Observations use in data assimilation and verification: *Similar but not the same*

Marion Mittermaier

with contributions from my co-members of the WMO Joint Working Group on Forecast Verification Research (JWGFVR) and other Met Office staff

 ECMWF-JCSDA Workshop on assimilating satellite observations of clouds and precipitation in NWP models

Outline

1. Basic concepts of verification
2. Observations – a nasty business?!
3. DA vs verification
4. Using analyses for verification
5. Dealing with observations errors (in verification)
6. A role of satellite observations?
7. Conclusions and recommendations

Basic verification concepts

Why verify?

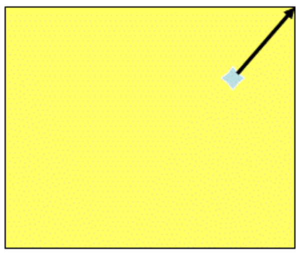
- **Administrative** purpose
  - Monitoring performance
  - Choice of model or model configuration (has the model improved?)
- **Scientific** purpose
  - Identifying and correcting model flaws
  - Forecast improvement
- **Economic** purpose
  - Improved decision making
  - "Feeding" decision models or decision support systems

JWGFVR training notes

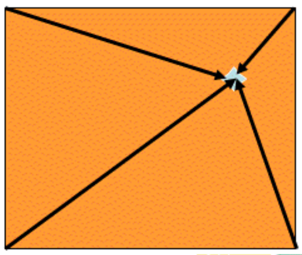
Matching forecasts and observations

- Point-to-grid and

  grid-to-point



- Matching approach can impact the results of the verification

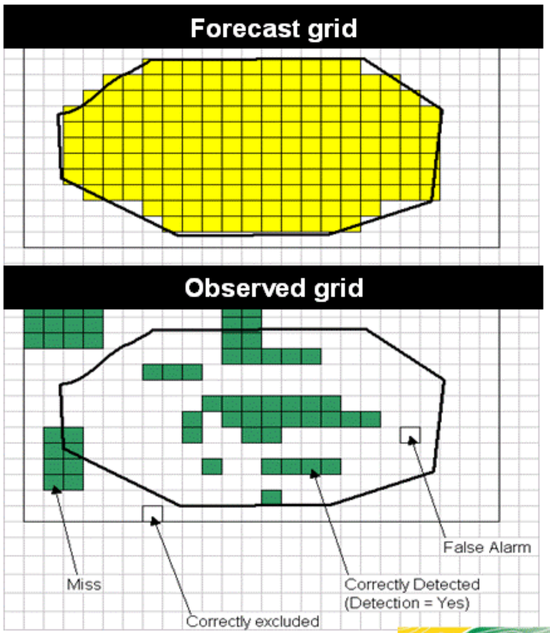


JWGFVR training notes

Met Office

Matching forecasts and observations

- Grid-to-grid approach
  - Overlay forecast and observed grids
  - Match each forecast and observation

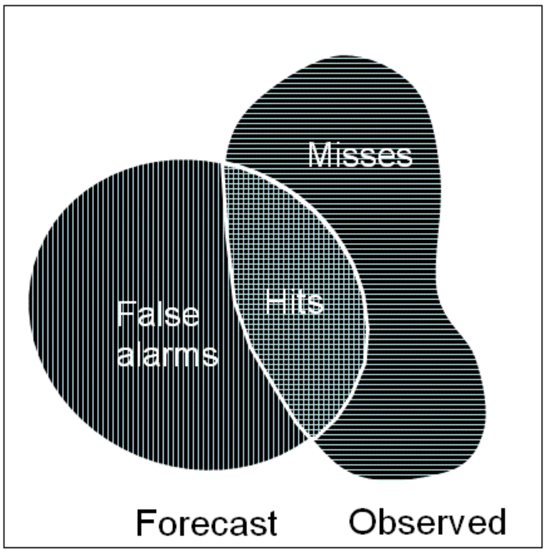


JWGFVR training notes

Met Office

Traditional spatial verification using categorical scores

Compute statistics on forecast-observation pairs



		Observed: Yes	Observed: no
Predicted	yes	*hits*	*false alarms*
	no	*misses*	*correct negatives*

$$POD = \frac{hits}{hits + misses} \quad FBI = \frac{hits + false\ alarms}{hits + misses}$$

$$FAR = \frac{false\ alarms}{hits + false\ alarms}$$

$$TS = \frac{hits}{hits + misses + false\ alarms}$$

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

JWGFVR training notes

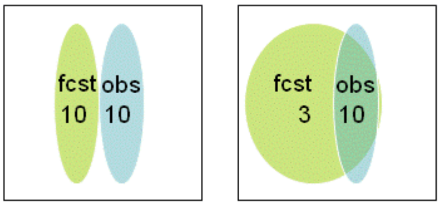
Met Office

Traditional spatial verification

- **Requires an exact match** between forecasts and observations at every grid point.

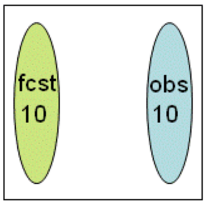
  - Problem of **"double penalty"** - event predicted where it did not occur, no event predicted where it did occur

- Traditional scores do not say very much about the source or nature of the errors



Hi res forecast
RMS ~ 4.7
POD=0, FAR=1
TS=0

Low res forecast
RMS ~ 2.7
POD~1, FAR~0.7
TS~0.3



JWGFVR training notes

Met Office

How parameter characteristics dictate the metrics

Precipitation

- **Positively bounded** quantity approximately log-normally distributed
- **Variety of sources**: gauges, radar, satellite
- **Highly discontinuous** in space and time, possibly sparse; difficult to verify due to potentially large space-time errors.
- Continuous metrics (e.g. rmse) not recommended
- Focus on rain areas, thresholds, spatial methods

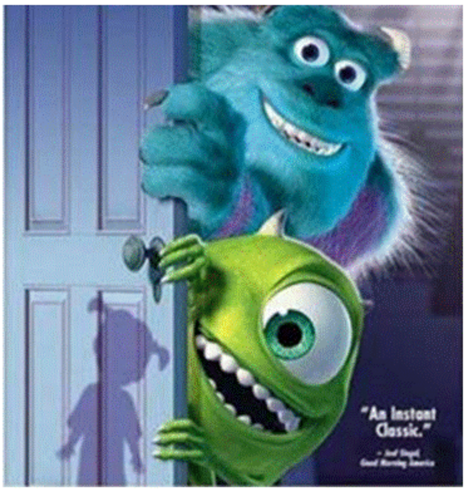
Cloud

- Cloud cover
  - **Bounded** (cloud fraction 0-1) but mostly discretised (0-8 okta)
  - **Complex 3-D structure** with discrete structures in space and time, usually simplified into total cloud amount (TCA)
  - Continuous metrics not recommended, ideally suited to 3 x 3 categorical contingency analyses.
- Radiances
  - Continuous parameter which could be assessed using continuous, categorical or spatial methods.

Observations

The monster(s) in the closet…

- **In attempting to assess model forecast skill, what are we losing/risking by ignoring observation uncertainty?**
- **What can we gain** by considering it?
  - (Confusion?)
- Can we afford to ignore it?
  - No!



Courtesy, Barb Brown

Observations are *NOT* perfect!

- **Observations error** vs **predictability** and forecast error/uncertainty
- **Different observation types** of the same parameter (manual or automated) can impact results
- Typical **instrument errors** are:
  - *For temperature: +/- 0.1°C*
  - *For wind speed: speed dependent errors but ~ +/- 0.5 m/s*
  - *For precipitation (gauges): +/- 0.1 mm (half tip) but 2 -- 50%*
  - *For cloud cover: ???*
- Then there are **further issues of shielding/exposure** etc
- In some instances "forecast" errors are very similar to instrument limits – *so, should the forecast get the blame?*

Met Office

Sources of error and uncertainty

- Biases in frequency or value ✓
- **Instrument error** ?
- Random error or noise ✓
- Reporting errors ✓
- **Reporting of errors** ?
- **Subjective obs** (e.g., impact-based observations) ?
- Representativeness error ✓
- Precision error ✓
- **Conversion/transformation error** ? ✓
- **Analysis error** ? ✓
- Other?

Met Office

Effects of observation errors

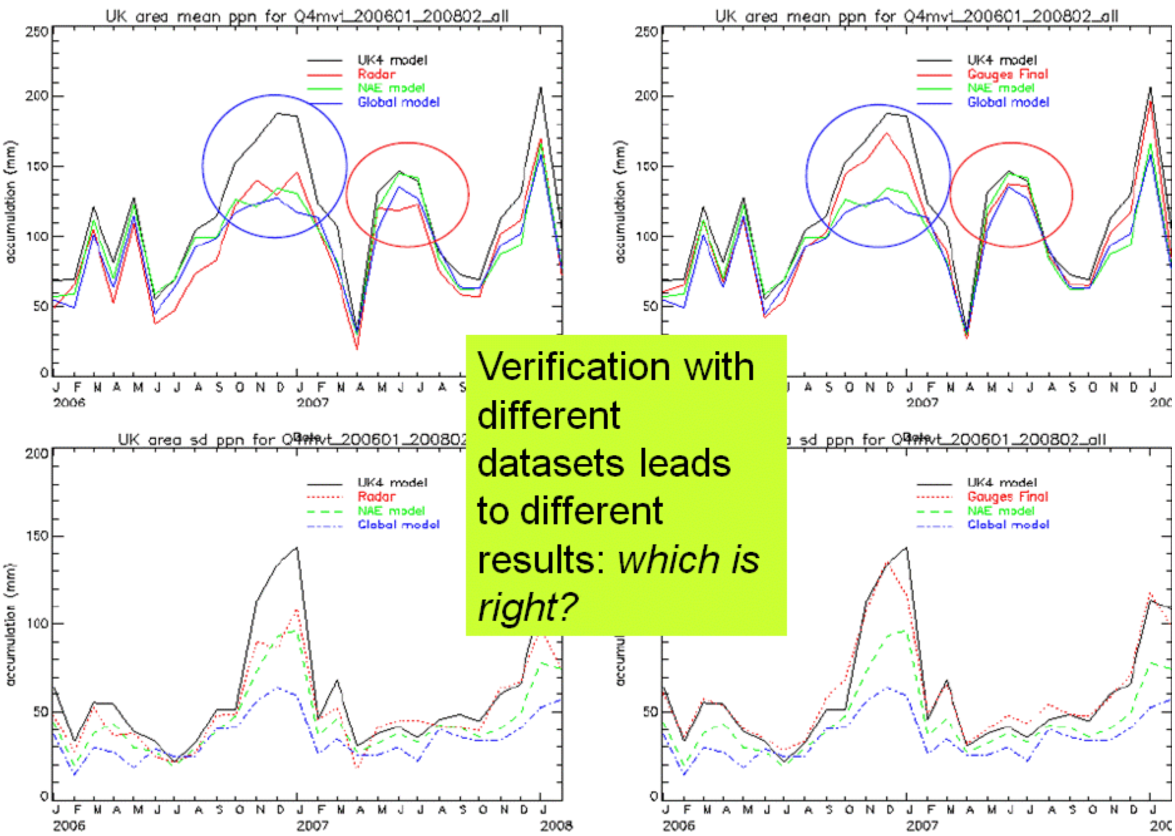
- Observation errors add uncertainty to the verification results
  - *True forecast skill is unknown (an imperfect model / ensemble may score better!)*
  - *Extra dispersion of observation PDF*
- Effects on verification results
  - *RMSE – overestimated*
  - *Spread – more ob outliers make ensemble look under-dispersed*
  - *Reliability – poorer*
  - *Resolution – greater in BS decomposition, but ROC area poorer*
  - *CRPS – poorer mean values*
- Can we remove the effects of observation error?
- More samples help with reliability estimates
- **Quantify actual observation errors as far as possible**

The pitfalls of observations type

Fact sheet

Met Office

- Manual surface observations are a “dying breed”.
- Using sparse and irregularly distributed observations for verifying high-resolution models leads to potentially disappointing results. *“Where is the benefit of high-resolution?”*
- **Cloud and precipitation are two of the most difficult parameters to predict accurately**, yet the impact of cloud biases (in particular) have huge **knock-on effects** on other parameters, such as temperature.
- Using different observation types for verifying the same model parameter will give different results. *[How does one deal with this?]*

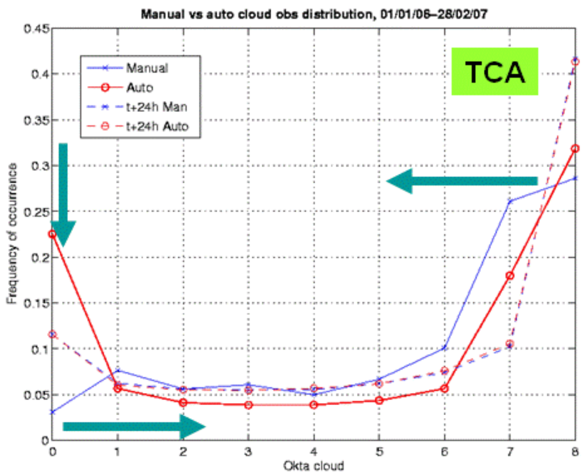


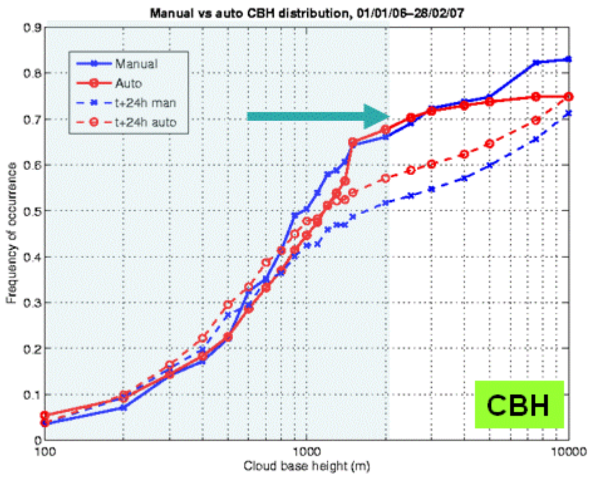
Courtesy of Clive Wilson

Met Office

TCA and CBH distributions

- 14 months of data for Block 03 stations
- Auto obs have greater proportion of no cloud (due to instrument limitations, can't see high cloud)
- Observers hedge away from the "boundaries".
- For CBH artificial cloud ceiling visible in cdf

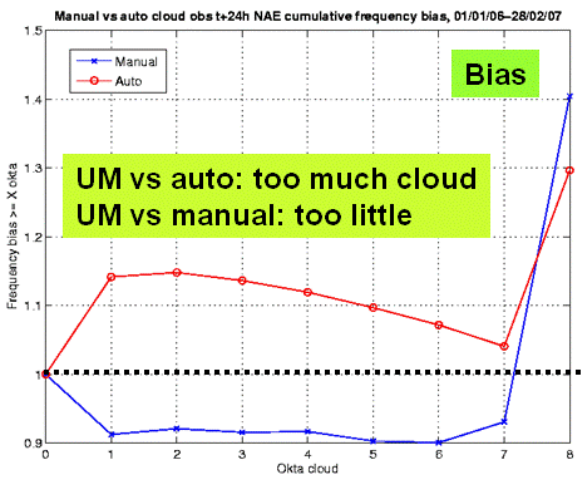


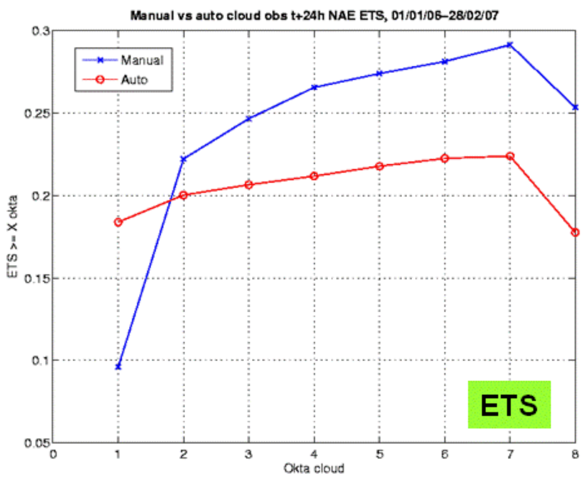


Met Office

How does ob type affect verification measures?

In the UM we discovered that use of manual and auto TCA leads to biases of equal but opposite magnitudes.





DA vs verification

Observations treatment

- DA and verification **both require observations** *BUT* the type, treatment, temporal resolution of observations used may be quite different.
- **Verification (in near real-time) relies heavily on the obs QC that DA provides**, using assigned flags to determine whether an ob is safe to use (other non-DA based obs QC takes a lot longer)
- **Independent observations analysis systems** (that do not rely on model background checking) are rarely available.

Observations treatment in DA

1. Observations received, check whether in time window, unit conversion and re-mapping
2. QC – **“probability of gross error”**
   - Updating of “reject lists”
   - Background checking (O-B) and buddy checking etc
   - Update obs QC flags
3. Data thinning for satellite obs (in both space and time) – all satellite obs tend to be QC’d

Impacts of observations handling

Met Office

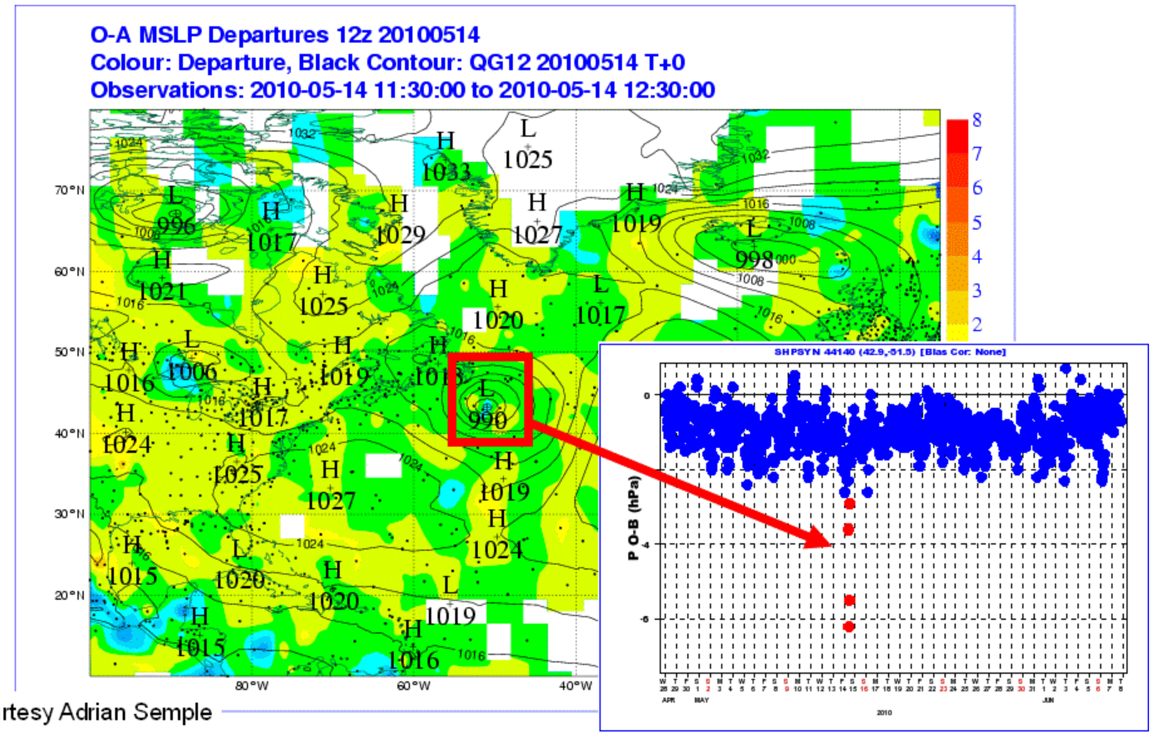
DA

- ✓ **Error tolerant** but sensitive to gross errors
- ✓ O-B at observation time
- ✓ PGE different for each model so observation sets may differ
- ✓ DA is run at **coarser** resolution than the forecast
- ✓ Linear model assumptions and interpolation methods
- ✓ Error inflation
- ✓ Thinning results in a self-selecting partial non-random sample

Verification

- ✗ **Error intolerant**, dependent on DA QC flags
- ✓ F-A at validity time
- ✗ Want the same obs for comparison of different models
- ✓ Forecast models are at **finer** resolution
- ✓ Impacts the QC flags so good observations may be rejected
- ✗ Forecast skill under-estimated
- ✓ Issues with non-independence

Data filtering for assimilation and QC



Courtesy Adrian Semple

Case study: OPERA European radar composite

- Two strands:
  - Data assimilation
  - Verification

- ***Stage 1: establish quality and advise on usefulness, make suggestions for improvements***

Region	OPERA vs Operational	No Precip vs Operational
NAE	-0.02%	+0.01%
Mes	-0.13%	-0.13%
UK Index List	-0.42%	-0.69%
WMO block 3	-0.27%	-0.32%
Scandanavia	-0.07%	0.0%
France	-0.26%	-0.02%
Iberia	+0.85%	+1.60%
Germany	-0.29%	-0.05%
Central Europe	-0.21%	-0.13%
Eastern Europe	+0.21%	+0.29%

Negative impact over UK –
OPERA degraded product compared to Nimrod

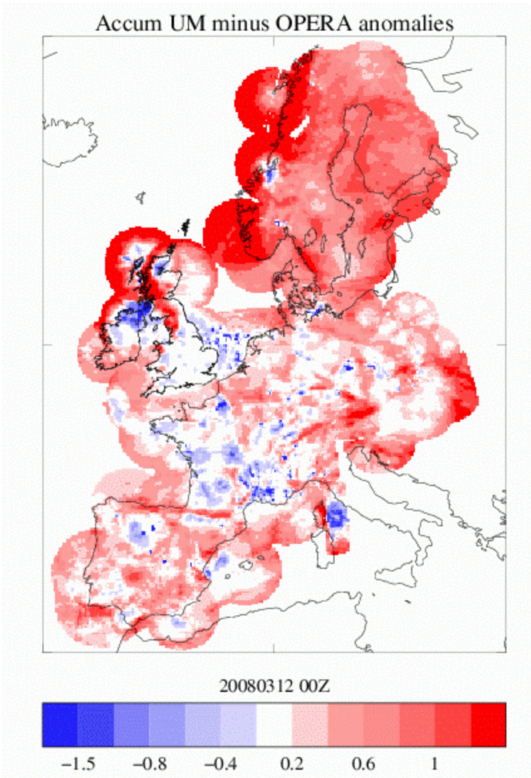
Negative impact over France –
OPERA is degraded product compared to Oper

Positive impact over Spain and Eastern Europe -
OPERA represents additional info available here

From Mittermaier et al, 2008

OPERA anomalies

- **Use the model forecasts as truth** to consider observations inconsistencies and errors.
- ***35 days*** accumulated normalised anomalies
- Computed from detrended model forecasts and OPERA accumulations.
- Pick out areas of:
  - ***Range problems*** and ***cold season*** bias
  - ***Anaprop***
  - ***Bright band***



From Mittermaier et al, 2008

The dreaded "verifying analysis"

Analyses: different flavours

Met Office

- **Forecast analysis**: here the purpose is provide the best estimate of the atmospheric state for the *model to produce the best possible forecast sequence*.
- **Observations analysis**: here the objective is to *match the observations as precisely as possible* to produce the best possible high-resolution estimate of the current atmospheric state. No forecast is produced from this. Variational and statistical techniques are used, but the use of model background fields is optional.
- **Re-analysis**: here the desire is to fix the method for creating the analysis, and produce a *retrospective dataset of analyses* which are used for model re-runs (of old case studies) and validation.

Why do we want to use gridded analyses for verification?

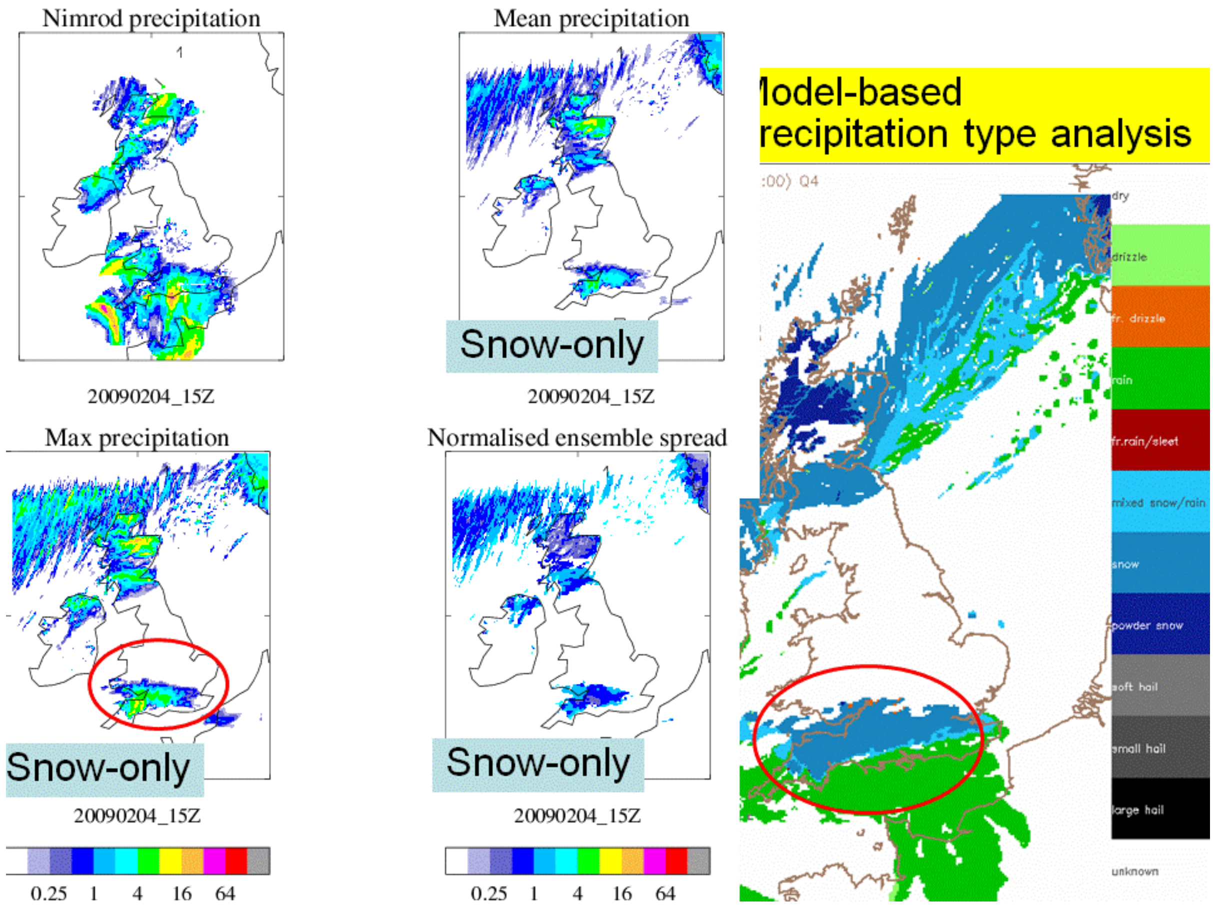
Met Office

- **Sample size and coverage** – get the "bigger picture"
- **Ease of use** – "hides" the observations, QC process has been done, consistent etc
- **Availability** - most created as part of the forecast process
- **Improved sampling of spatially discontinuous parameters** e.g. cloud and precipitation
- High-resolution models suffer from poor verification results when compared at isolated points

Met Office

The issues with using analyses

- **Non-independence**, adjacent grid points correlated in space and time. This reduces the degrees of freedom of verifying sample.
- **Local effects** not always well captured, or too much local (spurious?) detail – **resolution**
- **Method** - created as part of the forecast process. Need to verify the analysis, can only do this at observations locations. Even so, **is this form of "truth" accurate elsewhere?** How does one know? **Need for cross-validation**; impact of **observations denial**?



Dealing with observations errors in verification

Approaches for coping with observational uncertainty

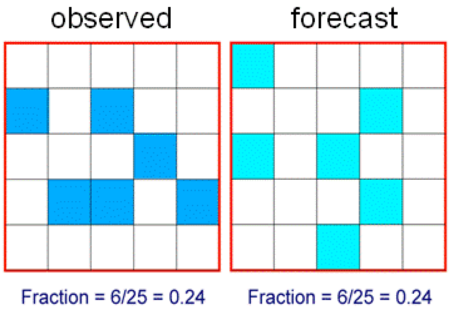
- **Indirect estimation** of observations uncertainties through verification approaches
- Incorporation of uncertainty information into verification metrics and **developing new methods that lessen the impact** (e.g. Roberts and Lean MWR, 2008, ICP special collection in WF)
- **Treat observations as probabilistic** (e.g. Candille and Talagrand)
- **Assimilation** approaches
- **Perturbing ensemble members** with observation error

Direct approaches for coping with observational uncertainty

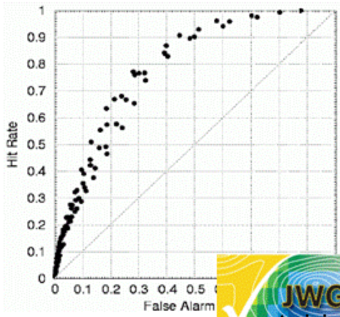
- Compare forecast error to known observation error. Can we be as simplistic as:
  - If forecast error is smaller than obs error then
    - ➢ A good forecast ✔
  - If forecast error is larger, then
    - ➢ A bad forecast ✘
- **What about testing improvements?** How can you know you are making the forecasts better when the improvement signal is in the “noise”?

Indirect approaches for coping with observational uncertainty

(Roberts and Lean, 2008)



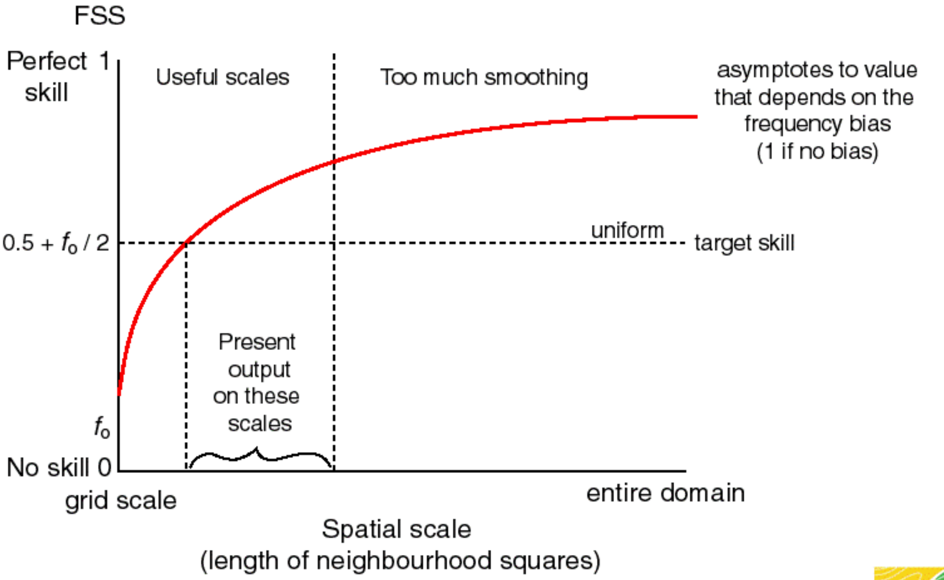
- **Neighbourhood** or fuzzy verification approaches
- Other spatial methods *(see the special collection in WF on the Inter-Comparison Project (ICP) of spatial verification methods)*



(Atger, 2001)

Fractions skill score

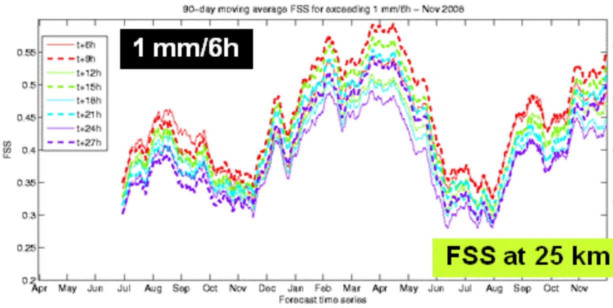
(Roberts and Lean, *MWR*, 2008)

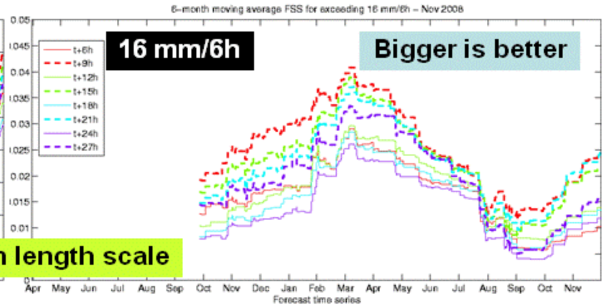


Met Office

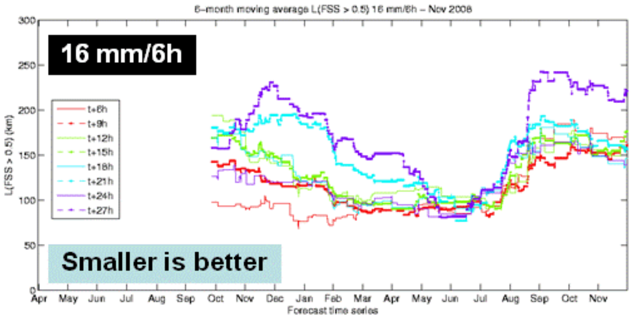
A “belt-and-braces” approach

Bundle up all sources of error, no direct attribution





- Is high-resolution (dashed) better than coarser resolution?
- Length scale which is skilful?

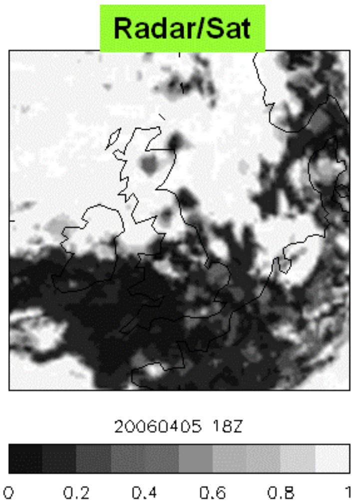


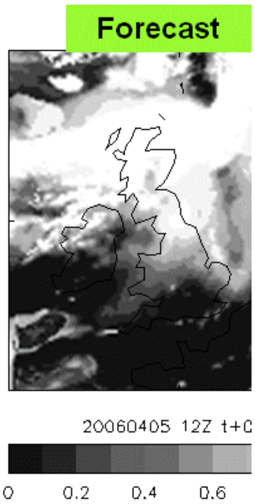
Remotely sensed cloud products: the way forward?

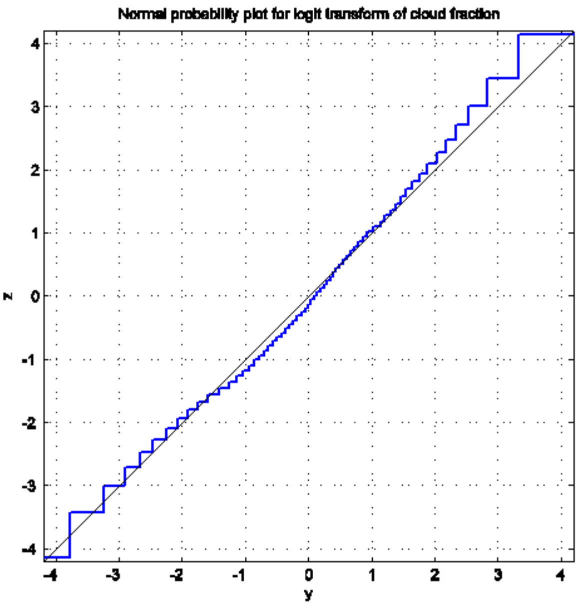
Met Office

Intensity-scale method

(Casati et al, 2004)







Met Office

Conclusions

1. Verification is much **more strongly dependent on the availability of quality observations**.

2. The **characteristics** of the (model) parameters and the observations required to assess them, **must be well understood** for verification, if the results are to be meaningful (i.e. assessing forecast skill).

3. **Interpretation of conflicting results** from different observation types present a considerable challenge and must be treated with care.

4. Increased horizontal (and vertical) model resolution necessitates a search for new verification data sources. **New data sources will require new verification tools and strategies. A LOT OF PROGRESS HAS BEEN MADE.**

5. **Satellites** may provide a useful dataset of remote cloud characteristics, both for the end user and the model physics developer.

Met Office

Strategic direction

- **The concept of using satellite observations needs to be proven to be computationally statistically viable.**
- **Spatial verification methods** need to be:
  - Used for routine verification of high-resolution precipitation forecasts, to prove that they are indeed getting better.
  - Proven for other variables, using analyses or gridded data sets.
- **Error sources** and magnitudes need to be better understood and quantified.
- **Prevent good observations from being rejected**! Investigate how observations are tagged.
- **Instigate best practice for data denial** to test credibility of analyses.
- Invest more in the development of **"independent" analyses**.
- **Generic uncertainty measures** need to be developed that can be sensibly incorporated into the standard routine verification processes.
- **Greater use of error bars and use of hypothesis testing** for assessing the impact of model changes.