Using lidar and radar to evaluate cloud forecasts in operational models.

Anthony Illingworth,
+ Robin Hogan, Ewan O’Connor, U of Reading, UK
and many others at the U of Reading and in the CloudNET team

DTC Workshop, Boulder, 2-4 Nov 2010
Plan of Talk

1. Using ground based radar and lidar.  
   (mean and pdf of clouds properties:  
   vertical profiles of cloud fraction, iwc, lwc)

2. Skill scores - the right cloud in the right place.  
   ETS? New skills score metrics.

3. ARM sites for model evaluation.

4. First results from ‘A train’: Cloudsat and Calipso
How skillful is a forecast?

- Most model evaluations of clouds test the cloud climatology.
  - What about individual forecasts?
- Standard measure shows ECMWF forecast “half-life” of ~6 days in 1980 and ~9 days in 2000.
  - But virtually insensitive to clouds!
• Cloud is noisier than geopotential height $Z$ because it is separated by around two orders of differentiation:
  - Cloud ~ vertical wind ~ relative vorticity ~ $\nabla^2$streamfunction ~ $\nabla^2$pressure
  - Suggests cloud observations should be used routinely to evaluate models
The Cloudnet methodology

Recently completed EU project; www.cloud-net.org

BAMS Article June 2007

• Aim: to retrieve and evaluate the crucial cloud variables in forecast and climate models using radar and lidar.
  - **Models:** Met Office (4-km, 12-km and global), ECMWF, Météo-France, KNMI RACMO, Swedish RCA model, DWD
  - **Variables:** target classification, cloud fraction, liquid water content, ice water content, drizzle rate, mean drizzle drop size, ice effective radius, TKE dissipation rate
  - **Sites:** 4 Cloudnet sites in Europe, 6 ARM including 3 for mobile facility
  - **Period:** Several years near-continuous data from each site

• Crucial aspects
  - **Common formats** (including errors & data quality flags) allow all algorithms to be applied at all sites to evaluate all models
  - Evaluate for months and years: avoid unrepresentative case studies

April 2011 onwards - more European sites, Italy, Ireland
Cloud profiling sites

Core instrumentation at each site
Cloud radar, lidar, microwave radiometers, raingauge

Physical Map of the World, June 2003
Standard CloudNET observations (e.g. Chilbolton)

Radar

Lidar, gauge, radiometers

But can the average user make sense of these measurements?
First step: target classification

- Combining radar, lidar and model allows the type of cloud (or other target) to be identified.
- From this can calculate cloud fraction in each model gridbox.
Cloud fraction in 7 models

- Mean & PDF for 2004 for Chilbolton, Paris and Cabauw

- Uncertain above 7 km as must remove undetectable clouds in model

- All models except DWD underestimate mid-level cloud; some have separate “radiatively inactive” snow (ECMWF, DWD); Met Office has combined ice and snow but still underestimates cloud fraction

- Wide range of low cloud amounts in models

- Not enough overcast boxes.

*Illingworth et al, BAMS, 2007*
**Ice water content**

- IWC estimated from radar reflectivity and temperature
  - Rain events excluded from comparison due to mm-wave attenuation
  - For IWC above rain, use cm-wave radar (e.g. Hogan et al., JAM, 2006)

- ECMWF and Met Office within the observational errors at all heights
- Encouraging: AMIP implied an error of a factor of 10!
- DWD (pink) far too low

- Be careful in interpretation: mean IWC dominated by occasional large values so PDF more relevant for radiative properties
- DWD (pink) pdf best – apart from max bin – so mean value worst.
A change to Meteo-France cloud scheme

- Compare cloud fraction to observations before and after April 2003
- Note that cloud fraction and water content are entirely diagnostic

But human obs. indicate model now underestimates mean cloud-cover!

Compensation of errors: overlap scheme changed from random to maximum-random

(a) Mean cloud fraction
(b) Cloud fraction

Before April 2003
After April 2003
Equitable threat score

- ETS is a widely used skill score
  - 1 = perfect forecast, 0 = random forecast

• Measure of the skill of forecasting cloud fraction > 0.05
  - Assesses the weather of the model not its climate
  - Persistence forecast is shown for comparison

• Lower skill in summer convective events
• Met Office global and mesoscale – equally good.
Contingency tables

For given set of observed events, only 2 degrees of freedom in all possible forecasts (e.g. \(a\) & \(b\)), because 2 quantities fixed:
- Number of events that occurred \(n = a + b + c + d\)
- Base rate (observed frequency of occurrence) \(p = (a + c)/n\)
Equitable Threat Score

- Number of events that occurred \( n = a + b + c + d \)
- Base rate (observed frequency of occurrence) \( p = (a + c)/n \)

Expectation value of \( a = E(a) = \text{total number of modelled cloud} \times p \)
\[ E(a) = (a+b)p = (a+b)(a+c)/n \]

ETS = \( \frac{\text{cloud hit} - E(\text{cloud hit})}{\{(\text{Obs or Mod cloud}) - E(\text{cloud hit})\}} \)
ETS = \( \frac{a - E(a)}{(a+b+c) - E(a)} \)
(but not meaningful as \( p \to 0 \))

Heidke Skill Score, given by \( ETS = HSS/(2-HSS) \);
(is unconditionally equitable, but not meaningful as \( p \to 0 \))

Log of Odds Ratio: \( \text{LOR} = \ln (ad/bc) \) (perfect forecast scores \( \infty \))
Desirable properties of verification measures

1. “Equitable”: all random forecasts receive expected score zero
   - Constant forecasts of occurrence or non-occurrence also score zero
   - Note that forecasting the right cloud climatology versus height but
     with no other skill should also score zero

2. Useful for rare events
   - Almost all measures are “degenerate” in that they asymptote to 0 or 1
     for vanishingly rare events

Extreme dependency score
- Stephenson et al. (2008) explained this behavior:
  - Almost all scores have a meaningless limit as “base rate” $p \to 0$
  - HSS tends to zero and LOR tends to infinity
- They proposed the Extreme Dependency Score:
  \[
  \text{EDS} = \frac{2 \ln[(a + c)/n]}{\ln(a/n)} - 1
  \]
  - where $n = a + b + c + d$
- It can be shown that this score tends to a meaningful limit:
Symmetric extreme dependency score

- EDS problems:
  - Easy to hedge (unless calibrated) (e.g. predict clouds all the time)
  - Not equitable

- Solved by defining a symmetric version:
  (same score if swap models and obs)
  - All the benefits of EDS, none of the drawbacks!

Solved by defining a symmetric version:
(same score if swap models and obs)
- All the benefits of EDS, none of the drawbacks!

\[
\text{SED} = \frac{\ln[(a + c)/n]}{\ln(a/n)} - 1
\]

\[
\sigma = \sqrt{\left\{ \frac{1}{a} - \frac{1}{a+c} \right\}} \left\{ -\ln[E(a)/n]/\ln(a/n)^2 \right\}
\]

Hogan, O’Connor and Illingworth (2009 QJRMS)
Skill versus height

- Most scores not reliable near the tropopause because cloud fraction tends to zero

- New score reveals:
  - Skill tends to slowly decrease at tropopause
  - Mid-level clouds (4-5 km) most skilfully predicted, particularly by Met Office
  - Boundary-layer clouds least skilfully predicted
Skill versus cloud-fraction threshold

- Consider 7 models evaluated over 3 European sites in 2003-2004

LOR implies skill increases for larger cloud-fraction threshold

HSS implies skill decreases significantly for larger cloud-fraction threshold
Skill versus cloud-fraction threshold

SEDS has much flatter behaviour for all models (except for Met Office which underestimates high cloud occurrence significantly)
Skill versus lead time

- Only possible for UK Met Office 12-km model and German DWD 7-km model
  - Steady decrease of skill with lead time
  - Both models appear to improve between 2004 and 2007
- Generally, UK model best over UK, German best over Germany
  - An exception is Murgtal in 2007 (Met Office model wins)
Key properties for estimating $\frac{1}{2}$ life

- We wish to model the score $S$ versus forecast lead time $t$ as:
  \[ S(t) = S_0 e^{-t/\tau} = S_0 \times 2^{-t/\tau_{1/2}} \]
  - where $\tau_{1/2}$ is forecast "half-life"

- We need linearity
  - Some measures "saturate" at high skill end (e.g. Yule's Q / ORSS)
  - Leads to misleadingly long half-life

- ...and equitability
  - The formula above assumes that score tends to zero for very long forecasts, which will only occur if the measure is equitable
Reality \((n=16, \ p=1/4)\)

Forecast

- **Best possible forecast**
  - 4 0
  - 0 12

- **Random unbiased forecast**
  - 1 3
  - 3 9

**SEDs has property of linearity**

**SED score 1**

**SED score decreases nearly linearly**
**Forecast “half life”**

- Fit an inverse-exponential: \( S(t) = S_0 \times 2^{-t/\tau_{1/2}} \)
  - \( S_0 \) is the initial score and \( \tau_{1/2} \) is the half-life

- Noticeably longer half-life fitted after 36 hours
  - Same thing found for Met Office rainfall forecast (Roberts 2008)
  - First timescale due to data assimilation and convective events
  - Second due to more predictable large-scale weather systems
What is the origin of the term “ETS”?  

- First use of “Equitable Threat Score”: Mesinger & Black (1992)
  - A modification of the “Threat Score” \( a/(a+b+c) \)
  - They cited Gandin and Murphy’s equitability requirement that constant forecasts score zero (which ETS does) although it doesn’t satisfy requirement that non-constant random forecasts have expected score 0
  - ETS now one of most widely used verification measures in meteorology

- An example of rediscovery
  - Gilbert (1884) discussed \( a/(a+b+c) \) as a possible verification measure in the context of Finley’s (1884) tornado forecasts
  - Gilbert noted deficiencies of this and also proposed exactly the same formula as ETS, 108 years before!

- Suggest that ETS is referred to as the Gilbert Skill Score (GSS)
  - Or use the Heidke Skill Score, which is unconditionally equitable and is uniquely related to ETS = HSS / (2 – HSS)

_Hogan, Ferro, Jolliffe and Stephenson (WAF)_
US Dept of Energy Climate Change Prediction Program recently funded 5-year consortium project centred at Brookhaven, NY
  - Implement updated Cloudnet processing system at Atmospheric Radiation Measurement (ARM) radar-lidar sites worldwide
  - Ingests ARM’s cloud boundary diagnosis, but uses Cloudnet for stats
  - New diagnostics being tested

Testing of NWP models
  - NCEP, ECMWF, Met Office, Meteo-France...
  - Over a decade of data at several sites: have cloud forecasts improved over this time?
Figure 3. Comparison of the seasonal composites of cloud fraction derived from observations (a) at the ARM SGP site for the years 2004 to 2009 with the values held in the ECMWF model (b), NCEP model (c) and the global version of the Met Office model (d).
Figure 4. Seasonal composites of the skill score SEDS at the ARM SGP site for the years 2004 to 2009 for ERA Interim (a), ECMWF (b), NCEP (c) and the global version of the Met Office model (d).
Figure 5. Same as Fig. 3 except for ice water content.
a) Obs  ARM SPG  b) ECMWF  c) NCEP  d) Met Office
Figure 6. Same as Fig. 3 except for liquid water content.

a) Obs ARM SPG  b) ECMWF  c) NCEP  d) Met Office
DARWIN: FIRST RESULTS CLOUD FRACTION

ECMWF MODEL

OBSERVED
SGP - cloud fraction -skill 2004-2010: ECMWF ‘ERA’ as ref.

And SEDS skill (with error) as a function of lead time?
The A-Train
- NASA
- 700-km orbit
- CloudSat 94-GHz radar (launch 2006)
- Calipso 532/1064-nm depol. lidar
- MODIS multi-wavelength radiometer
- CERES broad-band radiometer
- AMSR-E microwave radiometer

EarthCARE (launch 2015)
- ESA+JAXA
- 400-km orbit: more sensitive
- 94-GHz Doppler radar
- 355-nm HSRL/depol. lidar
- Multispectral imager
- Broad-band radiometer
- Heart-warming name
Spaceborne Overview

• What do spaceborne radar and lidar see?
• Towards a “unified” retrieval of ice clouds, liquid clouds, precipitation and aerosol
  - Variational retrieval framework - GIVES ERRORS
• Results from CloudSat-Calipso ice-cloud retrieval
  - Consistency with top-of-atmosphere radiative fluxes
  - Evaluation and improvement of models
  - Spatial structure of mid-latitude and tropical cirrus
• Test new version of ECMWF model which has prognostic cloud fraction, iwc and lwc. Current operational model has only prognostic total water content: iwc/lwc ratio diagnosed from temperature.
What do CloudSat and Calipso see?

- **Radar**: $\sim D^6$, detects whole profile, surface echo provides integral constraint.
- **Lidar**: $\sim D^2$, more sensitive to thin cirrus and liquid clouds but attenuated.

Delanoe and Hogan (2008, 2010)
Ingredients of a variational retrieval

- **Aim:** to retrieve an optimal estimate of the properties of clouds, aerosols and precipitation from combining these measurements
  - To make use of integral constraints must retrieve components *together*
- **For each ray of data,** define *observation vector* \( y \):
  - Radar reflectivity values
  - Lidar backscatter values
  - Infrared radiances
  - Shortwave radiances
  - Surface radar echo (provides two-way attenuation)
- **Define *state vector* \( x \) of properties to be retrieved:**
  - Ice cloud extinction, number concentration and lidar-ratio profile
  - Liquid water content profile and number concentration
  - Rain rate profile and number concentration
  - Aerosol extinction coefficient profile and lidar ratio
- **Forward model** \( H(x) \) to predict the observations
  - Microphysical component: particle scattering properties
  - Radiative transfer component
The cost function

The essence of the method is to find the state vector $\mathbf{x}$ that minimizes a cost function:

$$J = \sum_{i=1}^{n_y} \frac{[y_i - H(\mathbf{x})]^2}{\sigma_{y_i}^2} + \sum_{i=1}^{n_x} \frac{[x_i - b_i]^2}{\sigma_{b_i}^2} + \text{Smoothness constraints}$$

- The forward model $H(\mathbf{x})$ predicts the observations from the state vector $\mathbf{x}$.
- Some elements of $\mathbf{x}$ are constrained by a prior estimate.
- Each observation $y_i$ is weighted by the inverse of its error variance.
- This term can be used to penalize curvature in the retrieved profile.
Lidar observations
Radar observations
Visible extinction
Ice water content
Effective radius
Lidar forward model
Radar forward model
Example ice cloud retrievals
Delanoe and Hogan (2010)
Evaluation using CERES TOA fluxes

- **Radar-lidar retrieved profiles** containing only ice used with Edwards-Slingo radiation code to predict CERES fluxes
- Small biases but large random shortwave error: 3D effects?

**Shortwave**
- Bias 4 W m\(^{-2}\), RMSE 71 W m\(^{-2}\)

**Longwave**
- Bias 0.3 W m\(^{-2}\), RMSE 14 W m\(^{-2}\)

Nicky Chalmers
A-Train versus models

- Ice water content
- 14 July 2006
- Half an orbit
- 150° longitude at equator

Delanoe et al. (2010)
Both models lack high thin cirrus.

- Met Office has too narrow a distribution of in-cloud IWC.
- ECMWF lacks high IWC values, remedied in new model version.
• Good agreement over six orders of magnitude.

• High iwp - models too low. Especially for operational ECMWF.

• New ECMWF model with prognostic iwc is much better than current operational model which has iwc/lwc diagnosed from total cloud water using temperature.
Summary and outlook

• Model comparisons reveal:
  - Rapid feed back to NWP centres of performance of new cloud schemes.
    Example: Meteo-France 2003:
    Example: testing of new ECMWF scheme before use operationally.
  - In Europe, higher skill for mid-level cloud and lower for boundary-layer cloud, but larger seasonal contrast in Southern US

• Findings applicable to other verification problems:
  - “Symmetric Extreme Dependency Score” is a reliable measure of skill for both common and rare events (given we have large enough sample)
  - Many measures regarded as equitable are only so for very large samples, including the “Equitable Threat Score”, but they can be rescaled

• Future work (in addition to CCPP):
  - CloudSat & Calipso: all four years - annual cycle - regional skill.
  - Cloudnet - extend to ARM sites plus more sites in Europe.
  - EU-FP7 ‘Actris" infrastructure for rapid feedback to NWP modellers.