Numerical weather prediction is fundamentally a probabilistic task owing to the growth of unavoidable uncertainties in the forecast’s initial conditions and in the forecast model itself (Sutton 1954; Lorenz 1963). A key question for the user is how certain they can be that a “10% probability of precipitation” really means that they will be unlucky to get wet. How would they assess the reliability of such a prediction? One approach would be for them to keep a record of the days when the forecast indicated a 10% probability of precipitation and to see if rainfall actually occurred on 10% of those days. If, in reality, it occurred on 20% of the days, this would indicate that these probabilistic forecasts were unreliable. On the other hand if, for a large set of occasions when the forecast gave a 40% chance of a hurricane making landfall in a particular region, a hurricane did make landfall 40% of the time, this would represent a “reliable” (Sanders 1958) set of forecasts. If the decision about whether to defend a vulnerable piece of infrastructure could be based purely on the forecast, then, since the forecast is reliable, a simple approach would be to defend the infrastructure if the cost to do so was less than 40% of the loss that would otherwise occur (Richardson 2000). In reality other factors will influence such a decision, but this example does highlight the importance of reliability. We may ask how we could do even better. Suppose we could partition the forecasts into two categories (based on the initial flow conditions) where the probabilities of landfall were, say, 60% and 20% (i.e., one flow situation is more likely to lead to landfall than the other), and suppose that the forecasts in both categories were reliable. It is straightforward to show (see “The benefits of flow-dependent reliability” sidebar) that, under the assumptions of this simple decision model, the expected expense in defending the infrastructure is always reduced (or matched). This result, which does not depend on the choice of numbers, emphasizes the potential utility of improving flow-dependent reliability. The key question to address here is, How do we assess and improve the flow-dependent reliability of our forecasts? Before attempting to address this question, it
is useful to discuss developments that have already improved forecast reliability and then to motivate use of the approach taken in this study.

**PREVIOUS PROGRESS IN ENSEMBLE FORECAST RELIABILITY.** Probabilistic weather forecasts are generally made through use of an ensemble of individual forecasts (Lewis 2005), each starting from a slightly different initial state and including a different realization of model uncertainty. Computing resources put a limit on the number of ensemble members that can be made at a chosen resolution, and thus on the fidelity with which the underlying forecast distribution can be estimated. This and other limitations make it impractical to assess forecast reliability for all possible events and probabilities—particularly for rare and/or extreme events—and so simpler approaches to diagnosing deficiencies in reliability are useful. One consequence of reliability in an ensemble forecast system is the so-called spread–error relationship (Leutbecher and Palmer 2008), whereby the average difference (or distance in state space) between an ensemble member and the ensemble mean (e.g., the ensemble standard deviation) should match the average difference between the eventual outcome and the ensemble mean [e.g., the root-mean-square error (rmse) of the ensemble mean] when averaged over a large-enough set of start dates and using a simple adjustment to account for finite ensemble size. As will be seen below, this relationship provides a practical first-order method for assessing ensemble reliability, although it is not so useful for the diagnosis of deficiencies in reliability.

Here, we focus mainly on the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble (ENS) forecast system, which became operational in May 1994 (Palmer et al. 1992; Molteni et al. 1996). The diagnostic approach discussed should, however, be more widely applicable. Figure 1a shows annual means of ENS spread (solid) and error (dashed) as a function of forecast lead time for Northern Hemisphere 500-hPa geopotential height (Z500) for the years 1996, 2005, and 2014. System developments have led to a reduction in errors and an improvement in the annual-mean agreement with the spread to the point that the spread and error curves are virtually indistinguishable by 2014. These changes represent substantial improvements in the reliability and sharpness of predictive distributions.

In 1996 (Fig. 1a, blue curves), the rapid rise in spread to day 2 reflected the inclusion of fast-growing “singular vector” (SV) perturbations to the initial conditions (Molteni and Palmer 1993), which helped the spread keep pace with error growth over the first 2 days, but it can be seen that this was not sustained beyond 2 days. Since 1996, four-dimensional variational data assimilation (Rabier et al. 2000) was developed and subsequently incorporated into an ensemble of data assimilations (EDA; Isaksen et al. 2010) to produce a set of equally likely initial conditions that take into account observation and model uncertainty and the growth of uncertainty from the previous set of initial conditions. These developments, together with many other incremental improvements (e.g.,

---

**THE BENEFITS OF FLOW-DEPENDENT RELIABILITY**

In the landfalling hurricane example, suppose that the overall forecast probability for landfall is \( p \) and that this probability is reliable (so that landfall does occur in a fraction \( p \) of these cases). For the simple decision model discussed, let \( C \) be the fixed cost of taking action and let \( L \) be the fixed loss that would be sustained if action was not taken and the event occurred. Writing \( \alpha = C/L \), the expected expense of defending or rebuilding the infrastructure would then be \( \min(p, \alpha) \) (in units of \( L \)). Suppose now that it is possible to partition the forecasts into two categories based on their initial flow conditions, with the fraction of forecasts in each category being \( w_1 \) and \( w_2 \) (where \( w_1 + w_2 = 1 \)) and the reliable probabilities of landfall being \( p_1 \) and \( p_2 \), respectively. Then \( w_1 p_1 + w_2 p_2 = p \). Without loss of generality, we can assume \( p < p_1 < p_2 \). The overall expected expense following the partition will be \( w_1 \min(p_1, \alpha) + w_2 \min(p_2, \alpha) \). There are four possible rankings of \( \alpha, p_1, p_2 \), and \( p \), and corresponding expected expenses for the partitioned and unpartitioned forecasts (Table SB1). In all cases, the expected expense is reduced or matched when a partitioning (with flow-dependent reliability) is possible.

<table>
<thead>
<tr>
<th>Table SB1. Rankings and comparison of expected expenses for partitioned and unpartitioned forecasts.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ranking</strong></td>
</tr>
<tr>
<td>( \alpha &lt; p_1 &lt; p &lt; p_2 )</td>
</tr>
<tr>
<td>( p_1 &lt; \alpha &lt; p &lt; p_2 )</td>
</tr>
<tr>
<td>( p_1 &lt; p &lt; p_1 &lt; \alpha )</td>
</tr>
<tr>
<td>( p_1 &lt; p &lt; p_2 \leq \alpha )</td>
</tr>
</tbody>
</table>
et al. 2000; Dee and Uppala 2009; Jung et al. 2010), have enabled a reduction in the magnitude of SV perturbations (Leutbecher and Lang 2014) along with the reduced errors and improved annual-mean agreement with the spread. Notice also the more realistic “exponential” shape (Molteni and Palmer 1993; Harlim et al. 2005) of the spread and error curves for 2014 over the first 5 or 6 days, which are much flatter at short ranges.

Obtaining annual-mean agreement between spread and error is an important first step in the development of a reliable forecast system. The primary reason for making ensemble forecasts, however, is to represent the day-to-day variations in forecast uncertainty. Figure 1b shows time series of spread and error for forecasts of European Z500 at a lead time of 6 days from several of the world’s operational forecasting centers. The concerted rise and fall in spread for all centers demonstrates that predictability varies with the prevailing flow situation. The partial agreement between the spread and error curves (it can never be perfect on a day-to-day basis; Whitaker and Loughe 1998) demonstrates that the forecasts do possess some degree of flow-dependent reliability, including for flow situations where intrinsic predictability is low (i.e., where uncertainty is growing rapidly). A key question is whether we can do any better.

**WHY FOCUS ON INITIAL UNCERTAINTY GROWTH RATES?** This section motivates the approach considered here, where flow-dependent reliability (out to ~10 days) is assessed and improved through the diagnosis of short-range (~12 h) ensemble forecasts.

The day-6 spread curves displayed in Fig. 1b represent the 6-day integral of the models’ instantaneous (and flow dependent) uncertainty growth rates (Smith et al. 1999), and it would be preferable if we could diagnose deficiencies in these instantaneous growth rates. Here we go some way to achieving this by diagnosing deficiencies at synoptic spatiotemporal scales. To illustrate this synoptic-scale growth rate, and with cases of large forecast uncertainty in mind, we focus on the jet stream level where uncertainties are likely to propagate most rapidly and be magnified by subsequent downstream cyclogenesis (Hoskins et al. 1985). A useful quantity for this purpose is potential vorticity (PV) on the \( \theta = 315\) K isentropic surface. We will estimate the growth rate by calculating tendencies in the ensemble standard deviation \( \hat{\sigma}_{PV} \) in the 12-h background (first guess) forecasts of the EDA (25 members). Because of the strong advection by the jet stream, flow features associated with strong amplification of uncertainty are better identified by calculating the growth rate following the horizontal ensemble-mean flow \( \bar{\mathbf{u}} \) along the isentropic surface:

\[
\frac{1}{\hat{\sigma}_{PV}} \left\{ \frac{\partial \hat{\sigma}_{PV}}{\partial t} + \bar{\mathbf{u}} \cdot \nabla \hat{\sigma}_{PV} \right\}.
\]
The time-derivative term in Eq. (1), which represents the local growth rate, is estimated using a 1-h time step. A spatiotemporal filter is then used to highlight uncertainty growth that projects onto the synoptic scales. The shading in Fig. 2 shows an example of the resulting field (see figure caption for more details), which is representative of the situations investigated later in a more systematic way. The red region over North America in Fig. 2 indicates a strong source of uncertainty. The 850-hPa winds in this region (arrows) indicate relatively warm moist southerly flow, which is associated with high values of convective available potential energy (CAPE; not shown) and an increased likelihood for intense convection and mesoscale convective systems (MCSs); the black dots show ensemble-mean precipitation—and thus the general location where precipitation is likely. The large uncertainty growth rates thus suggest that the ensemble places considerable uncertainty in the location(s) and magnitude(s) of this convection, possibly reflecting inherent predictability limits or the model’s limited ability to represent such convection with ~18-km grid resolution. The red line indicates the tropopause position [2 PV units (PVU; 1 PVU = 10^{-6} \text{K kg}^{-1} \text{m}^{2} \text{s}^{-1}) on the 315-K surface], and this highlights a trough feature over the Rocky Mountains. The jet stream (generally just to the south of the red line) will be affected by the uncertainty generated by this baroclinic and convective situation, and it will propagate this uncertainty (as well as the “signal”; Grams et al. 2011) downstream. The flow over North America in Fig. 2 is reminiscent of the “trough/CAPE” synoptic flow situation discussed by Rodwell et al. (2013), which often precedes deterministic forecast busts experienced over Europe 6 days later—the approximate time scale for errors to propagate across the Atlantic. However, using the spread–error relationship, Rodwell (2016) was not able to say conclusively that the ensemble was unreliable in these cases (the spread increased less than the error, but the difference was not statistically significant). This was thought to be partly due to an unavoidable limit on the number of cases available (when evaluating a given version of the ECMWF forecast system) and the complicating effects of interactions with other features of the flow over the North Atlantic by day 6. For example, Fig. 2 also indicates enhanced uncertainty growth rates over the North Atlantic, which appear to be associated with a “warm conveyor belt” (WCB; Browning 1990; Madonna et al. 2014). WCBs are associated (in the ECMWF model) with more slantwise ascent than the MCSs discussed above, although they also include embedded convection. As the flow evolves over the next few days (not shown), the uncertainty associated with the MCS activity is advected northward into a developing strong ridge, and the uncertainty associated with the WCB is colocated with a developing strong trough. The subsequent interaction of the ridge and trough, and the “pooling” of two major sources of uncertainty, seems likely to be a major factor in the very large uncertainties seen at day 6 over Europe for forecasts starting between 5 and 7 March 2017 in Fig. 1b. [See Lillo and Parsons (2017) for further discussion of the causes of deterministic forecast busts.] See "A case for international cooperation" sidebar for more discussion of this particular case.

The collocation of strong uncertainty growth rates with moist processes (in both the MCS and WCB regions) is interesting. Does this result represent, for example, an improvement over the baroclinic singular vectors highlighted by Molteni and Palmer (1993)

![Fig. 2. The synoptic-scale uncertainty growth rate (shaded) for PV on the 315-K isentropic surface following the horizontal flow as defined in Eq. (1) and derived from the background forecasts of the ECMWF EDA. Also shown, for the background forecast of the unperturbed “control” member of the EDA are PV = 2 potential vorticity units (PVU) on the 315-K surface (red contour; indicating where this surface intersects the tropopause) and horizontal winds on the 850-hPa surface (vectors) and for the ensemble mean the 24-h accumulated precipitation (dots; radius is linear up to 48 mm and then constant). The synoptic spatial filter applied to the growth rate is also applied to the PV = 2-PVU contour and involves a \nabla^{-2} smoothing on spectral harmonics with a total wavenumber greater than 21. All fields have a 1-day mean applied, which is centered at 0600 UTC 7 Mar 2017.](http://journals.ametsoc.org/bams/article-pdf/99/5/1015/3749385/bams-d-17-0027_1.pdf)
when nonlinear effects and moisture are included, or does it reflect deficiencies in the representation of model uncertainty? Note that model uncertainty is represented in the EDA through “stochastic perturbations to physical tendencies” (SPPT; Buizza et al. 1999), but there are no SV initial perturbations to the background forecasts. In short, how well does the forecast system represent intrinsic uncertainty growth rates and maintain reliability in these (and other) synoptic flow situations? This work thus complements the more idealized studies on predictability by Durran and Gingrich (2014) and Sun and Zhang (2016), and results such as those presented in Fig. 2 may provide useful information for such studies on the sources and scales of uncertainty that dominate present-day forecast initialization.

The trough/CAPE initial flow situation represents here a useful example with which to demonstrate the more general utility of evaluating flow-dependent reliability at short (~12 h) forecast ranges. If we can show that short-range forecasts from an initial flow situation, such as the trough/CAPE synoptic pattern, are unreliable owing to deficiencies in uncertainty growth rates (too much or too little), then this will imply that ensemble forecasts that predict a high likelihood of a trough/CAPE situation at any lead time will become unreliable after this lead time (in an area expanding from the likely trough/CAPE event). Such arguments would apply to other synoptic flow types, for which initial uncertainty growth rates might be found to be deficient. Putting the argument the other way around, modeling developments that improve initial uncertainty growth rates for a range of flow types are a necessary requirement for more reliable (and indeed more skillful) ensemble forecasts. Here we discuss a diagnostic strategy that should help achieve this goal.

**EVALUATION OF INITIAL UNCERTAINTY GROWTH.** The standard spread–error relationship is not suitable for the evaluation of flow-dependent uncertainty growth in the 12-h EDA background forecasts because, at these short lead times, uncertainty in our knowledge of the truth cannot be neglected. Instead, we use the observation–space variance budget developed in Rodwell et al. (2016). This budget is simply an extension of the spread–error relationship to account for bias and observation uncertainty, and is analogous to the equations of data assimilation. Rather than quantifying ensemble-mean forecast “errors” from the truth, the budget calculates the ensemble-mean background forecast “departures” from the uncertain observations. When averaged over a set of EDA cycles, the mean-squared departures, Depar², from a given set of observations can be decomposed as

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

where EnsVar is the mean ensemble variance [scaled by \((m + 1)/(m - 1)\), i.e., the squared spread]. For a perfect ensemble forecast system, with no observation error, the other terms on the right-hand side of Eq. (2) would be zero, and the budget would revert to the standard spread–error relationship. If there is observation error, then this should be accounted for by ObsUnc², the estimated observation error variance, as modeled by the data assimilation system (here it is based on the observation perturbations actually applied within the EDA). The Bias² term is calculated as the squared mean difference between the observations and the ensemble mean. A nonzero Bias² indicates a bias in the background forecasts and/or the assimilated observations. The Residual term is calculated as the residual in the budget. A nonzero Residual term indicates a deficiency in the ensemble variance and/or a deficiency in the modeling of observation error variance. Both the Bias² and Residual terms are associated with a lack of reliability, and, because of this, Eq. (2) is called here the EDA reliability budget. Importantly from a diagnostic point of view, Rodwell et al. (2016) found that the Residual term was able to highlight mean changes in reliability associated with the inclusion of model uncertainty (SPPT) and with improvements to the modeling of observation error variance. In the present study, the aim is to determine whether the Residual term is able to highlight statistically significant deficiencies in flow-dependent reliability.
COMPOSITE STUDY: THE TROUGH/CAPE FLOW TYPE. As an example of a flow-dependent application of Eq. (2), we consider “jet stream winds” (zonal winds at 200 ± 15 hPa—as observed by aircraft, together with collocated model background values) for the trough/CAPE synoptic situation discussed above. A total of 54 cases were found during the period 19 November 2013–12 May 2015 (when version 40r1 of ECMWF’s Integrated Forecast System was operational) for which the initial conditions of the background forecast of the EDA unperturbed control closely matched the trough/CAPE pattern, using the same method as in Rodwell et al. (2013). Figure 3f shows the aircraft observation density for these cases. Aircraft observations are numerous over central North America at this cruising altitude, and, indeed, they are particularly influential in the data assimilation system.

Before calculating the terms in the EDA reliability budget in Eq. (2), all data are first aggregated onto an approximately equal-area grid (~125 km²). Figures 3a–e show the resulting terms for the trough/CAPE composite. As discussed above, the budget decomposes the squared departure term (Depar²; Fig. 3a) into contributions from the bias (Bias²; Fig. 3b), ensemble variance (EnsVar; Fig. 3c), observation uncertainty (ObsUnc²; Fig. 3d), and the Residual term (Fig. 3e).

The spatial structure of the squared departure term (Fig. 3a) indicates increased ensemble-mean forecast departures from the observations in the Great Lakes–Mississippi River region of North America in the trough/CAPE composite. This is to be expected because of the strong (and less predictable) MCS activity liable to be taking place in this region. The ensemble variance (Fig. 3c) does indicate more uncertainty (relative to surrounding regions), but notice that this does not fully account for the increased departures. The observation uncertainty term (Fig. 3d) scales roughly as the reciprocal of the number of observations aggregated into a grid cell (Fig. 3f) and is thus relatively small over North America. The residual (Fig. 3e; note the different shading interval) is particularly large in the region associated with MCS activity; it has roughly twice the magnitude of the ensemble variance and is statistically significant at the 5% significance level (as indicated by the more saturated colors).

A key question here is whether the residual (Fig. 3e) in the Great Lakes–Mississippi River region is due to an underestimation of observation errors (or observation error correlations) or a lack of modeled ensemble variance (or model representativity of point observations). At the ~125-km² aggregation scale used here, it is unlikely that deficiencies in observation error modeling or representativity are the most important issue since similar (high) observation densities are seen over western North America (Fig. 3f) where the residual (Fig. 3e) is much smaller. Hence it is more likely that the main deficiency is an underestimation of ensemble variance in the jet-stream winds during trough/CAPE situations. From a diagnostic point of view, the key result here is that the EDA reliability budget is able to identify statistically significant flow-dependent deficiencies in reliability. Note that the EDA uses the variances of the background forecasts directly in its background error covariance matrix. This is likely to make the EDA responsive to the flow of the day, and this could be important for the success of this targeting of a given flow type.

It is possible that an enhanced representation of model uncertainty in this convective region could improve the reliability budget—consistent with the results of Rodwell et al. (2016), who showed that turning off SPPT reduced mean EnsVar in convective regions by ~60%.

There could, however, also be a role for flow-dependent systematic model error. For the trough/CAPE composite, Fig. 4c shows the convective heating at 300 hPa within the EDA control (unperturbed) background forecast. It highlights the increased likelihood for MCS activity over the Great Lakes–Mississippi River region. This convective heating is largely balanced by dynamical cooling (Fig. 4a). The “hole” in radiative cooling in this region is perhaps indicative of higher cloud tops, and, indeed, the cloud term (Fig. 4d) does also indicate active grid-scale microphysics. While there is some cancellation between all these terms, there is a positive mean “analysis increment” (Fig. 4e; note the reduced shading interval). The new observational data being assimilated evidently suggest that the background forecast at 300 hPa is systematically too cold. This could indicate that the modeled convection does not extend high enough, with the consequence that the interaction between the mesoscale convection and the jet stream is underestimated. This could be another reason for the apparent lack of ensemble variance in zonal winds at 200 hPa (Fig. 3). Investigation will continue to better identify the reasons for the strong positive residual in Fig. 3e.

The sum of all the tendency terms and the increment is the (analyzed composite) “Evolution” of the flow in the background forecast (Fig. 4f):

\[
\text{Evolution} = \text{Dynamics} + \text{Vertical Diffusion} + \text{Radiation} + \text{Convection} + \text{Cloud} + \text{Increment.}
\]
The downstream evolution seen over the North Atlantic may be associated with large-scale Rossby waves (which would include the trough over the Rockies) and their amplification by the composite convection. The vertical diffusion term is not shown in Fig. 4 since it is negligible in the temperature budget at 300 hPa. The “almost equal to” symbol in Eq. (3) reflects other generally negligible terms in the model’s tendency budget (when averaged over sufficient data assimilation cycles) such as horizontal diffusion and semi-implicit adjustment. By highlighting mean deficiencies, and possibly pointing to their causes, Eq. (3) is also an important budget to consider when diagnosing flow-dependent reliability. Further discussion of Eq. (3), which is a development of the ideas of Klinker and Sardeshmukh (1992), can be found in Rodwell and Palmer (2007), Rodwell and Jung (2008), and Klocke and Rodwell (2014).

**MODEL OR OBSERVATION UNCERTAINTY?**

The residual in Fig. 3e is likely to be associated with deficiencies in the forecast model (or the representation

---

![Fig. 3](http://journals.ametsoc.org/bams/article-pdf/99/5/1015/3749385/bams-d-17-0027_1.pdf)

**Fig. 3.** Short-range variance budget based on upper-tropospheric winds in trough/CAPE situations. (a)–(e) The terms in the EDA reliability budget in Eq. (2) applied to zonal winds at 200 ± 15 hPa [as observed by aircraft, together with collocated model background values aggregated to an octahedral (O80) grid resolution of 125 km²] for the trough/CAPE composite. (f) The density of the aircraft observations that are assimilated within the EDA control. To reduce noise, an average of at least one observation per grid cell per 12-h analysis cycle is required for the budget to be plotted. The color bar uses a dual-color palette; statistically significant values are determined using a pointwise Student’s t-test at the 5% level and indicated with the more saturated colors. Autocorrelation is not accounted for in the t-test since the composite dates are not necessarily consecutive. The significance of Bias² is based on that of its unsquared value.
of model uncertainty). However, the EDA reliability budget is also sensitive in general to the modeling of observation uncertainty; this is useful because a good modeling of observation uncertainty is also important for the reliable initialization of ensemble forecasts.

There can be situations where the residual term is most clearly associated with the modeling of observation uncertainty. For example, composites have been produced based on the existence of WCBs (objectively identified through trajectory calculations; Wernli and Davies 1997; Madonna et al. 2014). The EDA reliability budget for this WCB composite, evaluated for satellite Microwave Humidity Sounder channels sensitive to midtropospheric humidity (not shown here), has an ObsUnc term that alone is around twice the magnitude of the Depar term. The budget thus suggests that the magnitudes of modeled observation uncertainty could be reduced. This would have the effect of drawing the EDA more strongly toward the observations, producing a sharper initial distribution.

It is possible that other factors will need to be improved at the same time, including cloud detection and the forecast model’s representation of sharp inversions—particularly when trying to assimilate observations with deep weighting functions. Note that with new observation types and developments to a data assimilation system (such as the all-sky developments active in this WCB example; Geer and Bauer 2011), it is better from a reliability point of view to first err on the side of overestimating observation uncertainty.

Fig. 4. Short-range mean tendency budget for upper-tropospheric temperatures in trough/CAPE situations. (a)–(d) Mean tendencies in temperature at 300 hPa for the trough/CAPE composite based on output from the model’s dynamics and physical parameterizations in the 12-h background forecast of the EDA control. (e) Analysis increment in response to the assimilation of new observations. (f) Flow evolution as in Eq. (3). Values that are significantly different from zero at the 5% level are indicated with more saturated colors; see Fig. 3 caption for more details.
In other situations, there may be more ambiguity as to whether it is the representation of observation or model uncertainty that is the main problem. Deficiencies in the model’s representativity of observations are a good example—which may be explored by calculating Eq. (2) without aggregating data beforehand. Additional information on observation errors (e.g., “Desroziers statistics”; Desroziers et al. 2005) might also help resolve some ambiguities.

**DISCUSSION: A DIAGNOSTIC DEVELOPMENT FRAMEWORK.** This study has been motivated by the idea that focusing modeling efforts on short-range flow-dependent reliability offers a practical framework for improving forecast skill (see "A framework for forecast system development" sidebar for a discussion based on the improvement of “proper” scores). To investigate the feasibility of this approach, the “reliability budget” in Eq. (2) of Rodwell et al. (2016), which is essentially an extension of the spread–error relationship to include observation uncertainty, has been applied here to the ECMWF EDA for a composite of cases where the trough/CAPE flow over North America was present in the initial conditions. This initial flow type is known to lead to increased forecast uncertainty for Europe at a lead time of 6 days (Rodwell et al. 2013). Results here indicate that uncertainty growth rates in the vicinity of the jet stream over North America are too weak, and this is likely to be associated with insufficient forcing by mesoscale convective systems MCSs, which, themselves, have low predictability. The immediate implication of this result is that, while forecast uncertainty is large for Europe in these situations, it may not be large enough. Partly this may be a consequence of systematic errors in the height that this convection attains (and thus how strongly it interacts with the jet stream), but deficiencies in the representation of model uncertainty in such convective situations are also likely to play a role. There is scope for providing advice to users based on such knowledge. For example, if a strong trough/CAPE situation exists over North America in the initial conditions, or is likely tomorrow, then advice to users stating that the current large forecast uncertainty for Europe next week is probably an underestimate and that delaying decisions (until details of the imminent MCS activity are known) might be a sensible course of action. It is also possible that calibration of forecast output (Hagedorn et al. 2008, 2012; Hamill et al. 2008; Hemri et al. 2014) could benefit in a limited way from such knowledge.

There remain slight differences in the way the ECMWF ENS is initialized and run compared to the background forecasts of the EDA (Lang et al. 2015), but as a more seamless EDA–ENS is developed, these

---

**A FRAMEWORK FOR FORECAST SYSTEM DEVELOPMENT**

The aim of forecast system development can be summarized as the improvement of proper scores (Gneiting and Raftery 2007). For example, the Brier score (BS), can be written and decomposed as follows:

\[
BS = \frac{1}{N} \sum_{t=1}^{N} \left( p(t) - o(t) \right)^2
\]

\[
= \sum_{k=1}^{K} w_k (p_k - o_k)^2 + \sum_{j=1}^{K} w_o (1 - o_k),
\]

\[ \text{(SB1)} \]

where, on the first line, \( t = 1, \ldots, N \) is an index of the forecast start time, \( p(t) \) is the forecast probability that some given event will occur, and \( o(t) = 1 \) if the event occurs and 0 if not. The score is proper (and difficult to improve by hedging) because its expected value is optimized (minimized) when the \( p(t) \) are equal to the expected values of \( o(t) \)

The second line in Eq. (SB1) shows the “reliability–refinement” decomposition of the Brier score (DeGroot and Fienberg 1983). Instead of the usual approach of binning directly on forecast probabilities, it is argued here, as a thought experiment, that this decomposition can be considered as a sum over a partition of \( K \) initial synoptic-scale flow types. The probabilities that arise from any forecast initialized from a given flow type should be similar enough to be represented by a single probability bin (Stephenson et al. 2008) if the flow types are defined tightly enough, if the events are local to the flow type, and if short-enough lead times are considered. Here \( p_k \) represents such forecast probabilities from each flow type, \( o_k \) represents the corresponding outcome frequencies, and \( w_k \) represents the fraction of forecasts within each initial flow-type bin. Modeling developments that improve short-range flow-dependent reliability will bring \( p_k \) closer to \( o_k \) and thus reduce the overall reliability term, but they should have less impact on the refinement term, which is only directly dependent on the initial flow types and the verifying observations. Hence, modeling efforts focused on improving short-range flow-dependent reliability (even if this involves increasing uncertainty growth rates) should lead to improvements in the Brier score. As an aside, the use of more observations, or the extraction of more information from the observations, is likely to be important for the improvement of the refinement term. Similar arguments follow for other proper scores (see, e.g., Bentzien and Friederichs 2014), which can measure forecast performance over a wider range of event definitions.
short-range diagnostics should become ever more relevant to the ENS. Forecast model developments (including developments to the representation of model uncertainty; Plant and Craig 2008; Berner et al. 2009; Christensen et al. 2017; Ollinaho et al. 2017), which improve short-range reliability for this and a variety of other flow types, should enable the ensemble to better maintain reliability out to ~10 days, as the phase-space trajectories “pass through” these different flow types. (At longer lead times, slower processes that are not assessable by the EDA reliability budget will start to become important.) Further work is required before the anticipated beneficial impacts on forecast skill can be verified for the trough/CAPE flow type. However, previous application of the EDA reliability budget (without flow dependence; Rodwell et al. 2016) indicated that there may be too much uncertainty growth in regions of subtropical anticyclones. The hypothesis was that the ECMWF representation of model uncertainty (Buizza et al. 1999) might be too active in clear-sky conditions. Motivated by this diagnostic result, the model uncertainty scheme at ECMWF has been adapted to reduce its impact in clear-sky conditions, and this development is about to be implemented in the operational forecast system. Further discussion of the impact of this change will be published shortly (S.-J. Lock 2017, personal communication). The demonstration that such diagnostics can have a real beneficial impact on the development of an operational forecast system is encouraging. More reliable initialization of ensemble forecasts can also come from improvements in the flow-specific modeling of observation error characteristics and “observation operators,” which map forecast model fields to observed quantities (Geer and Bauer 2011). Again, such modeling developments should lead to improved forecast skill.

Of course, such modeling work already takes place (at ECMWF and in other forecasting centers around the world), but the thought is that this diagnostic framework, based on short-range flow-dependent reliability, might help to organize and prioritize efforts. For example, initial flow types could be identified through localized cluster analysis and prioritized for attention based on their contribution to the overall reliability term of a proper score (see “A framework for forecast system development” sidebar) or, analogously, on their frequency-weighted residual in the EDA reliability budget in Eq. (2). This budget should also be useful in monitoring the overall development process (including the assimilation of more observational information as well as improvements to modeling), the aim being to reduce EnsVar while maintaining small Bias2 and Residual terms for a range of initial flow types.

ACKNOWLEDGMENTS. The authors thank Elias Hólm, Heather Lawrence, Sam Lillo, and Ashton Robinson for the helpful discussions, and Alan Thorpe for initiating links between ECMWF and the University of Oklahoma. They would also like to thank Brian Etherton, Tom Hamill, Tim Palmer, and an anonymous reviewer for their insightful comments. One author, D.B.P., would especially like to thank Steven Cavallo and members of his research group.

REFERENCES


Durran, D. R., and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not of practical


Rodwell, M. J., 2016: Using ensemble data assimilation to diagnose flow-dependent forecast reliability. ECMWF Newsletter, No. 146, ECMWF, Shinfield Park, Reading, United Kingdom, 29–34.


