 2007. The DWD values of 2.67–2.97 days in 2004 do appear to increase significantly by 2007 (calculated for the first 1.5 days of the forecast). Interestingly, the characteristic half-life for DWD 1.5–2.5-day forecasts is in the range 4.00–4.51 days, around 1 day greater than the corresponding 0–1.5-day forecasts. The same behaviour was found for Met Office rainfall forecasts over the UK by Roberts (2008), which he attributed to the skill at short lead times being dominated by the assimilation of radar data and the predictability of convective-scale events with intrinsically short time-scales, whereas after around a day the skill is determined by the predictability of larger-scale weather systems that have longer time-scales.

Another demonstration of this phenomenon is in Figure 13, which shows the half-life as a function of hours averaged for the 2004 data. A threshold of $f_t = 0.2$ has been used for $\ln\theta$ and SEDS, rather than 0.1 as previously. This is because, as the averaging period is increased, the frequency of occurrence p will tend towards the long-term average cloud fraction, which is around 0.2 in this dataset in the troposphere. It can be seen in Figure 7(a) that for no averaging, $f_t = 0.2$ results in $p \sim 0.2$, and therefore we expect p to remain approximately constant with averaging time. Hence any dependence of score on p will not have an effect on the calculation. Figure 13 shows that up to 6-hour averaging there is little increase in predictability, but for larger time-scales, associated with larger-scale cloud systems, there appears to be a strong increase in predictability. This is associated with the corresponding increase in absolute skill that was demonstrated in Figure 9. It should be noted, however, that as the number of hours averaged increases, the number of independent data points decreases, resulting in an increase in the error in any particular skill score estimate. This leads to even larger errors in fitting the inverse exponential, and explains the poorer agreement between the half-lives estimated by the three different skill scores for 24-hour averaging.

6. Conclusions

In this paper we have demonstrated how cloud forecasts may be verified using continuous cloud radar and lidar observations, utilizing suitable skill scores. The properties of a number of new and existing scores have been evaluated, with particular attention being paid to their *linearity,* a property important for estimating forecast 'half-life' that has not been considered in this context before. The findings are summarized in Table II. Of particular general interest for verification of binary forecasts is our new

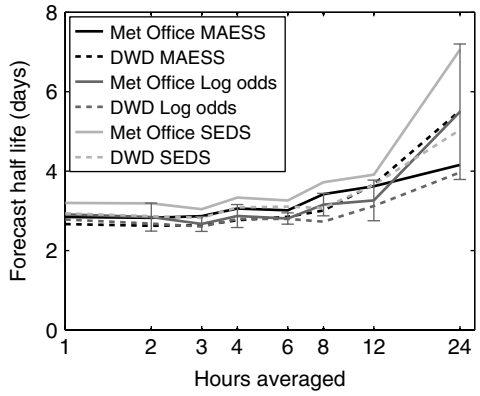

Figure 13. Forecast half-life for the same data as shown in Figure 12(a), but versus the number of hours averaged. For the Log of Odds Ratio and the Symmetric Extreme Dependency Score (SEDS), a threshold of $f_t = 0.2$ was used. Error bars indicate the approximate 95% confidence intervals, and are calculated from the standard error in the coefficients of the least-squares fit.

DOI: 10.1002/qj

score, the *Symmetric Extreme Dependency Score* (SEDS), a simple modification of the one proposed by Stephenson *et al.* (2008) that has the advantage of being equitable (for large samples), transpose symmetric (and therefore not easily hedged), while retaining the the ability to reliably verify forecasts of rare events.

The skill of single-site cloud-fraction forecasts from seven European models has been estimated as a function of cloud-fraction threshold, height, spatial scale and (for two of the models) forecast lead time. Using SEDS, it has been found that skill is essentially constant with cloud-fraction threshold, while the other scores considered yielded misleading results due to their intrinsic dependence on the frequency of cloud occurrence, p (except Log of Odds Ratio for p greater than around 0.05).

Models are found to be least skilful at predicting the occurrence of boundary-layer clouds and most skilful at predicting mid-level clouds. We stress that this is a statement about the ability to get clouds correct at the right time rather than with the right amount on average; indeed, almost all models tend to underestimate mean mid-level cloud fraction when cloud is present (Illingworth *et al.*, 2007). From what has been learned in this paper, we are now in a better position to interpret the results of Wilkinson *et al.* (2008), who calculated the Equitable Threat Score (ETS) and Log of Odds Ratio ($\ln\theta$) for ECMWF forecasts under the track of a spaceborne lidar. Considering ETS to be unreliable due to its dependence on frequency of occurrence p, and ignoring $\ln\theta$ when $p < 0.05$, we see that of all cloud types globally, tropical boundary-layer clouds were the least skilfully predicted and midlatitude clouds between 5 and 10 km altitude were the most skilfully predicted.

Forecasts with a lead time from 0 to 2.5 days have been used to estimate forecast 'half-life', i.e. the time over which the skill of a forecast would be expected to be halved. Values in the range 2.5–3.5 days are typically found when calculated from the first 1.5 days of the forecast, increasing to greater than 4 days for forecast lead times of 1.5–2.5 days. It is interesting that these values are considerably less than the half-life of 9 days estimated in section 1 for 500 hPa geopotential height. There are two main reasons for this difference. Firstly, we do not have cloud forecasts out to ~5 days to confirm the lead time at which the scores actually reach half of their initial value; rather, the 'half-lives' calculated in this paper are a measure of the rate of decay of skill in around the first two days of a forecast. The results for the DWD model in Figure 12(b) do show predictability time-scales tending to increase for larger lead times, so it would be interesting to apply this analysis to much longer forecasts. Secondly, very different variables are being evaluated: clouds are intrinsically of smaller scale and therefore more difficult to predict than larger-scale variables such as geopotential height. Temporal averaging of the data before verification indeed demonstrates that larger-scale features are more reliably predicted. We would expect clouds to have a similar predictability to a variable such as vertical wind, which has intrinsically much more small-scale structure than geopotential height.

We conclude that, for the ongoing development of mesoscale models, whose pressure fields are largely determined by the global model used to provide the boundary conditions, it is important that routine verification makes use of high-resolution observations of clouds, which are now becoming available in near-real time. It is also possible to apply many of the scores discussed in this paper to other cloud variables such as water content.

Acknowledgements

We are indebted to Ian Jolliffe, Chris Ferro and David Stephenson for very useful discussions that led to significant improvements to this paper. We thank Damian Wilson and Peter Clark for providing the Met Office model data, Axel Seifert and Thorsten Reinhardt for providing the DWD model data, Adrian Tompkins for providing the ECMWF model data, Erik van Meijgaard for providing the KNMI RACMO model data, Jean-Marcel Piriou for providing the Météo-France model data and Ulrika Willén for providing the SMHI-RCA model data. The Chilbolton radar and lidar observations were were provided by the Rutherford Appleton Laboratory, the Lindenberg observations by Ulrich Görstof of DWD, the Palaiseau observations at the 'Site Instrumental de Recherche par Télédétection Atmosphérique' by the Institut Pierre-Simon Laplace, the Cabauw observations by Henk Klein Baltinck of KNMI and the Murgtal observations by the Atmospheric Radiation Measurement programme. Malcolm Brooks carried out some of the software development and Peter Henderson is thanked for useful discussions. This work was supported by the European Union (grant EVK2-2000-00065) and NERC (grant NE/D005205/1).

Appendix

Confidence intervals in the Heidke Skill Score

Estimation of confidence intervals on categorical skill scores is not as common in the literature as it should be (Stephenson, 2000), partly because it is not obvious how to model the correlation between the errors in the elements of the contingency table. We assume that the base rate, p, is fixed, and therefore that the elements of the contingency table satisfy

$$a + c = pn \quad \text{and} \quad b + d = (1 - p)n.$$

This means that errors in a are perfectly anti-correlated with errors in c, and likewise for b and d, but errors in a and c are uncorrelated with errors in b and d. In order to work out an analytic formula for the error in a skill score, it is convenient to first redefine it in terms of constants (e.g. p and n) and uncorrelated variables (e.g. a and d).

DOI: 10.1002/qj

Doing this for the Heidke Skill Score yields

$$\mathrm{HSS} = \frac{2(1-p)a + 2pd + 2p(p-1)n}{(1-2p)(a-d) + (1-2p+2p^2)n} = \frac{X}{Y}. \quad \text{(A.1)}$$

It may be assumed (e.g. Mason, 2003) that a is a binomially distributed random variable drawn from a sequence of pn independent events each with probability $H = a/(a+c)$, i.e. $a \sim B(pn, H)$, and therefore its error variance (and hence also the variance in c) is given by

$$\sigma_a^2 = \sigma_c^2 = pnH(1-H) = \frac{ac}{(a+c)}.$$

Likewise,

$$\sigma_b^2 = \sigma_d^2 = \frac{bd}{(b+d)}.$$

The standard error of HSS, σ_{HSS}, may be given by

$$\left(\frac{\sigma_{\mathrm{HSS}}}{\mathrm{HSS}}\right)^2 = \left(\frac{\sigma_X}{X}\right)^2 + \left(\frac{\sigma_Y}{Y}\right)^2 + 2\frac{\overline{\epsilon_X \epsilon_Y}}{XY}, \quad \text{(A.2)}$$

where ϵ_X and ϵ_Y are the instantaneous errors in X and Y. The final covariance term may be calculated by writing X and Y as functions of a and d, i.e. $X = X(a, d)$, and hence

$$\epsilon_X = X(a + \epsilon_a, d + \epsilon_d) - X(a, d),$$

and similarly for ϵ_Y. Substituting these into $\overline{\epsilon_X \epsilon_Y}$, and noting that $\overline{\epsilon_a^2} = \sigma_a^2$, $\overline{\epsilon_d^2} = \sigma_d^2$ and $\overline{\epsilon_a \epsilon_d} = 0$, we obtain

$$\left(\frac{\sigma_{\mathrm{HSS}}}{\mathrm{HSS}}\right)^2 = 4\frac{(1-p)^2\sigma_a^2 + p^2\sigma_d^2}{X^2} + \frac{(1-2p)^2(\sigma_a^2 + \sigma_d^2)}{Y^2} + 4(1-2p)\frac{(1-p)\sigma_a^2 - p\sigma_d^2}{XY}. \quad \text{(A.3)}$$

References

Agresti A. 1996. *An Introduction to Categorical Data Analysis*. John Wiley and Sons:.

Bodas-Salcedo A, Webb MJ, Brooks ME, Ringer MA, Williams KD, Milton SF, Wilson DR. 2008. Evaluating cloud systems in the Met Office global forecast model using simulated CloudSat radar reflectivities. *J. Geophys. Res.* **113**: DOI: 10.1029/2007JD009620.

Coles S, Heffernan J, Tawn J. 1999. Dependence measures for extreme value analyses. *Extremes* **2**: 339–365.

Déqué M. 2003. Continuous variables. *Chapter 5 of Forecast verification: A practitioner's guide in atmospheric science.* Jolliffe IN, Stephenson DB (eds.) John Wiley &Sons: Chichester, UK.

Doswell CA III, Davies-Jones R, Keller DL. 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather Forecasting* **5**: 576–585.

Gandin LS, Murphy AH. 1992. Equitable scores for categorical forecasts. *Mon. Weather Rev.* **120**: 361–370.

Gilbert GK. 1884. Finley's tornado predictions. *Amer. Meteorol. J.* **1**: 166–172.

Heidke P. 1926. Calculation of the success and goodness of strong wind forecasts in the storm warning service. *Geogr. Ann. Stockholm* **8**: 301–349.

Henderson PW, Pincus R. 2009. Multi-year evaluations of a cloud model using ARM data. *J. Atmos. Sci. in press*.

Hogan RJ, Illingworth AJ. 2000. Deriving cloud overlap statistics from radar. *Q. J. R. Meteorol. Soc.* **126**: 2903–2909.

Hogan RJ, Jakob C, Illingworth AJ. 2001. Comparison of ECMWF winter-season cloud fraction with radar-derived values. *J. Appl. Meteorol.* **40**: 513–525.

Hogan RJ, Bouniol D, Ladd DN, O'Connor EJ, Illingworth AJ. 2003. Absolute calibration of 94/95-GHz radars using rain. *J. Atmos. Oceanic Technol.* **20**: 572–580.

Illingworth AJ, Hogan RJ, O'Connor EJ, Bouniol D, Brooks ME, Delanoe J, Donovan DP, Eastment JD, Gaussiat N, Goddard JWF, Haeffelin M, Klein Baltink H, Krasnov OA, Pelon J, Piriou J-M, Protat A, Russchenberg HWJ, Seifert A, Tompkins AM, van Zadelhoff G-J, Vinit F, Willén U, Wilson DR, Wrench CL. 2007. Cloudnet – Continuous evaluation of cloud profiles in seven operational models using ground-based observations. *Bull. Am. Meteorol. Soc.* **88**: 883–898.

Jakob C, Pincus R, Hannay C, Xu K-M. 2004. The use of cloud radar observations for model evaluation: A probabilistic approach. *J. Geophys. Res.* **109**: D03202, DOI: 10.1029/2003JD003473.

Jolliffe IT. 2008. The impenetrable hedge: A note on propriety, equitability and consistency. *Meteorol. Appl.* **15**: 25–29.

Lorenz EN. 1969. The predictability of a flow which posesses many scales of motion. *Tellus* **21**: 289–307.

Mace GG, Jakob C, Moran KP. 1998. Validation of hydrometeor occurrence predicted by the ECMWF model using millimeter wave radar data. *Geophys. Res. Lett.* **25**: 1645–1648.

Martin GM, Bush MR, Brown AR, Lock AP, Smith RNB. 2000. A new boundary layer mixing scheme – 2. Tests in climate and mesoscale models. *Mon. Weather Rev.* **128**: 3200–3217.

Mason IB. 2003. Binary Events. *Chapter 3 of Forecast verification: A practitioner's guide in atmospheric science,* Jolliffe IN, Stephenson DB (eds.) John Wiley &Sons: Chichester, UK.

Mass CF, Ovens D, Westrick K, Colle BA. 2002. Does increasing horizontal resolution produce more skillful forecasts? *Bull. Am. Meteorol. Soc.* **83**: 407–430.

Miller MA, Slingo A. 2007. The ARM Mobile Facility and its first international depolyment: Measuring radiative flux divergence in West Africa. *Bull. Am. Meteorol. Soc.* **88**: 1229–1244.

Miller SD, Stephens GL, Beljaars ACM. 1999. A validation survey of the ECMWF prognostic cloud scheme using LITE. *Geophys. Res. Lett.* **26**: 1417–1420.

Murphy AH. 1988. Skill scores based on the mean square error and their relationship to the correlation coefficient. *Mon. Weather Rev.* **116**: 2417–2424.

Palm SP, Benedetti A, Spinhirne JD. 2005. Validation of ECMWF global forecast model parameters using GLAS atmospheric channel measurements. *Geophys. Res. Lett.* **32**: L22S09, DOI: 10.1029/2005GL023535.

Randall DA, Abeles JA, Corsetti TG. 1985. Seasonal simulations of the planetary boundary layer and boundary-layer stratocumulus clouds with a general circulation model. *J. Atmos. Sci.* **42**: 641–676.

Roberts N. 2008. Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorol. Appl.* **15**: 163–169.

Ryan BF, Katzfey JJ, Abbs DJ, Rotstayn LD, Jakob C, Lohmann U, Rockel B, Stewart RE, Szeto KK, Tselioudis G, Yau MK. 2000. Simulation of a cold front using cloud-resolving, limited-area, and large-scale models. *Mon. Weather Rev.* **128**: 3218–3235.

Simmons AJ, Hollingsworth A. 2002. Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.* **128**: 647–677.

Stephenson DB. 2000. Use of the 'Odds Ratio' for diagnosing forecast skill. *Weather Forecasting* **15**: 221–232.

Stephenson DB, Casati B, Ferro CAT, Wilson CA. 2008. The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorol. Appl.* **15**: 41–50.

Tiedtke M. 1993. Representation of clouds in large-scale models. *Mon. Weather Rev.* **121**: 3040–3061.

Wilkinson JM, Hogan RJ, Illingworth AJ, Benedetti A. 2008. Use of a lidar forward model for global comparisons of cloud fraction between the ICESat lidar and the ECMWF model. *Mon. Weather Rev.* **136**: 3742–3759.

Wilson DR, Ballard SP. 1999. A microphysically based precipitation scheme for the Meteorological Office Unified Model. *Q. J. R. Meteorol. Soc.* **125**: 1607–1636.

Wilson DR, Bushell AC, Kerr-Munslow AM, Price JD, Morcrette CJ. 2008. PC2: A prognostic cloud fraction and condensation scheme – 1. Scheme description. *Q. J. R. Meteorol. Soc.* **134**: 2093–2107.

Wulfmeyer V, Behrendt A, Bauer H-S, Kottmeier C, Corsmeier U, Blyth A, Craig G, Schumann U, Hagen M, Crewell S, Di Girolamo P, Flamant C, Miller M, Montani A, Mobbs S, Richard E, Rotach MW, Arpagaus M, Russchenberg H, Schlüssel P, König M, Gärtner V, Steinacker R, Dorninger M, Turner DD, Weckwerth T, Hense A, Simmer C. 2008. The convective and orographically induced precipitation study. *Bull. Am. Meteorol. Soc.* **89**: 1477–1486.

Yule GU. 1900. On the association of attributes in statistics. *Phil. Trans. R. Soc. London* **194A**: 257–319.

DOI: 10.1002/qj