WP5–NA5: Clouds and aerosol quality-controlled observations

Deliverable D5.10: Implementation of new metrics

Objectives

The major objective for Task 5.4 is the routine and automated evaluation of the representation of clouds in numerical models. Evaluating the representation of clouds in climate and numerical weather prediction (NWP) models is not straightforward. For NWP models, this task is compounded by the expectation of a good forecast, as well as the reliable representation of the specific cloud parameters themselves. Task 5.4 of WP5 developed and implemented a comprehensive suite set of objective metrics for the evaluation of model cloud parameters, in continual joint collaboration with operational modellers.

The set of evaluation metrics was designed to investigate both the climatological aspects required of a climate model, and the ability to forecast the correct cloud at the right time, a necessary validation for NWP. The evaluation metrics cover the following sets of parameters:

- Model climatology (means, and distributions, of cloud variables)
- Forecast skill scores with appropriate characteristics (SEDS/SEDI)
- Composites (seasonal, diurnal, omega at 500 mb) of cloud variables and skill scores,
- Choice of vertical grid for model evaluation: height or temperature
- Model intercomparisons;

and the goal, which has been achieved, is a framework providing:

- routine , automated evaluation,
- near real time capability,
- online dissemination.

Evaluation procedure

Evaluation of model products is performed in levels 2b, 3, and beyond, of the Cloudnet processing chain shown in Fig. 1. In level 2b, the observations are placed on the grid of each individual model, and in level 3, amassed into monthly and yearly files containing a wide range of statistical measures. These include: means, distributions, and joint-pdfs for creating the contingency tables used for deriving the skill score of choice. From these files, a wide range of metrics can then be routinely plotted and analysed.

All standard statistical metrics that are described in this report are created by the automated system, in near-real-time wherever possible. In addition, the functionality to create custom metrics through user-selection of time-period and other choices, is technically feasible. The back-end system for the Cloudnet database incorporates a self-discovery mechanism for identifying new datasets. Standard metrics are created for the new datasets; the system identifies any current metrics that the new datasets may potentially belong to and updates the standard metrics as required.

The automatic procedures cover the entire production chain, from processing observations, through to evaluating model data and the generation of statistical metrics and figures. Incoming new observational and model datasets are identified, and all evaluation metrics are then created. The definition of new datasets can include near-real-time daily updates, a month/year of batch-processed model data, or reprocessed data from a previous period. The procedures identify and update any current metrics that are potentially dependent on the new data; this could include a yearly metric being updated due to the addition of a new month of data, or a seasonal composite being updated due to the addition of a new year of data.

Observed product on model grid

The model evaluation process begins at the Level 2b stage in the Cloudnet processing chain (see Fig. 1). Here, Cloudnet products at high temporal and spatial resolution are averaged to the specific grid of the model being evaluated. Temporal averaging is used to create the equivalent of a two-dimensional slice through the three-dimensional model grid box. As discussed in Illingworth et al. (2007), the appropriate amount of averaging time is given by the advective timescale, which describes how long it takes for a cloud structure to advect through the grid box and is given by the horizontal wind speed. The advective timescale is obtained as a function of height from the vertical profile of horizontal wind speed taken from a model, radiosonde, or a combination of radar and Doppler lidar wind profilers. Thus, given the 16 km horizontal resolution of the current ECMWF model grid, for example, a wind speed of 20 m s^{-1} corresponds to an averaging time of 800 seconds (13 minutes), centred on the model timestep.

Cloud fraction is calculated as the fraction of pixels diagnosed as cloud in the categorization product for the observed profiles that fall within the advective timescale. For observations with a resolution of 30 m and 30 seconds, and an advective timescale of 800 s, the observed cloud fraction corresponding to an ECMWF model grid box 100 m deep is calculated from about 150 independent pixels. For other bulk properties, such water contents, the observations are averaged according to the advective timescale (the mean and variance are the properties retained).

Note that, since the wind varies with both height and time, the advective timescale, and thus the number of points averaged, changes with height and time. As each model has different horizontal and vertical resolutions, a separate level-2b product is produced for each model.

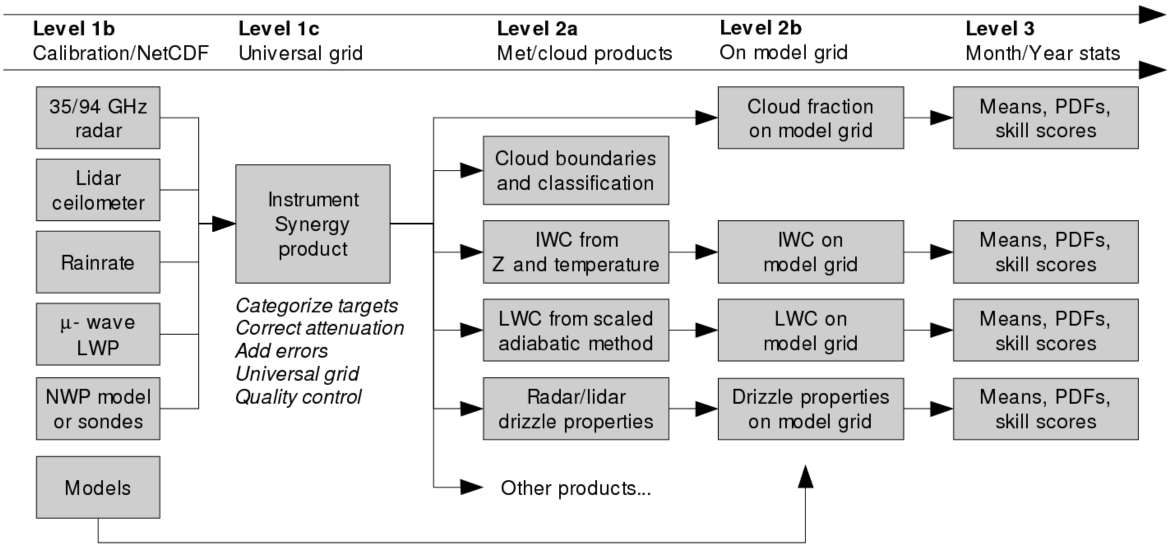


Figure 1: Overview of Cloudnet processing chain

Level 3 products – evaluating model climatology

Mean, frequency-of-occurrence and amount-when-present

The obvious metric to test first is the model climatology, or how the model performs on average, by comparing the mean observed and modelled values of the parameter of interest versus height over an extended period of time (Fig. 2). In addition, we also examine how often the given cloud parameter occurs (above a given threshold), and what is its value when it does occur, known as frequency-of-occurrence and amount-when-present (Hogan et al. 2001). For instance, a model may forecast clouds often enough but underestimate the amount of cloud when present; therefore the mean cloud fraction will be underestimated (Morcrette et al., 2012). The metrics are calculated for a range of threshold values and available in the level3 data files; a standard threshold value is selected to create the quicklooks that are provided routinely on the website. Some metrics, such as cloud fraction, are not that sensitive to the choice of threshold (see Fig. 5); others, such as water contents, may show more sensitivity, especially when increasing or decreasing the threshold value by an order of magnitude.

Distribution

The probability distribution function, PDF, of a given parameter is also calculated for each month and year (Fig. 3). These examine whether the model produces the correct range or distribution of the cloud parameter of interest. For example, the distribution of ice water content may be very close to matching the observed distribution, but a slight overabundance of very high modelled ice water contents would severely bias the mean value; merely scaling the modelled mean to match the observed mean would not be an appropriate correction for such a model scheme.

The PDFs are provided for broad height ranges (e.g. low, mid or high-level clouds), with the specific height range chosen to be appropriate for the cloud parameter; the height ranges for liquid water content are not the same as for ice water contents for example, since liquid water clouds should not exist at temperatures below -40 C, and ice should not exist above freezing (strictly speaking, ice is possible up to 4 C). PDFs are also possible to calculate for narrow height/temperature ranges as requested.

Vertical grid

There is a choice of vertical grids: altitude, or temperature. Certain features are more suited to a height grid, such as evaluating the representation of boundary-layer clouds; a temperature grid (Fig. 4) is more suited to identifying the presence of supercooled liquid water and temperature-specific ice growth processes.

Marked changes in model parameters have been noted at specific temperatures, and are understandable when the physical basis for certain parametrizations may be functionally dependent on temperature. For instance, some microphysical schemes use temperature to partition water condensate into liquid and ice; ice density, growth rates and fall speeds may also be defined as a function of temperature. Hence, calculating metrics on both the temperature and standard height grid in conjunction with each other should help identify where model deficiencies are due to parametrizations involving temperature, and thus where to concentrate investigating and improving specific shortcomings in microphysical parameterization schemes.

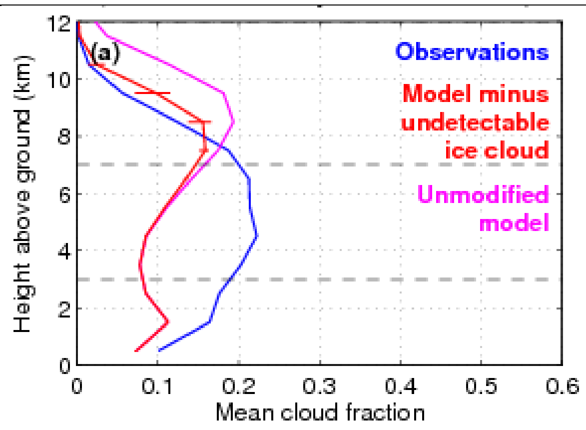
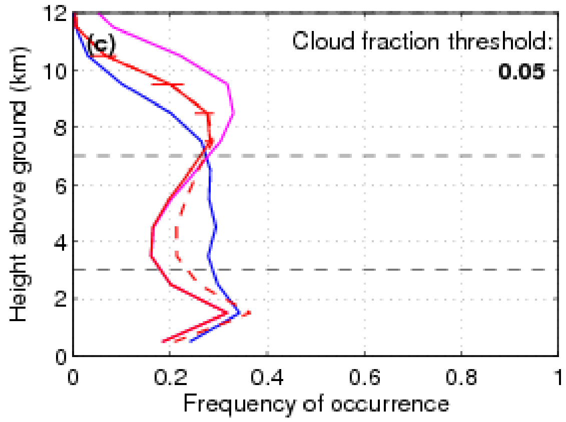
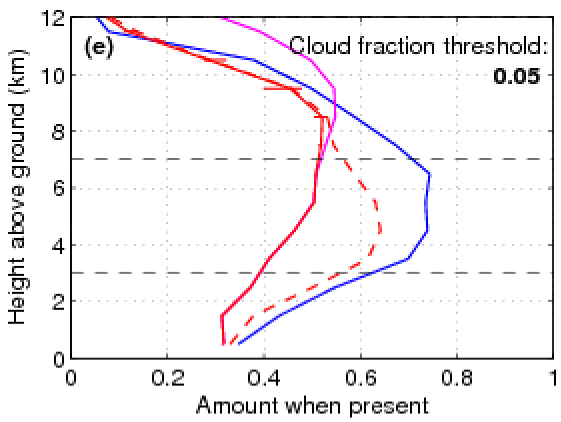
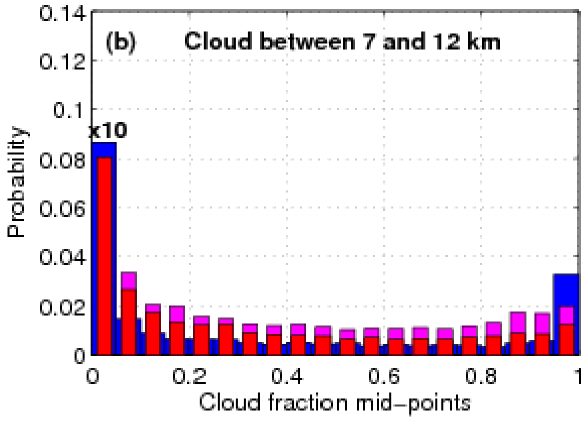
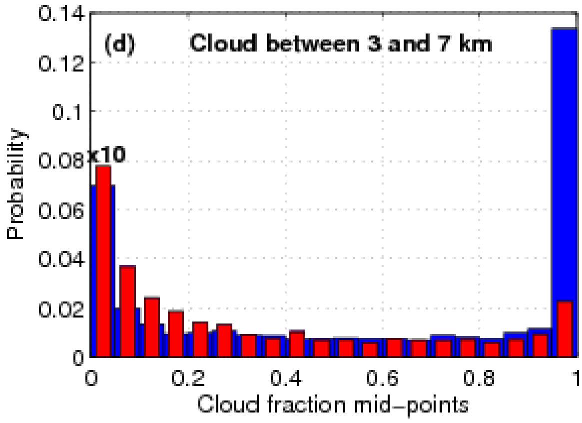
<table>
<tr><td>
</td><td>Mean cloud fraction</td></tr>
<tr><td>
</td><td>Frequency of occurrence</td></tr>
<tr><td>
</td><td>Amount when present</td></tr>
</table>

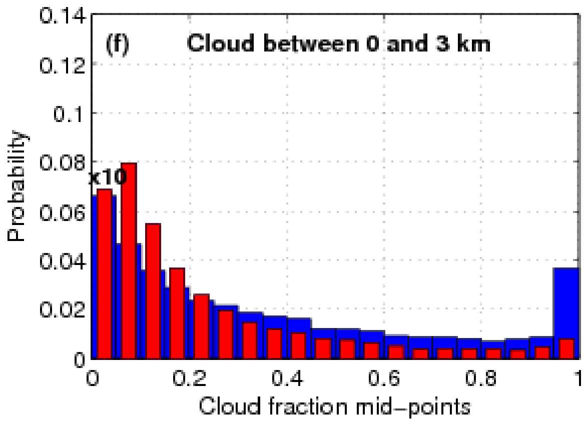
Figure 2: Evaluation of ECMWF cloud fraction (12-35 hour forecast range) at Juelich, Germany during 2014, comprising 309 days of data. The dashed red lines indicate model cloud fraction if snow is included. The vertical coordinate is height above ground.



High Cloud



Mid-level cloud



Low cloud

Figure 3: Evaluation of ECMWF cloud fraction (12-35 hour forecast range) at Juelich, Germany during 2014, comprising 309 days of data. This figure is generated from the same data as displayed in Fig. 2.

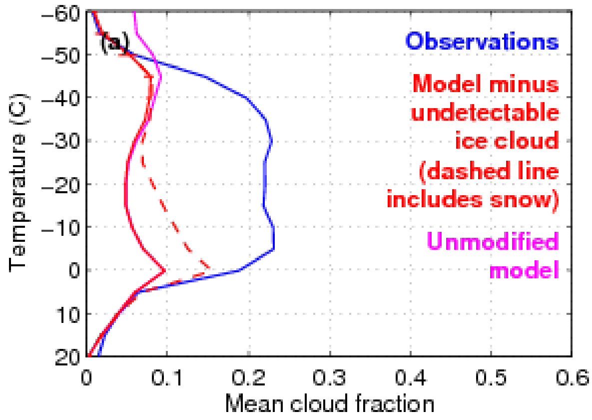
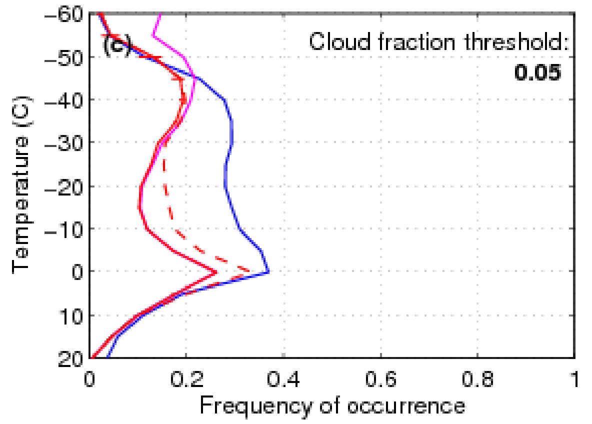
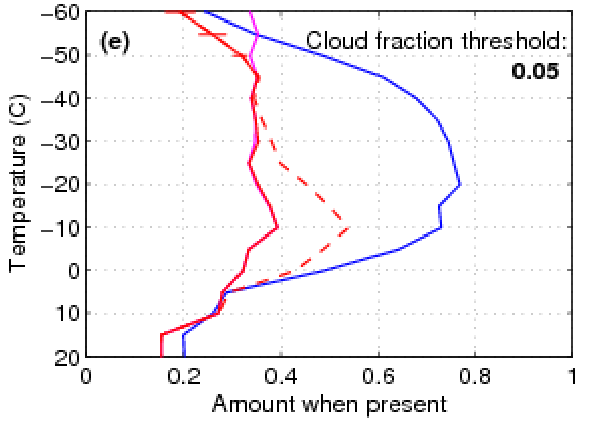
	Mean cloud fraction
	Frequency of occurrence
	Amount when present

Figure 4: Evaluation of ECMWF cloud fraction (12-35 hour forecast range) at Juelich, Germany during 2014, comprising 309 days of data. The dashed red lines indicate model cloud fraction if snow is included. The vertical coordinate is temperature. This figure is generated from the same data as displayed in Figs. 2-3.

Level 3 products – evaluating model forecast skill

Forecast skill

Evaluating the forecast skill of a model requires a skill score with suitable attributes (Hogan et al., 2009, 2010). Within Cloudnet we use the Symmetric Extreme Dependence Skill Score, SEDS, and Symmetric Extreme Dependence Index, SEDI. In addition, the original contingency tables and joint-pdfs are provided in the file, allowing the user to derive their skill score of choice. As for the climatological metrics, these scores are create for two sets of vertical coordinates: height above ground, and temperature (Fig. 5).

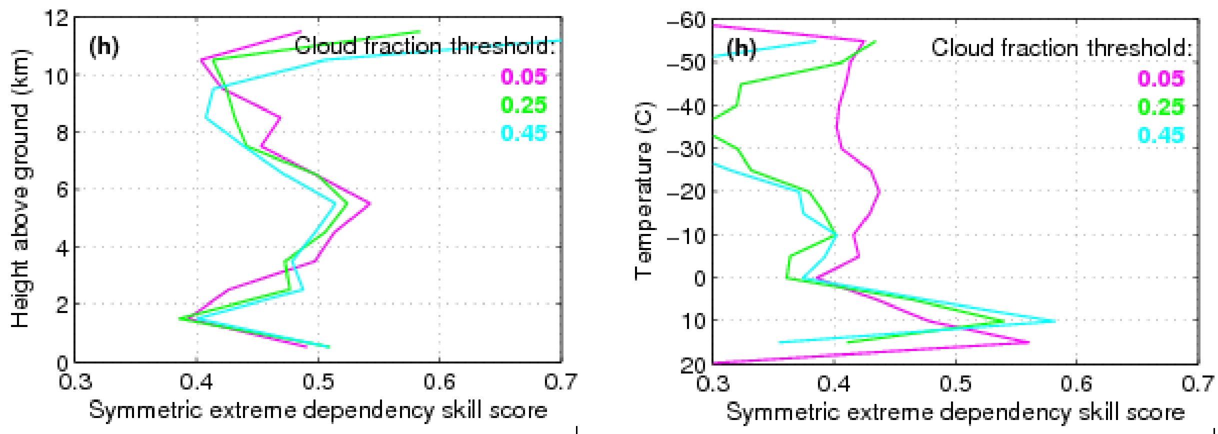


Figure 5: Evaluation of ECMWF skill in forecasting cloud fraction (12-35 hour forecast range) at Juelich, Germany during 2014, comprising 309 days of data. The vertical coordinate is height above ground (left panel) or temperature (right panel). This figure is generated from the same data as displayed in Figs. 2-4.

Forecast lead time

The climatological metrics and skill scores can be calculated for the same model at different forecast lead-times (Fig. 6). These metrics can indicate whether there is a change in model climatology between the initial state and later forecasts (Fig. 6a), which would show the impact of model spin-up, and whether the model has a preferred state within the analysis or data-assimilation cycle that is different to the free-running model state. These metrics also show how the model forecast degrades with time (Fig. 6b, 6c), and whether the rate of degradation is the same at all heights, temperatures, and for initial cloud parameter state.

Note that the ability to create this particular metric depends on the particular forecast output supplied by the modelling centre, and how long the model forecast runs for. Certain models are initialised twice a day with each forecast running out to 72 hours and more, whereas others, especially high-resolution models designed to capture high-impact weather, are often initialised more often (maybe 6-8 times a day) but only run for 24 hours.

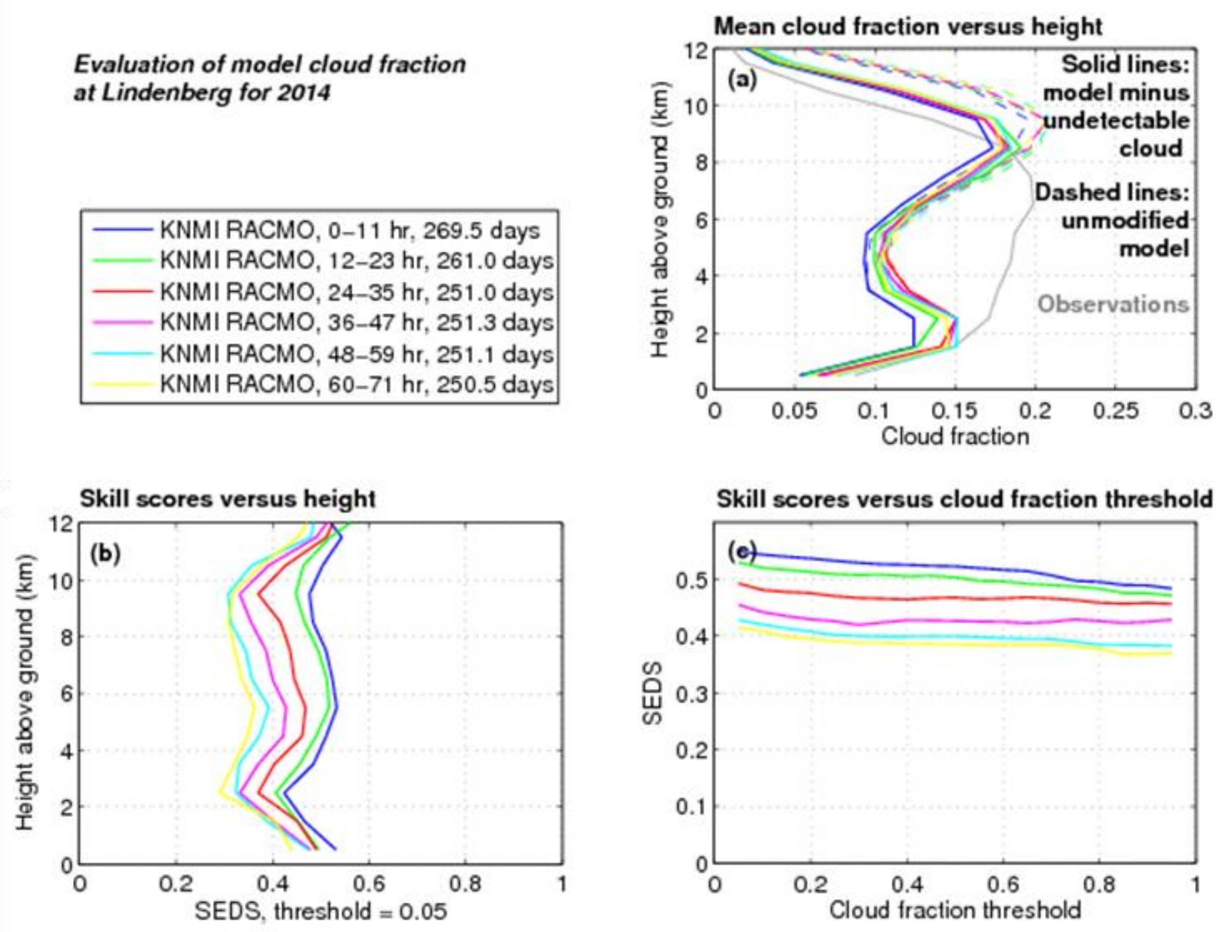


Figure 6: Evaluation of (a) KNMI RACMO cloud fraction and (b, c) skill in forecasting cloud fraction for various forecast lead-times at Lindenberg, Germany during 2014, comprising more than 250 days of data. The vertical coordinate is height above ground for panels (a) and (b). Panel (c) displays how sensitive the skill score is to choice of threshold.

Level 3 products – Composites

To identify and understand location-specific and model-specific deficiencies with more confidence, composites of cloud-fraction, ice water content, liquid water content and skill scores are now also available in both temperature and height vertical coordinates. Composites include seasonal, diurnal, and omega at 500 hPa.

An example of seasonal composites of cloud-fraction, ice water content, liquid water content and skill scores are shown in Fig. 7. These composites are available in both temperature and height vertical coordinates. These have revealed that, for many situations within the free troposphere, models show substantially higher forecast skill for clouds at temperatures below freezing than above. However, this increase in skill does not necessarily extend to clouds that lie within the boundary-layer

An additional metric is the composite of cloud fraction by vertical motion (omega) at 500 hPa, thereby examining the dependence of model behaviour on synoptic regime, i.e. large scale ascent (frontal) or descent (anticyclonic subsidence. A similar approach was taken in comparison of radiative variables by Bony and Dufresne (2005).

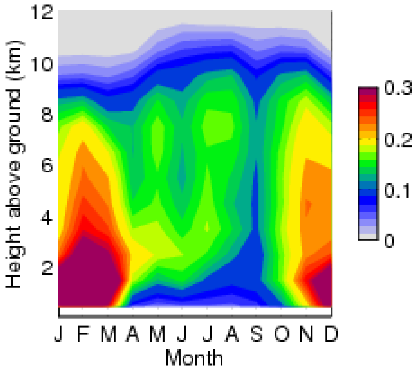
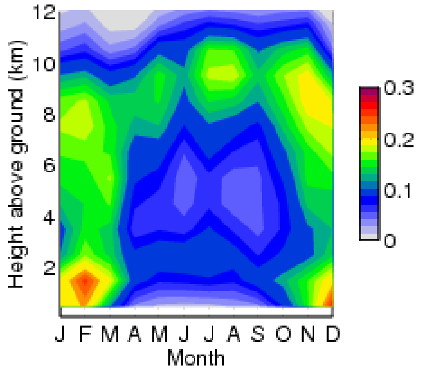
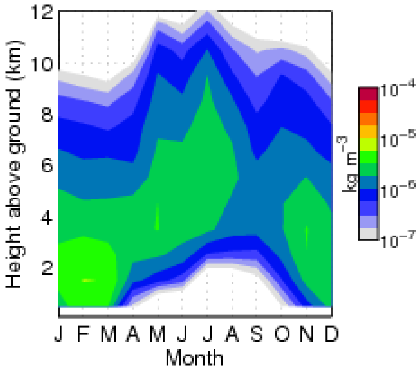
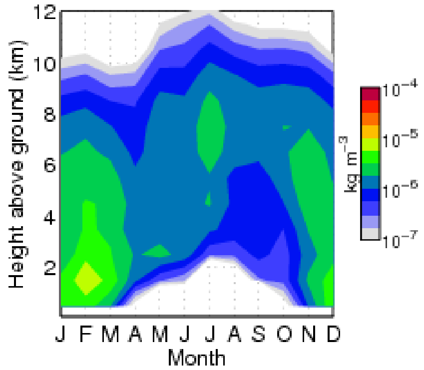
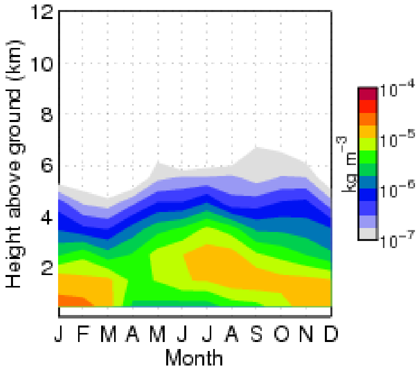
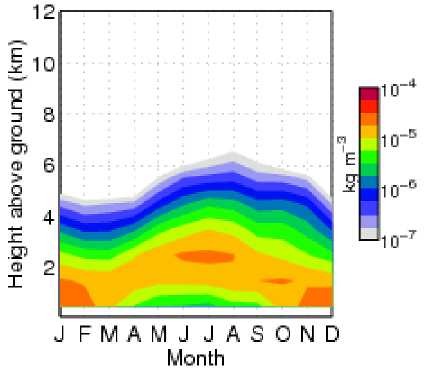
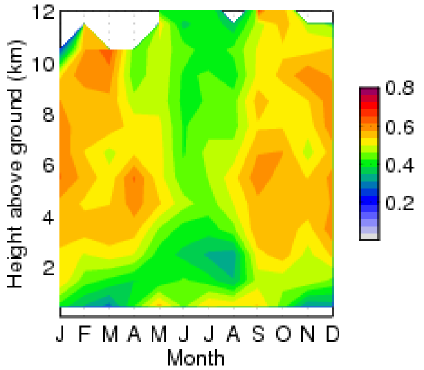
<table>
<tr>
<td>
</td>
<td>
</td>
<td>Cloud fraction</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td>Ice water content</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td>Liquid water content</td>
</tr>
<tr>
<td></td>
<td>
</td>
<td>Cloud fraction skill score,
SEDI</td>
</tr>
<tr>
<td>Observations</td>
<td>Model</td>
<td>Variable</td>
</tr>
</table>

Figure 7: Examples of new metrics. Top 3 panels: observed (left) and modeled (right) seasonal composites of cloud fraction, ice water content and liquid water content. Bottom panel: seasonal composite of the model skill in forecasting cloud fraction. The site is Lindenberg, Germany and the model is ECMWF.

Model intercomparisons

Additional metrics that have been implemented include timeseries of skill-score intercomparisons (Fig. 8), which can be used to display seasonal/annual variability in predictability. In other words, some years are easier to forecast than others. These (and standard metrics) are also routinely calculated for all available sites and models, although not necessarily in real time.

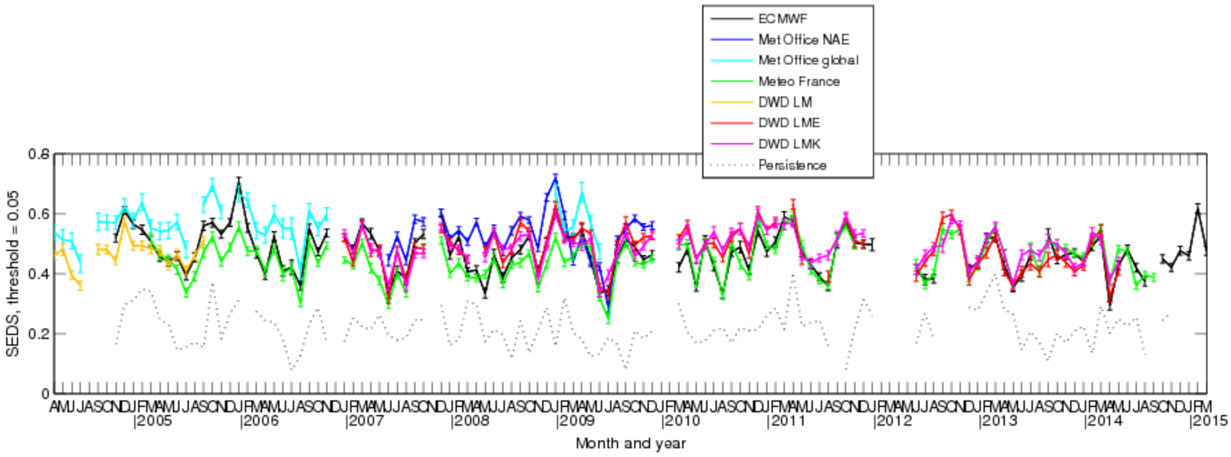


Figure 8. Timeseries of cloud fraction skill score SEDS for 7 NWP models at Lindenberg, Germany, over an 11 year period (2004-2015). A persistence skill score is also included.

Engagement with users

There is continual feedback from the users of these products, namely those responsible for microphysical parametrization schemes in operational model centres, with discussion helping to define metrics capable of judging model performance, and evolving or improving metrics to answer specific questions and identify model shortcomings. The most recent meeting, in December 2014, resulted in requests by modellers for rapid feedback on new trial model versions, and new avenues for model evaluation including ranking of regime by contribution to model bias.

Conclusions

A suite of comprehensive metrics for evaluating the representation of clouds in models has been implemented. The implementation is automated, covering the entire production chain from processing observations, through to evaluating model data and the generation of statistical metrics and figures, and can operate in near-real-time, providing updates continuously to the web. Incoming new observational and model datasets are identified, and all evaluation metrics are then created. The definition of new datasets can include near realtime daily updates, a month/year of batch-processed model data, or reprocessed data from a previous period. The procedures identify and update any current metrics that are potentially dependent on the new data; this could include a yearly metric being updated due to the addition of a new month of data, or a seasonal composite being updated due to the addition of a new year of data. Currently, the model evaluation operates in near-real-time for specific site-model combinations only; the limitation is due to the necessity of input data being provided in near-real-time.

The evaluation scheme has been extended to cover multiple versions of the same model, where model versions may refer to:

- operational models run at different resolutions – global, regional, local-area versions
- operational model versus new updated pre-release version

- test runs with different microphysics (prior to pre-release)

The long-term model evaluation has already elucidated many features requiring improvements, but has also shown that models already provide a 'reasonable' forecast of many cloud properties. The outlook for the future is for Cloudnet to provide rapid evaluation feedback for model test-runs, this requires the ability to perform arbitrary model evaluation, as long as the model data is uploaded in a specific format.

References

Bony, S., and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models, *Geophys. Res. Lett*., **32**, L20806, doi:10.1029/2005GL023851.

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

Hogan, R. J., E. J. O'Connor and A. J. Illingworth, 2009: Verification of cloud-fraction forecasts. *Q. J. R. Meteorol*. Soc., **135**, 1494-1511.

Hogan, R. J., C. A. T. Ferro, I. T. Jolliffe and D. B. Stephenson, 2010: Equitability revisited: Why the 'equitable threat score' is not equitable. *Weath. Forecasting*, **25**, 710-726.

Illingworth, A. J., R. J. Hogan, E. J. O'Connor et al., 2007: Cloudnet – continuous evaluation of cloud profiles in seven operational models using ground-based observations. *Bull. Am. Meteorol. Soc.*, **88**, 883-898.

Morcrette, C. J., O'Connor, E. J. and Petch, J. C., 2012: Evaluation of two cloud parametrization schemes using ARM and Cloud-Net observations. *Q.J.R. Meteorol. Soc.*, **138**, 964–979.