Flow-dependent reliability: A practical route to more skilful ensemble forecasts

Mark Rodwell

With

David Richardson
Heini Wernli
Dave Parsons
Simon Lang … and many others

5th WGNE workshop on systematic errors in weather and climate models

21 June 2017
Montréal, Canada
Forecast reliability

In a “reliable” forecast system, the truth can be considered as another ensemble member. Reliability is very useful: if we predict an event with probability 70%, it will happen with frequency 70%. A testable consequence of reliability is that:

\[ \text{average Error} = \text{average Spread} \]

(averaged over many forecast start dates)

Given we had a reliable system, progress would be …

Predicting “sharper” (tighter) distributions while retaining reliability
In 1996, Spread \( \approx \) RMSE to D+2 due to Singular Vector (SV) initial perturbations. Since then, Error and Spread have converged and reduced throughout forecast range due to improved observations, model, representation of observation and model uncertainty, and introduction of Ensemble of Data Assimilations (EDA); allowing weaker SV perturbations.

… but uncertainty varies from day-to-day. The real reason we make ensemble forecasts. What causes this, and how can we evaluate it in our forecasts?

500 hPa geopotential height (Z500). "Error" is RMS of ensemble-mean error

Spread = ensemble standard deviation (scaled to account for finite ensemble size)
Potential Vorticity on the Potential Temperature = 320K surface. 20110410 0 UTC. Step (days, hours) = 0 00.0
Ensemble members start from very similar conditions. Differences account for our uncertainty in the truth and are almost imperceptible to the eye here. Differences then grow with lead-time and the members become completely different beyond about day 4.

Member 32 agrees well with the observed outcome. Simply a case of low predictability? How do we make progress?
Potential Vorticity on the Potential Temperature = 320K surface. 20110404 0 UTC. Step (days, hours) = 0 00.0
For this start-date, there were many more good forecasts.
Ensembles represent the integral of instantaneous uncertainty growth-rate; would like to diagnose deficiencies in the instantaneous growth-rate. Here consider 1h tendencies in EDA background forecasts. Filter to T21, 1d to emphasise growth-rate within synoptic-scale systems.

$\delta_{PV_t}$ used to highlight dynamical (e.g. baroclinic) and physical (e.g. stochastic) sources of uncertainty. 315K $\Rightarrow$ strong propagation in Jetstream.

Results emphasise role of moist processes (MCS, WCB); intrinsic to real system or artefact of e.g. deficiencies in model uncertainty? Key question.
Average initial conditions of 584 single forecast “busts” over Europe at day 6

Rodwell et al, 2013, BAMS

A Trough over the Rocky mountains, with high convective potential ahead
Conducive to the formation of mesoscale convection
Interesting flow regime to evaluate ‘instantaneous’ uncertainty growth-rate
(Using different set of dates to avoid misleading results)

‘CAPE’ = Convective Available Potential Energy
Enhanced uncertainty (EnsVar) around the Great Lakes / Mississippi Region, large ‘errors’ (Depar$^2$)
Observation uncertainty (ObsUnc$^2$) quite small so a statistically significant positive Residual ⇒
ENS does not inject enough uncertainty into global circulation. Forecasts will be too confident
Short-range mean error assessment for T300 in “trough/CAPE” situations using DA

54 cases

Model-space process tendency and analysis increment budget

(a) Dynamics  
(b) Radiation  
(c) Convection

(d) Cloud  
(e) Increment  
(f) Evolution

Likelihood of (mesoscale) convective heating, dynamical cooling
Cloud microphysics, ‘hole’ in radiative cooling
Statistically significant Increment implies flow-dependent model (or observation) bias

Dynamics + Radiation + Convection + Cloud + Increment ≈ Evolution
Reliability ⇒ E[Increment]=0

European Centre for Medium-Range Weather Forecasts  
Mark J Rodwell
If we don’t hit the string hard enough, the wave in the string will be too weak.

If we hit the string at the wrong time, the wave will arrive over Europe at the wrong time.

We do not know when to press the key (mesoscale convection itself involves chaotic uncertainty).

What we want is that the ensemble members generate such convection with the “right” uncertainty.
Short-range variance assessment for mid-tropospheric temperatures: Control

Relative to AMSUA channel 5 microwave brightness temperatures

(a) Depar$^2$

(b) Bias$^2$

(c) EnsVar

(d) ObsUnc$^2$

(e) Residual

(f) Observation density (2°x2°, 12h)

Residual suggests EnsVar underestimates (overestimates) uncertainty in convective (clear-sky) regions (and/or deficiencies in ObsUnc$^2$)
Short-range variance assessment for mid-tropospheric temperatures: No stoch. phys.

Experiments by Simon Lang

20110812-20111116, 2 members

Relative to AMSUA channel 5 microwave brightness temperatures

(a) Depar$^2$  (b) Bias$^2$  (c) EnsVar

(d) ObsUnc$^2$  (e) Residual  (f) Observation density (2°x2°, 12h)

A role for deterministic and stochastic physics in improving flow-dependent reliability

EnsVar is highly sensitive to representation of model uncertainty. Turning it off deteriorates (improves) Residual in convective (clear-sky) regions

European Centre for Medium-Range Weather Forecasts

Mark J Rodwell
Top 50 Warm Conveyor Belt inflow events in box indicated from Nov 15 – Oct 16

From Heini Werni. Based on trajectories ascending by more than 600 hPa in 2d
EDA variance assessment with ASCAT surface v wind: Non-WCB composite

87 cases

Bias² and Residual are not significant in absence of WCBs.

ObsUnc² a large component
EDA variance assessment with ASCAT surface v wind: WCB (inflow) composite

50 cases

Larger EnsVar and Depar\(^2\) so a more uncertainty situation
Increased Bias\(^2\) (first moment error)
Positive Residual but not significant: possibly insufficient spread associated with cyclogenesis, or underestimation of observation uncertainty within the WCB region?
EDA variance assessment with MHS “all sky” mid-tropospheric humidity: Non-WCB

87 cases

Bias and residual are not significant in absence of WCBs

Microwave channel 5
EDA variance assessment with MHS “all sky” mid-tropospheric humidity: WCB events

50 cases

Increased $\text{Depar}^2$ and EnsVar for inflow composite
Difficult to say if EnsVar is reasonable since negative residual is dominated by the overestimation of $\text{ObsUnc}^2$ in WCB region (it is larger than the departures!)

The aim of the diagnostic is to indicate deficiencies and chart our progress towards the ultimate target, but it is not practical to get there immediately

Microwave channel 5
Summary, implications and future directions

- Forecast system improvement ≡ Minimisation of a proper score
- Proper Score = Reliability – Resolution + Uncertainty
- Red curve (constant forecast) = Black curve (climatological distribution)
  Perfectly reliable (Reliability = 0), but no resolution (Resolution = 0)
- Green and blue forecasts based on (local) partition of phase-space
  If we work to make these reliable (Residual = 0 in EDA budget) then:
    Model uncertainty is likely to be well-represented and …
    Shaded regions ⇒ Resolution > 0 ⇒ More skill (into medium-range?)

‡ Practical but not trivial. Need to improve model and representation of model and observation uncertainty
† Better (and more numerous) observations will help increase sharpness; Can view as a separate important task

The instantaneous uncertainty growth-rate can be written as:
\[
\frac{1}{\sigma_x} \frac{\partial \sigma_x}{\partial t} = \rho_{XX_t} \frac{\sigma_{X_t}}{\sigma_x}
\]

My animation showed the \(\sigma_{X_t}/\sigma_x\) part. Do the random numbers used in some model uncertainty formulations deteriorate the correlation aspect?
Thank you
Reliability in ensemble forecasting

\( \text{Error}^2 = \text{EnsVar} + \text{Residual} \)

Error \( \approx \) Spread

(when averaged over enough start dates - a consequence of statistical “reliability”)

Progress: While maintaining this relationship, we wish to reduce both error and spread (i.e. “shrink diagram in x,y”)

Proper scores (Brier, CRPS, Ignorance etc.) help ensure that development decisions lead us in this direction

(Cross-terms on squaring have zero expectation. EnsVar is scaled variance to account for finite ensemble-size)

Adapted from Rodwell et al. (2015) QJRMS
Reliability in ensemble data assimilation

Chaos makes it difficult to identify problems in the medium-range using the spread-error relationship. Go to much shorter lead-times – within ensemble data assimilation process. Need to take account for observation error. Obtain diagnostic equation to evaluate “instantaneous” growth of uncertainty.

\[ \text{Depar}^2 = \text{Bias}^2 + \text{EnsVar} + \text{ObsUnc}^2 + \text{Residual} \]

(Cross-terms on squaring have zero expectation. EnsVar is scaled variance to account for finite ensemble-size)
Summary
Graphs of overall forecast skill, reliability and sharpness are useful for monitoring performance.

Short-range, flow-dependent diagnostics offer a more direct method of identifying/improving underlying deficiencies in our representation of model uncertainty and observation uncertainty. For example, increasing model uncertainty in convective situations and decreasing it in clear-sky situations could improve reliability. Sharpness can be improved with better (use of) observational information.

CRPSS, extratropical precipitation against observations

Progress: 2.8 day lead-time gain in the last decade
You will know better than me what this means to forecast users