Feature-Based Diagnostic Evaluation of Global NWP Forecasts

MARION MITTERMAIER, RACHEL NORTH, AND ADRIAN SEMPLE

Weather Science, Met Office, Exeter, United Kingdom

RANDY BULLOCK

Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 15 April 2015, in final form 3 September 2015)

ABSTRACT

With the resolution of global numerical weather prediction (NWP) models now typically between 10 and 20 km, forecasts are able to capture the evolution of synoptic features that are important drivers for significant surface weather. The position, timing, and intensity of jet cores, surface highs and lows, and changes in the behavior of these forecast features is explored using the Method for Object-based Diagnostic Evaluation (MODE) at the global scale. Previously this was only possible with a more subjective approach. The spatial aspects of the forecast features (objects) and their intensity can be assessed separately. The evolution of paired forecast–analysis object attributes such as location and orientation differences, as well as area ratios, can be considered. The differences in the paired object attribute distributions from various model configurations were evaluated using the $k$-sample Anderson–Darling (AD) test. Increases or decreases in hits, false alarms (forecast-not-observed), and misses (observed-not-forecast) features were also assessed. It was found that when focusing purely on the forecast features of interest, differences in seasonal spatial extent biases emerged, intensity biases varied as a function of analysis time, and changes in the attribute distributions could be detected but were largely insignificant, primarily due to sample size. As has been shown for kilometer-scale NWP, results from spatial verification methods are more in line with subjective assessment. This type of objective assessment provides a new dimension to the traditional assessment of global NWP, and provides output that is closer to the way in which forecasts are used.

1. Introduction

Operational numerical weather prediction (NWP) centers typically update model configurations on a regular basis. Typical changes include physics improvements, data assimilation changes such as the incorporation of new satellite data, or increases in horizontal or vertical resolution. These changes are preceded by periods of intense testing and evaluation, where the stability, robustness, and overall impact of the proposed changes is assessed. The metrics used range from the traditional (routine) point- or grid-based verification scores (e.g., root-mean-square error skill score) to the newer spatial-based methods [e.g., fractions skill score (FSS); Mittermaier et al. (2013)] to name but one that is used extensively at the Met Office. Certainly for kilometer-scale NWP these newer methods are becoming as routine as the traditional metrics (Ebert et al. 2013).

Many verification metrics such as skill scores used traditionally in verification of NWP forecasts have the characteristic that they are by necessity some form of average statistical measure over some predefined area. As an example consider the World Meteorological Organization (WMO) Commission for Basic Systems (CBS) exchange of scores between global NWP modeling centers (http://apps.ecmwf.int/wmolcdnv/) for different regions: Northern Hemisphere ($90^\circ$–$20^\circ$N), Southern Hemisphere ($20^\circ$–$90^\circ$S), and tropics ($20^\circ$N–$20^\circ$S). Although incredibly valuable as performance measures, such scores sometimes have the disadvantage that they can provide an insufficient, and possibly misleading representation of model performance as many significant characteristics of the model may be poorly represented in any particular verification score.

With many global NWP models now running with grid spacings below 20 km, it has also been the case that higher-resolution model versions often have larger rmse...
(and smaller skill scores) than their coarser-resolution counterparts, even for relatively smooth fields such as 250-hPa winds or temperature. This is because by definition rmse favors a smoother field, and higher horizontal resolution invariably implies more detail. This detail may not be correct, in which case it acts like noise (Mass et al. 2002). Higher horizontal resolution will also pick up smaller features, previously subgrid scale for global NWP, which can add to the noise via small errors in displacement for example. Hence, the use of traditional metrics may not accurately reflect the apparent subjective improvement in the forecast gained from increased horizontal resolution. This phenomenon was first documented in conjunction with kilometer-scale NWP precipitation forecasts, and that sparked the development of spatial verification methods and an intercomparison of such methods (e.g., Ebert 2008; Ahijevych et al. 2009; Gilleland et al. 2009, 2010).

For this reason, spatial methods, and specifically the object-based subclass of spatial methods, can provide a means of objective verification focused on synoptic features of interest. The Method for Object-Based Diagnostic Evaluation (MODE) (Davis et al. 2006a,b) is one such method that enables one to extract features of interest from a forecast (and analysis) field and compare only these features based on a number of attributes such as the location and orientation error, and area ratio. Like the other spatial methods MODE was originally developed to consider precipitation forecast performance in convective-scale NWP models, and was one of the methods participating in the spatial methods intercomparison project mentioned earlier (Davis et al. 2009). Other examples of object-based methods include the contiguous rain area (CRA) method (Ebert and McBride 2000), structure–amplitude–location (SAL) method (Wernli et al. 2008), and a multiscale method proposed by Lack et al. (2010). MODE has also been applied to convective-scale ensemble forecasts of precipitation (Gallus 2010), and more recently to other forecast fields such as for the comparison of convective-scale cloud forecasts (Mittermaier and Bullock 2013). This included a first exploration of using MODE in the time dimension.

Although the application of spatial methods addresses many of the limitations of traditional skill scores, for global models in particular a comprehensive diagnosis and understanding of model performance still requires further consideration of the model in terms of its dynamical and thermodynamical forecast fields. This allows individual forecasts to be assessed in detail to consider both large- and small-scale meteorological features such as the 250-hPa jet stream, 850-hPa temperature gradients, and driving shortwave and longwave upper troughs; any of which can be significant and critical features in the evolution of a forecast.

When performed manually, such verification procedures are often termed “subjective” because they rely on interpretation of numerical model products, as opposed to “objective,” which are mathematical techniques that require no interpretation in their production. The interpretation necessary for subjective verification techniques is based both on a sound knowledge of meteorology and modeling and also on the experience gained in the assessment of many previous forecasts. It is in the application of combined knowledge and experience throughout a forecast that invaluable insight can sometimes be provided into model problems that are difficult to disclose through objective verification alone. These subjective verification techniques together with traditional and object- or feature-based methods provide a comprehensive understanding of model performance.

This underpins a key aim of this paper, which is to show that object-based spatial methods can replicate aspects of a subjective assessment process in a quantifiable manner. During testing of a model upgrade, different model configurations (generally two but possibly more) need to be compared. To enable this they are typically run over the same prespecified period, of between 1–3 months. When using a method such as MODE, how can the impact on object attributes due to model configuration changes be quantified? This has to date not been fully explored. In this paper a new score is introduced to compare paired object attribute distributions from two or more model configurations, both against each other, and also to what would be the perfect paired object attribute distribution. Section 2 provides an outline of the model configurations. Section 3 is devoted to the process of subjective evaluation. Section 4 provides an introduction to MODE and the methodology used, including the skill score definition. This is followed by the results in section 5. A discussion and conclusions are provided in section 6.

2. Model configurations

The Met Office Unified Model (UM) provides a nested modeling system that enables global long-range simulations for climate projections and regional short-range NWP forecasts at the kilometer-scale. In this paper a three-way comparison of global NWP configurations is used to illustrate an evaluation methodology, designed to test the full impact of a combined model resolution upgrade on top of a dynamical core change, alongside the pure impact of only changing the dynamical core at the same horizontal resolution, and just
changing resolution using the same dynamical core (though this was only tested for the new dynamical core). The horizontal resolutions are N512 (~25 km in the midlatitudes, ~40 km in the tropics) or N768 (~17 km in the midlatitudes, ~26 km in the tropics). The two dynamical cores are the operational version up to 15 July 2014 [new dynamics (ND)] (Davies et al. 2005) and the new dynamical core [even newer dynamics (EG)] as described by Wood et al. (2014). For the trials all three configurations were run twice a day at 0000 and 1200 UTC. Forecasts were produced to \( t + 144 \) h. There were two trial periods: late June to mid-September [Northern Hemisphere summer (NHS) and Southern Hemisphere winter (SHW)] and November–December 2012 [Northern Hemisphere winter (NHW) and Southern Hemisphere summer (SHS)].

An independent mean sea level pressure (MSLP) analysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecast System (IFS) (Hortal 2002) cycle 38r1 with a native resolution T1279 (16 km) was used as reference data, regridded to match the two horizontal resolutions listed earlier. The ECMWF spherical harmonics data are converted to the regular UM latitude–longitude grid. The spherical harmonics are truncated to a wavenumber dependent on the requested grid spacing using bilinear interpolation. This was done to ensure that all configurations were compared to the same analysis. The model configurations are therefore labeled as follows: N512ND (control, previous operational), N512EG (new dynamical core only), and N768EG (new dynamical core and resolution increase).

3. Subjective assessment

The subjective assessment of model performance documented here consists of two main approaches. The first is the direct comparison of model fields of variables that are important to understand and diagnose the model evolution: these include MSLP, geopotential height (GPH), temperature, wind, vorticity, thicknesses, and humidity. Random sampling of the forecasts during a model trial or daily monitoring of the forecasts enables the general differences between the model characteristics to be highlighted.

The second approach aims to identify and quantify the forecast errors of the models and use this information as a vehicle to disclose more about the model characteristics than can be discovered by direct comparisons alone. This method will be discussed in detail here and requires the subjective assessment and interpretation of forecast errors of selected parameters that are compared with those of a reference model. In this instance the model under study is the model configuration with the new N768EG dynamical core and this is compared with the N512ND configuration of the global model (GM), which was operational at the time of the assessment. Forecast errors (relative to the verifying UM analysis) of 500-hPa GPH, 1000–500-hPa thickness, and MSLP for each forecast in the trial period have been calculated for all forecast ranges to give a clear and representative signal of model error. Being absolute quantities these errors are less susceptible to the problems originating from the characteristics of any one skill score and are here initially used to obtain the location and magnitude of the largest positive and negative errors within selected domains. The errors are plotted as a function of verification date for each forecast range and provide an indication of the consistency of model performance at that forecast range in each domain. By plotting the positive and negative errors independently in this way it is often clear from the resulting time series whether the forecast error is a result of a displacement or an over- or underdevelopment or a combination of both; for example, overdevelopment of a depression results in a large negative error and underdevelopment results in a large positive error, while an error in the location of a depression will be associated with both large positive and negative errors. Further, sharp departures of the forecast error from the more usual error magnitudes are indicative of forecast “busts” and provide a means to their identification. Objective assessment of these quantities for both the trial model and the reference model enables significant departures in behavior of the models in terms of the forecast errors at that forecast range to be identified and this provides a basis from which to perform further subjective investigations.

Figures 1 and 2 show the \( t + 120 \)-h 500-hPa GPH and MSLP forecast errors for Europe (30°–80°N, 10°W–40°E), respectively. The data show that in general the two model configurations are comparable in terms of the magnitude of the largest 500-hPa GPH forecast errors and this is reassuring in the assessment of the new model. However, while positive MSLP errors for the trial model are generally similar or smaller, negative MSLP errors for the new N768EG model configuration are generally slightly larger than those of the operational model. These larger MSLP errors for the N768EG model show that this particular model configuration has a greater tendency to increase errors of over-deepening relative to the old N512ND configuration, but decrease errors associated with underdeepening.

The \( t + 120 \)-h error time series show a few instances in which there are significant differences in the forecast errors of the two models: 14 November, 18–20 November, 24–26 November, and 7 December. To illustrate the
The methodology of further investigation the forecast of
0000 UTC 7 December will be discussed.

Figure 3 shows the $t_{120}$-h 500-hPa GPH error fields
verifying at 0000 UTC 7 December, which corresponds
to the large negative 500-hPa GPH error detected in the
N768EG configuration but not in the N512ND configuration.
The error can now be observed as arising from
an overamplification of a trough extension (negative
errors) over the Bay of Biscay, which appears in the
N768EG configuration but not in the N512ND model.
Further, this error has some longevity to it as it propa-
gates through to the next main cycle of the model (at an
improving rate) as identified by the forecast error veri-
fying 12 h later on the time series (Fig. 1) and verified
through the error fields (not shown).

Once a significant forecast error in the trial model has
been identified the origins of model errors may be de-
duced by use of tracking techniques using a suitable
model field such as 500-hPa GPH. These tracking
techniques involve the calculation of differences be-
tween the model forecast under investigation and a
chosen reference; the reference being either the verifying
analysis throughout the forecast or another model
forecast.

Figure 4 shows a sequence of forecast minus analysis
differences for the N768EG model at 0000 UTC
2 December forecast leading to the $t_{120}$-h error
identified in Fig. 3. These fields have been generated using
autoscaling throughout the forecast range so as to enable the errors to be tracked as they develop. The negative error over Biscay at $t_{120}$ h can be traced back
through the forecast into the North Atlantic to the south
of Iceland at $t_{96}$ h; here the model has incorrectly
developed an upper trough where the analysis shows a
vortex. The forecast upper trough is a more progressive
feature than the analyzed vortex and creates a positive/
negative dipole in the error field. The $t_{72}$-h error field
shows that the negative error, which eventually be-
comes the Biscay error, develops on the leading edge of
an error wave train that extends eastward from the negative
error over the Hudson Bay at this time. This Hudson
Bay error is due to an overextension of the upper
trough in this location that results in an overamplification
of the downstream synoptic waves and in the eventual
overextension of the Biscay trough. This error can be
traced back further through the forecast and is associated
with a strong error signal at $t_{24}$ h when the trough is at
52°N, 117°W; at this point the trough is significantly too
extended. The first frame of the error fields at $t + 12$ h shows what appear to be minor errors in the trough situated over the west coast of North America.

With the source region now clearly identified, 250-hPa GPH analysis differences between the trial model and the operational model (Fig. 5) show clearly that the major upper trough in the eastern Pacific has been analyzed differently in the two models and that these differences are the origins of the $t + 120$-h forecast error of the trial model (this is further west than in the first error field of Fig. 3 being 12 h earlier). In the trial model the trough is slightly sharper than in the operational model as indicated by positive differences on the trough’s forward flank. Further, a comparison of 250-hPa jet cores between the two models (not shown) indicate that the sharper trough in the new model is accompanied by slightly stronger wind speeds running around the base of the trough. These two features are consistent with the characteristics deduced from both the error time series and from the objective verification scores.

The differences at $t + 0$ h identified here between the two models are not a result of differences in the assimilation schemes as these are identical in the two model configurations. Differences between the two models, however, can occur at $t + 0$ h because each model configuration is running independently with differences in the background states used in the assimilation scheme naturally arising and propagating from cycle to cycle. These differences could either be due to the model’s new dynamical core or else a consequence of the new model’s higher resolution. Indeed, the repeated use of an earlier forecast within the data assimilation cycle is likely to amplify any overdevelopment characteristics of a model as observed here because the model background will itself contain developments that are overamplified. Irrespective of the exact cause at this stage, this is valuable information because if it is found that the same characteristic errors occur over a number of cases it can lead to systematic causes of forecast error to be established and research priorities to be determined. Further, the systematic origins of forecast error established through these means can be used to better interpret the model characteristics identified through the objective verification. Model characteristics are not only important for development purposes in focusing research on model problems, but are also important for operational forecasters in assessing model guidance and adding value to NWP output.

4. Objective evaluation methodology

a. Method for Object-based Diagnostic Evaluation

MODE (Davis et al. 2006a,b) is a spatial verification method that provides the ability to identify features
FIG. 3. The 500-hPa GPH errors (m) for the Northern Hemisphere, 0000 UTC 7 Dec 2012 + 120h. A significant difference between the two model configurations is identified in western Europe associated with a trough extension.
(referred to as objects) of interest in the forecast field, and focus only on these features when assessing forecast performance. The process of identifying objects involves several steps. Initially a convolution filter may be applied to the raw data field to smooth it, if required. Next, a threshold is applied to the raw or smoothed field, giving a binary or mask field. The original data are then put back into the masked field, restoring the intensity
information. Once the objects have been identified, a variety of object attributes can be calculated, such as area, centroid, orientation angle, and intensity percentiles.

The object attributes facilitate the processes of matching and merging. The process of merging considers the association of objects in the same field with each other. For example, MODE enables two forecast simple objects that are in close proximity and belong together to be grouped in to what is called a cluster object. Cluster object attributes are also calculated. The process of matching on the other hand is the association of objects in different fields, allowing forecast objects to be paired with observed objects. Single cluster objects were used for matching in this study. Attributes for single objects are used to calculate attributes for pairs of objects, one object from the forecast field, and one from the observed field. There are exceptions to this rule (e.g., area overlap, which cannot be inferred by considering the two single objects in isolation). Paired statistics such as hits and misses can be calculated from the matched objects. Both the matching and merging processes use a fuzzy logic engine to calculate a total interest value for each possible forecast–observed object pair. This engine uses the simple object attributes calculated, as well as paired attributes such as overlap area, centroid distance, difference in orientation angle, and differences in intensity distributions. Finally a threshold is applied to the total interest to determine which are the best matches.

b. Comparing distributions

When it comes to using MODE for assessing the impact of model changes on forecast accuracy and skill there are arguably several matched pair attributes that are of primary interest:

1) **Centroid difference**—the location error (expressed in grid lengths that can be multiplied by the grid spacing to get real distance);

2) **Angle difference**—the orientation of the objects with respect to each other [i.e., the rotation error (in degrees)];
3) Area ratio—an estimate of the spatial extent bias between paired forecast–analysis objects exceeding a given intensity (bounded to fall between 0 and 1);
4) Intersection-over-union area—the physical overlap between paired forecast–analysis objects in spatial terms (bounded to fall between 0 and 1). This is analogous to the critical success index (CSI) or threat score (TS) (see e.g., chapter 3 in Jolliffe and Stephenson 2011).

A distribution approach can be useful in establishing the differences in model configurations, to determine whether the distributions have shifted, contracted, or expanded relative to each other. The individual paired object attribute values across all forecasts in the sample are turned into empirical cumulative relative distributions. Simple quantile–quantile plots of attribute distributions (not shown) indicate they are not normally distributed, eliminating the use of any parametric goodness-of-fit tests for comparing distributions, these assume normality.

The classic Anderson–Darling (AD) test (Anderson and Darling 1952) is a nonparametric test designed to establish whether the empirical cumulative distribution of a sample $F_n$ (size $n$) was drawn from a continuous population based on a fully described or known distribution function $F_{ref}$. The test statistic $A^2$ is a sum of the integrated squared differences between two distributions functions:

$$A^2_n = n \int_{-\infty}^{\infty} \frac{(F_n - F_{ref})^2}{F_{ref}(1 - F_{ref})} dF_{ref}. \hspace{1cm} (1)$$

The two-sample test was introduced by Darling (1957) and compares samples from two independent populations $F_n$ and $G_m$, using a combined distribution $H_N = (mF_n + mG_m)/N$ with $N = m + n$. In the two sample case the test considers the hypothesis that $F = G$ without specifying a common (continuous) distribution function so that the test statistic becomes

$$A^2_{nm} = \frac{mn}{N} \int_{-\infty}^{\infty} \frac{(F_n - G_m)^2}{H_N(1 - H_N)} dH_N. \hspace{1cm} (2)$$

The $k$-sample AD test (Scholz and Stephens 1987) is a further extension of the test, and is applicable to both continuous and discrete parent populations. This has been implemented in the R package kSamples. One of the purposes of this test is to determine whether differences in several sampled populations are significant, with particular sensitivity toward the tails of the pooled sample. This is in contrast to the more popular and well-known Kolmogorov–Smirnov (KS) test that only considers the maximum difference between two distributions, both in its continuous and discrete forms (e.g., Arnold and Emerson 2011). For the $k$-sample AD test the null hypothesis is that empirical distribution functions $F_1 = F_2 = \cdots = F_k$, without specifying a common distribution $F$. The test statistic then becomes a further generalization:

$$A^2_{kN} = \sum_{i=1}^{k} n_i \int_{B_N} \frac{(F_i - H_N)^2}{H_N(1 - H_N)} dH_N, \hspace{1cm} (3)$$

where $B_N \in \mathbb{R}$ and $H_N < 1$. For $k = 2$ Eq. (3) reduces to Eq. (2). Under $H_0$ and assuming a continuous common distribution $F$, the expected value of $A^2_{kN}$ is $k - 1$. The test statistic is standardized using the expected value and the variance of $A^2_{kN}$, $\sigma_N$:

$$T_{kN} = \frac{A^2_{kN} - (k - 1)}{\sigma_N}. \hspace{1cm} (4)$$

The variable $H_0$ is rejected at significance level $\alpha$ if $T_{kN}$ exceeds the critical value $t_{k-1}(\alpha)$. The samples do not need to be the same size and can be as small as 5 for the asymptotic standardized percentiles to be used as approximate critical values. For more detail the reader is referred to Scholz and Stephens (1987) and the R package kSamples.

In subsequent sections the $A^2_{kN}$ statistic is calculated for two- and three-sample cases where either pairs of model configurations are compared, or they are considered all together. The analysis focuses on the $p$ values at an $\alpha$ of 0.05 to determine whether, as a function of lead time, any of the differences in the attribute empirical distributions are significant.

c. Experimental design

The evaluation is focused on analyzing the driving synoptic features that are associated with distinct surface weather and associated impacts. Table 1 lists the features and thresholds used for this study. Strong jet cores can trigger rapid cyclogenesis. Deep surface low pressures can be associated with potentially damaging winds, heavy precipitation and can contribute to coastal impacts associated with storm surges. Strong high pressures can be associated with a range of surface weather impacts depending on the season, from heat waves accompanied by generally poor air quality, to cold snaps with reduced visibility and persistent fog.

The experimental design enables the evaluation of the following:

- **Spatial** biases, considering the extent of features. Are they over or underestimated?
- Changes in **intensity**, are features deeper or stronger?
- Changes in the number of analysis and forecast objects, as a “contingency table” view of hits, false alarms, and misses in a feature-based context.
Changes in the attribute distributions, are the forecast attribute distributions of the “test” configurations closer to perfection?

Note that MODE only outputs a bounded area ratio by default, whereby the reciprocal is taken when values are greater than 1. While mathematically elegant, this means there is no way of distinguishing between under- and overestimation of spatial extent. For this evaluation an unbounded area ratio was calculated, in order to assess the spatial over/underestimation, and is considered alongside the bounded version.

5. Results

This section covers but a subset of the total results, and attempts to provide an overview of key findings. Forecasts are compared at 12-hourly intervals from $t + 12$ to $t + 144$ h, with both 0000 and 1200 UTC initializations assessed.

Figure 6 provides a global view of identified MODE forecast objects with analysis outlines. The identified low and high pressure centers, and jet cores for a single N768EG $t + 48$-h forecast valid at 0000 UTC 9 December 2012 are shown. The different colors simply help identify distinct objects. Despite being austral summer, pressures are much lower over the Southern Ocean in Fig. 6a. By contrast there are much smaller distinct low pressure centers in the Northern Hemisphere storm track. For this reason a lower threshold of 970 hPa was also considered for the Southern Hemisphere. The Siberian high in Fig. 6b dominates over large parts of northern Asia, with the largest jet core in Fig. 6c to the east of the Tibetan Plateau.

The area overlap (union) of individual paired forecast and analysis objects can be aggregated over the November–December 2012 trial period to understand the prevalence and coverage of the identified features, and to investigate the proportion of the time during the trial period that the forecast and the analysis had a feature exceeding the threshold at a particular location. This is shown in Fig. 7. Here the N768EG $t + 24$-h results for the 0000 UTC runs are used to illustrate (summer trial period not shown). In Fig. 7a lows deeper than 990 hPa highlight that despite it being austral summer there is a prevalence of deep lows in the Southern Ocean. The impact of the landmasses in the Northern Hemisphere is also clear, preventing a continuous storm
track encircling the Northern Hemisphere. Figure 7b shows the frequency of 1030-hPa highs for the same trial. Generally the North Atlantic storm track is surrounded by higher pressure, with a frequency of 30%–40% for the trial period. High pressure dominates large parts of the Arctic and continental parts of the Northern Hemisphere. The jet cores in Fig. 7 show a much more even distribution between the summer and winter hemispheres, with the persistent presence of the jet emerging off the Tibetan Plateau.

Considering Figs. 6 and 7 together it would appear that the N768EG configuration (in this case) is capturing the events well (as defined by the analysis). A review of the trials using figures such as Fig. 6 would suggest that at the earlier lead times this is fairly typical. This is the first time that global forecasts have been considered in this way.

a. Spatial biases

For the spatial extent bias, all identified features in the binary forecast and analysis exceedance field are considered. Results were not stratified hemispherically but aggregated globally for the two trial periods, therefore the distribution of objects shown in Fig. 7 is important to establish which hemisphere dominates the results. The features occur naturally within a limited range of northern and southern latitudes (i.e., are naturally stratified). For the spatial biases the 970-hPa threshold is shown for the low pressures, as Fig. 7 demonstrated that 990 hPa is potentially too high a threshold for the Southern Hemisphere. This means that the biases shown in Figs. 8a and 8b are predominantly for the Southern Hemisphere. For the November–December period the extent bias is fairly neutral though N768EG 0000 UTC runs in particular are favoring slightly larger values at longer lead times, the bias decreasing with decreasing lead time so that at $t + 12$ h all the configurations are underestimating extent slightly. Other configurations are more neutral overall. For the June–September period there are signs that the forecast lows become comparatively larger closer to analysis time. Trends between the configurations as a function of lead time are mixed with some hints that the N768EG configuration has larger lows from day 6.

For the high pressures, results are dominated by the Northern Hemisphere for the November–December 2012 trial and by the Southern Hemisphere for the June–September 2012 trial (see Fig. 7b). Figures 8c and 8d show very distinct behavior between the two trials and hemispheres. In Fig. 8c the extent of high pressures stronger than 1030 hPa is underestimated at all lead times, with the bias improving with decreasing lead time. The N768EG results appear to be clustered at the lower
FIG. 8. Spatial extent biases for selected thresholds for identified lows, highs, and jet cores for the specified trial periods, aggregated globally. (a),(c),(e) The November–December 2012 trial and (b),(d),(f) the June–September 2012 trial. The Northern and Southern Hemisphere results are combined. The lead time is in hours.
end of the spatial extent bias, with substantial differences to N512ND beyond day 3 of the forecast. During the Northern Hemisphere (boreal) summer the trend is reversed with high pressures slightly too extensive at long lead times, though the bias is relatively stable beyond day 3. There is a steady reduction in the bias with decreasing lead time from days 3 to 1, so that by day 1 the extent is underestimated. The differences between the ND and EG configurations become more marked beyond day 2. It is hypothesized that the reasons for the changes in bias are related to the response of the Eurasian landmass to heating and cooling, and the comparative lack of land at the same latitudes in the Southern Hemisphere.

The clearest impact in spatial extent biases is seen for the 250-hPa jet cores, where Figs. 8e and 8f show a substantial swing, and for the most part, a change from underestimating to overestimating the spatial extent for EG configurations. The EG dynamical core is increasing the spatial extent of the jet cores. The trend in spatial extent is also more stable as a function of lead time. Note that both hemispheres contribute more evenly for the jet cores, as seen in Fig. 7c.

b. Intensity

MODE retains the intensity information for all identified objects, which can be assessed as percentile distributions, below (above) the object threshold. In this case Northern and Southern Hemisphere results were kept separate. Though the overall minimum (for lows) and maximum values (for highs and jets) are identified, statistically it is more robust to consider the 10th or 90th percentiles to determine whether the intensity of identified synoptic features has changed. To illustrate this the percentiles comparing N768EG and N512ND for the NHW objects are plotted as scatterplots in Fig. 9. In each of the scatterplots the analyses are indicated by “0” with 0000 UTC initializations in black and 1200 UTC initializations in gray. For the lows deeper than 990 hPa the difference between the 0000 and 1200 UTC 10th percentile EC analyses pressures is more than 5 hPa. The forecast ranges are more scattered for the 1200 UTC initializations than for the 0000 UTC initializations. Generally N768EG 10th percentiles are lower by as much as ~5 hPa, suggesting a substantial shift in the tail of the distribution.

For the high pressures the opposite is true. Here Fig. 9 shows that the analyses are closer together (~1 hPa), and the 0000 and 1200 UTC initializations are generally more clustered together. There is a clear upward shift to higher pressures in the N768EG 90th percentile values for the NHW by ~2 hPa.
For the 250-hPa NHW jet cores, the results suggest a predominant strengthening in the N768EG 90th percentile wind speeds, especially for the 1200 UTC initializations, but for the 0000 UTC initializations there are indications that for the early lead times the differences are more marginal.

c. Events

The potential drawback of having a more active model is, for example, the spinning up of spurious low pressures. A contingency table–based events assessment can reveal to what extent this may be the case. This is shown for NHW objects in Figs. 10–12. Each figure shows the number of matched objects (hits), as well as the number of unmatched forecast objects (false alarms), and the number of unmatched analysis objects (misses). Also shown are the matched and unmatched object areas, which refers to the total of all the union areas for the matched pairs, and the total of all the unmatched forecast and analysis objects. Generally it is desirable to maximize the hits, and minimize both the false alarms and misses. Missed events are generally undesirable especially if they can be associated with high-impact weather. False alarms can also be undesirable, especially if excessive (also called the “cry wolf” problem).

For the low pressures below 990 hPa Fig. 10 shows there is a clear increase in the number of matched objects (hits) for the EG configurations, especially at longer lead times where the N512ND matched objects tail off more quickly. There is a commensurate increase in the number of false alarms in the EG configurations with N768EG often having the largest number of false alarms. Reassuringly N768EG often has the lowest number of missed events. The matched area shows a substantial improvement in terms of matching the analysis, while N768EG has the largest unmatched areas suggesting that some of the missed events are rather large in spatial extent. This is in line with the spatial extent biases shown earlier.

For the surface high pressures in Fig. 11 the total number of identified N768EG objects is also substantially increased, with the margin of increase maintained for most lead times. The EG configurations have more false alarms though the ordering is reversed with N512EG suggesting more false alarms compared to N768EG. The misses suggest an intriguing diurnal signal, with the misses increasing with lead time, and configurations again very similar. This pattern originates from the analyses where there are distinct fluctuations between 0000 and 1200 UTC. MODE is a threshold-based methodology. It would appear as though the high pressures are more sensitive to these small 1–2-hPa fluctuations. This may again be attributed to the impact of continental heating and cooling. The N768EG configuration has the largest matched area, but also the largest amount of unmatched N768EG area for almost all lead times, suggesting an overall increase in the area above the threshold, and a significant proportion that has not been matched in the analysis. The N512 configurations are less distinguishable from one another.

Figure 12 shows the results for the 250-hPa jets. There is an increase in the number of matched objects (hits) with the EG dynamical core, with an additional increase with resolution. There is a commensurate increase in the number of unmatched forecast objects (false alarms), with a comparatively large increase between N512ND and N512EG. On the other hand there is no discernible difference between the configurations for the missed jet cores. The matched area shows an increase with N768EG at the upper end for all lead times. There is a steady upward shift across all lead times between N768EG unmatched area relative to the other configurations.

d. Attribute distributions

MODE produces a range of attributes for the forecast–analysis pairs. As described in section 4 one such attribute is the area ratio. The 1200 UTC initialization area ratio attribute distributions for NHW 990-hPa lows are shown in Fig. 13 for all three configurations. Six lead times are shown separately to highlight how the distributions evolve (i.e., as the forecast error grows). The perfect attribute distribution for area ratio would be where there is a perfect overlap between all forecast and analysis objects yielding area ratios of 1 (i.e., a cumulative distribution function that is a step function that goes from 0 to 1 at area ratio equal to 1). Even at $t + 24$ h it is clear that the area ratios already have a substantial tail, which increases with lead time. Superficially there appears little difference between the distributions for the three configurations. As described in section 4 the $p$ value of a $k$-sample AD test is used to determine whether the differences are significant or not. While previous sections indicate that the number of matched objects decreases with increasing lead time, statistically sample sizes are in excess of 50 at all lead times, which is more than sufficient for $p$-value calculation.

The results of the $k$-sample AD test for NHW 990-hPa lows are shown in Fig. 14. Here the 0000 and 1200 UTC initializations have been combined. Considering the five attributes there are very few instances of statistically significant differences between the attribute distributions from the three model configurations. Only angle difference and the unbounded area ratio show any specific signals. The N512EG-N512ND comparison of...
FIG. 10. Events assessment of the number of hits, false alarms, and misses for NHW lows deeper than 990 hPa as a function of forecast lead time (h).
FIG. 11. As in Fig. 10, but for NHW highs stronger than 1030 hPa.
FIG. 12. As in Fig. 10, but for NHW jets stronger than 60 m s$^{-1}$. 
angle difference distributions highlights two lead times at $t + 60$ and $t + 32$ h as being statistically significant, suggesting there is something in the dynamical core change at the same resolution that is causing a change or shift. The $p$ values for N768EG-N512ND are also comparatively low while the difference between EG configurations at different resolutions yields a very large $p$ value. A very interesting outcome is the unbounded area ratio that suggests a very large block of lead times from $t + 12$ to $t + 84$ h with statistically significant differences, driven it would appear by the change in the dynamical core, as the resolution change tested in N768EG-N512EG shows large $p$ values by contrast. While the $k$-sample AD test cannot say whether the distributions are getting better, or closer to our understanding of perfection, it is highlighting a statistically significant shift or change. In the case of the unbounded area ratio the perfect distribution would be a step function going from 0 to 1 at an unbounded area ratio of 1. Why would the unbounded area ratio show such strong results and the bounded area ratio show nothing? Plotting the distributions it would appear that the area ratios that are greater than 1 and are allowed to remain greater than 1 (rather than take the reciprocal), provides a different distribution shape, especially in the extremities, and as the $k$-sample AD test is more sensitive to differences in the tails it would suggest that this is where the significant differences arise.

Figure 15 shows the results for NHW 1030-hPa high pressures. Here too statistically significant differences are hard to find with only isolated lead times with $p$ values less than 0.05. There are a scattering of lead times and attributes with $p$ values between 0.05 and 0.1. There is nothing systematic that would support the assertion that either the resolution change or the dynamical core change have had the bigger, more statistically significant impact on the attribute distributions.

Finally Fig. 16 shows the results for the NHW 250-hPa jet cores. Four of the five attributes show some statistically significant differences, though arguably only two hint at more systematic differences with lead time: centroid difference and the unbounded area ratio. Again the dynamical core change yields $p$ values less than 0.05 between $t + 36$ and $t + 60$ h, though the resolution increase in N768EG-N512EG shows no significant difference. The combined dynamical core and resolution change is also statistically significant. The unbounded
area ratio also shows statistically significant differences between $t + 24$ and $t + 72$ h with the same pattern as for centroid difference. The dynamical core change is significant and the combined change is also significant at these lead times. Unlike in Fig. 14, the (bounded) area ratio also has a scattering of lead times with $p$ values less than 0.05 and a few in the range of 0.05–0.1, suggesting a slightly different distribution of area ratios for the jets than the lows.

6. Conclusions and discussion

In this study MODE has been used to investigate how a change in dynamical core and resolution affects
the evolution of synoptic features that drive surface weather, and how this compares to a subjective assessment. Rather than consider each point in the forecast grid for evaluation only those features of interest were selected and assessed. Surface lows and highs as well as 250-hPa jet cores were evaluated.

The results suggest that the spatial extent biases change with the dynamical core and with an increase in horizontal resolution. Jet cores and low pressures are bigger, while high pressures are smaller.

In addition the object intensities suggest that the new dynamical core combined with the resolution increase is providing deeper lows and stronger highs and jet cores. Combined with the increase in spatial extent of low pressures and jets, and a decrease in the extent of high pressures, stronger larger jets are leading to larger deeper lows and stronger smaller highs, producing tighter pressure gradients, and resulting in a more energetic and active model atmosphere.

FIG. 15. As in Fig. 14, but for NHW 1030-hPa high pressure attribute distributions.
The number of hits is increased for the new dynamical core, especially N768EG. False alarms also increase, while misses are reduced or remain predominantly neutral.

The $k$-sample AD test shows only a limited number of the lead times with statistically significant differences between the various model configurations’ attributes. The $k$-sample AD test was chosen because it considers the whole distribution. The alternative was to restrict the discussion to a comparison of the medians and variances, which is analogous to the use of, for example, box-and-whisker plots. The question then arises, why so little significance? Perhaps expectation of more is unrealistic. In reality model development is based on an incremental approach, so while there is the aspiration that changes in the model have a significant impact, individual steps are often relatively small, potentially difficult to detect with confidence, and backed by statistical significance. For this analysis sample sizes were well in excess of 100 in most instances, and statistically

![Table](image)

**Fig. 16.** As in Fig. 14, but for NHW 250-hPa jet core attribute distributions.
the $k$-sample AD test can provide significant results with good power for samples as small as 5. But beyond the statistics, the sample size is determined by the length of the trials, which are possibly insufficient in length to sample a variety of flow patterns. The zonal wave-number and the general state of amplification of the zonal flow, phases of the North Atlantic Oscillation (NAO) and similar large-scale teleconnection patterns can persist for long periods and provide consistent (disturbed or very static) conditions for several weeks at a time, producing fairly homogenous samples. Trials are unlikely to get longer, both in terms of computing time and cost, nor are the length of time windows available for trialing between operational upgrades. Therefore, sample size in terms of sampling different global flow patterns remains a challenge, and feature-based methods are not necessarily exempt from these sampling problems. In the case of MODE object attributes, sample size also depends on the thresholds used to define the objects.

As seen in Figs. 10–12 the number of matched objects decreases with lead time. The main reason is clear: the forecast error grows with time, or put differently, over time forecasts converge on the final observed outcome. Low pressures may take a different track, or change in intensity and drop below the threshold, thus dropping out of the analysis, and reappear again. For this study the search radius for finding matches is kept constant with lead time. If the emphasis had to try to capture and track the same low pressures as far as possible, then a larger search radius would be needed at longer lead times. However, as this analysis is not intent on tracking features such as low pressures over their lifetimes, this issue is less important. Another question may then be whether the longer lead times are dominated by cases that are more strongly forced. This is not clear, but features identified and matched beyond day 5 certainly appear to be more predictable. A feature-based method such as MODE may be vulnerable to spurious matches, but otherwise the method is insensitive to or independent of forcing or predictability.

It is also worth dwelling on the uncertainties that exist in using the evaluation approach described in this paper. First, as already alluded to, there is the dependence on the attributes derived by MODE. The biggest uncertainties arise in the merging and matching approach that MODE utilizes, which is controlled by setting a range of tunable parameters (e.g., intensity and size thresholds and search radii). It is true to that for any given case visual inspection might make a human, using a more subjective approach, decide to merge and match objects differently. The key to MODE and methods like it is to test the tunable parameters so that, on average, they give the same result as would be obtained in a more subjective manner. The whole premise of a method such as MODE is to be able to automate this process and consider a much larger volume of data.

Another source of uncertainty is the gridded truth data. Model analyses are often frowned upon as a form of “truth,” because of the dependence of the analysis on the underlying model (e.g., Ebert et al. 2013). Here ECMWF analyses were used from an independent modeling system [as, e.g., also used by Clayton et al. (2013)], in an attempt to mitigate against this, and this is the preferred solution if analyses are used. A threshold-based method can still be sensitive to subtle biases in the different modeling systems, therefore, care is required when interpreting results, especially for surface fields as the treatment of the near-surface varies between the models from different centers. There are also diurnal variations, which would be present in any model analysis. The fluctuations in the 0000 and 1200 UTC analyses have certainly had an impact on results, introducing diurnal variations in a usable statistic, particularly for the high pressures. One possible future development of MODE is the incorporation of percentile thresholds that would remove any visible bias.

Spatial and temporal correlation could also be considered a source of uncertainty. Some of the features are fairly temporally and/or spatially persistent (e.g., some jet cores and the Siberian high), while others have a much stronger life cycle and track and downstream progression. More recent versions of MODE are incorporating the ability to also track objects in time [as first explored by Mittermaier and Bullock (2013)]. In this study the implicit assumption is that successive forecast ranges are independent of each other. Future analyses may explore the impact of this assumption more carefully.

Some of the objects are rather large so that the position of the centroid alone is perhaps not sufficient to quantify the location error and any difference between various model configurations producing forecasts valid at the same time. This is to some extent mitigated against by considering a range of attributes together (e.g., centroid difference with area ratio). There will undoubtedly be cases where the overlap between forecast and truth objects might be quite good but the calculated centroids are quite some distance apart. There is another attribute that was not considered appropriate for use in this study, which considers the shape of the object as well.

MODE is a tool that continues to be developed and refined. As such MODE was not designed for global applications, but this study proves that it can be used. One particular issue is that MODE has no notion of...
cyclical boundaries, which can lead to the splitting of objects at whichever longitude the grid is “unwrapped” (e.g., the date line). This was not found to be a significant problem for the datasets used for this study, with very few split objects. Split objects, if correctly forecast, would still be analyzed (provided they exceeded the minimum size) but as two separate entities rather than as one. Another grid-relative issue arises near the poles where grid lines converge, but in this study objects are primarily stratified latitudinally and objects are subject to the same effects within a latitudinal band. It was agreed that some attributes should not be used as part of this study because of the latitudinal dependence. The shape attribute is one.

The study demonstrates that MODE provides results consistent with subjective assessments and that the attributes of the forecast features can provide valuable information to the forecast user for how to best interpret forecasts on a day-to-day basis.

Acknowledgments. Marion Mittermaier would like to thank NCAR-DTC for the support provided.

REFERENCES


