A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites

MARION P. MITTERMAIER
Numerical Modelling, Weather Science, Met Office, Exeter, United Kingdom

(Manuscript received 26 July 2012, in final form 19 July 2013)

ABSTRACT

Routine verification of deterministic numerical weather prediction (NWP) forecasts from the convection-permitting 4-km (UK4) and near-convection-resolving 1.5-km (UKV) configurations of the Met Office Unified Model (MetUM) has shown that it is hard to consistently demonstrate an improvement in skill from the higher-resolution model, even though subjective comparison suggests that it performs better. In this paper the use of conventional metrics and precise matching (through extracting the nearest grid point to an observing site) of the forecast to conventional synoptic observations in space and time is replaced with the use of inherently probabilistic metrics such as the Brier score, ranked probability, and continuous ranked probability scores applied to neighborhoods of forecast grid points. Three neighborhood sizes were used: ~4, ~12, and ~25 km, which match the sizes of the grid elements currently used operationally. Six surface variables were considered: 2-m temperature, 10-m wind speed, total cloud amount (TCA), cloud-base height (CBH), visibility, and hourly precipitation. Any neighborhood has a positive impact on skill, either in reducing the skill deficit or enhancing the skillfulness over and above the single grid point. This is true for all variables. An optimal neighborhood appears to depend on the variable and threshold. Adopting this probabilistic approach enables easy comparison to future near-convection-resolving ensemble prediction systems (EPS) and also enables the optimization of postprocessing to maximize the skill of forecast products.

1. Introduction

There are many benefits of high-resolution NWP, including more realistic precipitation (e.g., compared to weather radar), wind, and cloud structures, with a closer tie to orography (e.g., Mass et al. 2002; Schwartz et al. 2009; Smith et al. 2014). While it has been relatively easy to show an increase in skill when comparing forecasts from models with from >30-km to <20-km horizontal resolutions, it has been somewhat more difficult to do so consistently for horizontal resolutions of >10 km to so-called convection-permitting grid resolutions (e.g., 4 km). This has been especially true for precipitation, where the forecast details are realistic but inaccurate. Mass et al. (2002) noted that precipitation scores were degraded from 12 to 4 km with an increased overestimation over windward slopes of higher ground, although for more intense precipitation 4-km forecasts provided some enhancement of forecast skill. Mass et al. (2002) also showed marginal increases in temperature and wind scores as grid spacing decreased from 12 to 4 km. They also noted that the improvements at 4 km were often masked by timing errors and were only evident when the events verified were filtered to only include occasions where the synoptic scale was well forecast. Mittermaier et al. (2013) also revealed that a spatial verification technique was required to show conclusively that Met Office Unified Model (MetUM) 4-km precipitation forecasts are more skillful than 12-km forecasts. Forecasters, on the other hand, intuitively make allowances for the space–time errors in high-resolution NWP.

Lorenz (1969) already pointed out that output from kilometer-scale models may not be accurate at the grid scale. This inaccuracy introduces the so-called double-penalty effect where the traditional approach to forecast verification of precise matching of forecasts to point observations means that even small displacement errors get penalized (increasing the false alarm rate), and closeness is not rewarded (decreasing the hit rate). This has led to a multitude of methods being developed to...
assess precipitation spatially. Ebert (2008) made the first attempt at categorizing these new methods into whether a single observation (SO) or neighborhood observation (NO) was used to compare to a neighborhood forecast (NF). The assumption is that neighboring forecast values are just (or nearly) as likely to provide the correct value of the single (central) forecast value. By considering a neighborhood (nb), a distribution of values is available instead of just one. When defining events, the fractional coverage of events within the neighborhood can then be treated probabilistically. Gilleland et al. (2009) provide an introduction to an intercomparison of spatial verification methods with a focus on gridded truth data. All of these methods were primarily developed to produce objective verification results that showed the benefits of convection-permitting model forecasts, because the traditional methods and metrics failed to match subjective assessment results.

More recently, near-convection-resolving model resolutions have been introduced, here loosely referring to resolutions ~2 km or less. The word near is used to prefix convection resolving because Bryan et al. (2003) suggest that resolutions of ~100 m or less are needed to resolve deep moist convection. Near-convection-resolving scales introduce even more detail, which has the potential to be wrong. Hohenegger and Schär (2007) also show that for 2.2-km forecasts the error growth rates are 10 times larger than for a 50-km forecast, further complicating the task of showing the benefits of near-convection-resolving forecasts at longer lead times. This also points toward a potentially even greater need for ensembles, conventional or poor man’s, at the near-convection-resolving scale (e.g., Johnson and Wang 2012), to quantify forecast uncertainty by capturing as many alternate outcomes as possible.

The MetUM provides a seamless nested forecasting system that can be run at multiple resolutions and time scales, from the kilometer scale to the climate scale. For further details see Davies et al. (2005), Lean et al. (2008), and Hewitt et al. (2011). At the Met Office, a 1.5-km (UKV) configuration has been running routinely for several years with objective verification back to April 2010, the aim being to replace the 4-km (UK4) configuration, introduced in 2007. Both configurations use 70 vertical levels with 20 levels below 850 hPa. Table 1 summarizes the parameterization differences between these two MetUM configurations. The two configurations only differ in their treatment of convection, where the UK4 uses a convection parameterization scheme to treat shallow convection only. Despite the apparent lack of differences, it has been impossible, using traditional verification metrics, to show a consistent signal of the UKV forecasts as being more skillful across the range of variables of interest.

While it is preferable to use gridded truth data for the verification of such high spatial resolution forecasts, these are rarely available for variables other than precipitation. For the most part, routine verification relies on the use of conventional synoptic observations, which are irregularly spaced and often sparse. Synoptic observations are point measurements, often representing snapshots in time. They are subject to representiveness errors in both space and time, particularly for highly variable quantities, for example, precipitation (Tustison et al. 2001). All verification results have an error component that is a function of the observations’ uncertainty, which is independent of the forecast. The World Meteorological Organization (2008) describes synoptic observations to be representative of areas up to 100 km but locally could be less than 10 km. This is a very “climatological view” of observations, which may be true on average but not true on the temporal and spatial scales relevant to NWP. The horizontal resolution and spatial variability in convection-permitting forecasts already suggest that now individual model grid spacings are typically smaller than the quoted representativeness of the observations! While the variability may not be accurate, it is real. As grid spacing shrinks, weather variables such as clouds and precipitation are tending to a binary response, in the limit, changing the meaning of the concept that a model grid value is an average of a subgrid-scale distribution. This may even be true for more smoothly varying variables such as temperature. Furthermore, observations may be instantaneous or time means. The same is true for forecasts. Precipitation is typically an accumulation, but other variables may be instantaneous time-step values or time means.

Precise matching of nearest or even bilinearly interpolated model grid points to synoptic (point) observation locations is exacerbated as model grid “points” represent progressively smaller areas, especially as the model still provides an area-average forecast, not a forecast for a point. Statistically, traditional point verification, which relies on precise matching of a forecast

<table>
<thead>
<tr>
<th>Scheme</th>
<th>UK4 (4 km)</th>
<th>UKV (1.5 km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convection</td>
<td>Gregory and Rowntree (1990)</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>area-scaled CAPE closure</td>
<td></td>
</tr>
<tr>
<td>Radiation</td>
<td>Edwards and Slingo (1996)</td>
<td></td>
</tr>
<tr>
<td>Boundary layer</td>
<td>Lock et al. (2000)</td>
<td></td>
</tr>
<tr>
<td>Cloud</td>
<td>Smith (1990)</td>
<td></td>
</tr>
<tr>
<td>Microphysics</td>
<td>Prognostic rain</td>
<td></td>
</tr>
</tbody>
</table>
grid point to an observation at a point is valid for kilometer-scale NWP, if this is the way the forecast is being used, but represents a very naïve (physically inappropriate) use of high-resolution NWP.

There are two issues with traditional verification applied to kilometer-scale NWP. First, many of the traditional metrics such as the root-mean-square-error (rmse)-and contingency-table-based scores have been much maligned, even blamed for the inability to show the benefit of convection-permitting models. There is some truth in this. Any rmse-based metrics mask issues such as the poorer representativeness of coarser-resolution models. To some extent, we have gotten away with this for variables such as mean sea level pressure or geopotential height, but is less than helpful for weather variables.

The other issue relates to the way in which the forecasts and the observations are paired, or matched. The extent to which the forecast and the observation are similar to each other in terms of their temporal and spatial characteristics is a part of this. These effects have been recognized for postprocessing of deterministic forecasts (e.g., Theis et al. 2005; Roberts (2003) used an NF approach to create probability of precipitation (PoP) forecasts, and compared these to gauge totals using standard probabilistic verification metrics such as the Brier score (Brier 1950). To test the benefit of using a neighborhood, they computed a Brier skill score (BSS) relative to what is termed direct model output (DMO), which is the nearest model grid point or bilinearly interpolated value.

The aim of this paper is to explore the use of the single-observation neighborhood-forecast (SO NF) concept as a framework for verifying variables from near-convection-permitting NWP models at standard surface observing sites. In addition to presenting a framework, this paper also aims to demonstrate the need for, and value of, using neighborhoods for creating forecast products. For the purposes of this study any reference to skill implies the skill of the forecast, and not just the model. As a weather service, we are interested in providing the most skillful forecasts to our customers and demonstrating or maximizing the skill of these forecasts as measured by a skill score, be it relative to another model or to a reference forecast is of primary interest. For the analysis the following questions are posed:

(i) What is the impact of using neighborhoods on practical (usable) skill compared to using a single grid point of the same model? Does this provide quantitative evidence of the representativeness mismatch and double-penalty effect, where detail is present but slightly displaced?

(ii) What does verification within neighborhoods suggest about the relative skill of near-convection-resolving and convection-permitting models?

(iii) Does the skill over persistence (24-h persisted observation) improve through the use of neighborhoods?

Section 2 provides a brief overview of the forecasts and observations used in this study, and how they are treated operationally. It also describes the new strategy that is employed, which includes the use of significance testing. In section 3, the results are discussed in the light of the questions posed above. The same framework is applied to two models, and the relative performance is considered. Finally, some conclusions are drawn in section 4.

2. Verification strategy

For the purposes of demonstrating the proposed framework, this study considers all 36-h UK4 and UKV forecasts initialized at 0300 UTC during December 2011 and focuses on comparing the current operational verification process to a new framework that is based on probabilistic metrics. December, being winter, should provide adequate samples of low visibility, low cloud bases, and precipitation.

For routine operational verification purposes, verifying surface observations are obtained from hourly surface synoptic observations (SYNOP) reports. Subject to quality control, ~130 World Meteorological Organization (WMO) synoptic observing stations around the United Kingdom are routinely used for forecast verification.

Verified variables included are 10-m wind, 2-m temperature, hourly precipitation (≥0.5, 1, and 4 mm h⁻¹), total cloud amount (TCA; ≥2.5, 4.5, and 6.5 oktas), cloud-base height (CBH; ≤100, 300, and 1000 m given 2.5 oktas), and visibility (≥200, 1000, and 4000 m). Cloud synoptic observations are reported as oktas, or eighths. CBH is measured using a low cloud-base recorder (LCBR), and binned in 30-m gates below 1500 m. Visibility is measured by visiometers that measure the transmittance of a sample volume of air, which is converted to meteorological optical range.

a. Precise matching and associated metrics

To assign the forecasts valid at these sites, a simple bilinear interpolation is used for continuous variables (temperature and wind), taking a distance-weighted average of the four surrounding model forecast grid-point values. For wind, the u and v components are interpolated separately. For the other variables (precipitation, TCA, CBH, and visibility) the nearest
model grid point is used. TCA, CBH, and visibility are all diagnostic variables, and are only as good as the method used to derive them from the prognostic model variables.

In the UK4 and UKV the cloud scheme described by Smith (1990) is used to diagnose a continuous model-level cloud fraction. Briefly, cloud fraction is diagnosed when supersaturation occurs at some point in the humidity distribution specified within the grid box. This scheme assumes a triangular humidity distribution over a grid box. Supersaturation is determined through the choice of a critical relative humidity (RHcrit). The cloud fraction is the integral of this distribution from saturation to the maximum value of relative humidity. The cloud fraction represents the proportion of the grid-box that contains cloud. TCA is derived using the maximum-random overlap assumption in the model column (Geleyn and Hollingsworth 1979; Räisänen 1998). The assumption states that vertically continuous cloud is arranged (in model levels) on top of each other (maximum overlap) and that clouds that are separated by at least one cloud-free level overlap in a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion. Cloud observations are in oktas (or eighths), so the diagnosed model TCA is binned to a random fashion.

CBH is reported as the height of the lowest model level with at least 2.5-oktas cloud fraction. A cloud fraction of ~0.3 is considered sufficient cloud cover to adequately diagnose cloud base for aviation purposes. The MetUM has 70 vertical levels with vertical resolution <50 m below 200 m (~975 hPa) and <100 m below 1.5-km height (~850 hPa). The visibility diagnostic is described by Clark et al. (2008). The diagnostic is a function of humidity and prognostic aerosol content, assuming a simple exponential scattering law and a visual range defined by a fixed liminal contrast. The aerosol concentrations are based on a range of emission sources. Aerosol can be advected, mixed by the boundary layer scheme, and washed out by rain. Visibility observations are also assimilated using a nonlinear function that has dependencies on temperature, humidity, and aerosol.

Forecasts at a site are only verified at times (hours) when quality-controlled observations are available. A model orography to station-height correction is applied to temperature observations using the average environmental lapse rate (ELR) of 6.5°Ckm⁻¹. The basic verification measure for temperature and wind forecasts is the monthly rmse (or root-mean-square-vector error, rmsve):

\[
\text{rmse} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (f_i - o_i)^2},
\]

where \(N\) is the number of matching pairs of forecasts and observations in the month, \(f_i\) is the forecast temperature, and \(o_i\) is the observed value. The root-mean-square-error skill score (RMSESS) or root-mean-square-vector error (RMSVESS) is computed relative to a reference forecast. In this case 24-h persisted observations are used as the reference forecast to avoid diurnal mismatches:

\[
\text{RMSESS} = 1 - \frac{\text{rmsve}}{\text{rmsref}},
\]

TABLE 2. A representation of a 2 \(\times\) 2 contingency table associated with an event threshold.

<table>
<thead>
<tr>
<th></th>
<th>Observed yes</th>
<th>Observed no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast yes</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Forecast no</td>
<td>(c)</td>
<td>(d)</td>
</tr>
</tbody>
</table>

The equitable threat score (ETS) has been the preferred score for categorical variables. It is but one of many metrics and scores that can be computed from 2 \(\times\) 2 contingency tables (see, e.g., Jolliffe and Stephenson 2003):

\[
\text{ETS} = \frac{a - a_r}{a - a_r + b + c},
\]

where
with forecaster perception, and potentially provide a quantitative estimate of representativeness effects. In the next section, the use of forecast neighborhoods is explored.

b. Probabilistic metrics applied to forecast neighborhoods with single observations

Atger (2001), Damrath (2004), and Theis et al. (2005) all used so-called SO-NF methods by computing the spatial fraction of values exceeding a threshold in a forecast neighborhood. This was for precipitation forecasts only, as per the classification done in Ebert (2008). The purpose here is to expand the use to the variables listed in the previous section: 2-m temperature, 10-m wind, TCA, CBH, and visibility.

When UKV DMO is verified at observing sites, the quality of the forecast is determined by using only ~130 model grid points from a 810 × 622 domain or 503 820 grid points. For this study three and two neighborhood sizes were applied to 1.5-km UKV and 4-km UK4 forecasts, respectively, to mimic the other operational MetUM resolutions: 4, 12, and 25 km. Figure 3 shows the dimensions of these neighborhood sizes applied to a UKV temperature forecast during April 2011. Even the 3 × 3 neighborhood corresponding to 4.5 km is barely visible. The temperature variations within the neighborhood become more apparent for the larger sizes. Note that in this study no attempt was made to mask out forecast values associated with large elevation differences or land–sea contrasts, as the idea is to create an “ensemble effect” and these variations add to the spread.

The variables that are currently verified using the ETS are well suited to being assessed using the Brier score (BS), which is “by far the most common” scalar accuracy measure for two-category (event–no event) probabilistic forecasts (Wilks 2006). By definition, the BS is analogous to the mean-squared error (MSE), but where the forecast is a probability and the observation is either a 0 or 1, corresponding to not observed or observed. The BS represents an error in probability space and is a measure of accuracy. It follows that the smaller the error, the smaller the BS, and the more accurate the forecast:

\[
BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2.
\] (5)

In Eq. (5), \(N\) represents the total number of forecast–observation pairs of a probabilistic forecast \(p_i\) and corresponding observation \(o_i\). The BS can be decomposed into three parts to provide separate information on the reliability, resolution, and uncertainty of the probabilistic
Fig. 1. Cumulative score differences between UK4 and UKV over the 0300 UTC 36-h forecasts from April 2010 to June 2012 (based on precise matching and the use of RMSESS and ETS): (from top to bottom) temperature, wind speed, TCA, CBH, visibility, and hourly precipitation. Included thresholds are hourly precipitation $\geq 0.5, 1, and 4 \text{ mm h}^{-1}$; TCA $\geq 2.5, 4.5, and 6.5$ oktas; and for CBH $\leq 100, 300, and 1000 \text{ m}$ given 2.5 oktas. The asterisk indicates scores for December 2011. Note these are UKV minus UK4 score differences.
The BS can also be calculated for a reference forecast, often the long-term climatological frequency of an event, which represents a constant reference forecast. Other reference forecasts can be used, for example, a sample climatology, or some other reference such as DMO. This enables the calculation of a Brier skill score like that described below:

$$BSS = 1 - \frac{BS}{BS_{ref}}, \quad (6)$$

While the BS is bounded by 0 (best) and 1 (worst), the BSS is positively oriented with an upper bound. A perfect forecast would have a BSS of 1.

Within the new framework the BS for a deterministic forecast at the grid scale (when a threshold is applied) simplifies to a binary error, as it can only be 0 or 1. It can only be 0 if both $p_i$ and $o_i$ are 0 or 1. An error results when $p_i$ and $o_i$ are not the same. Therefore, at the grid scale the new framework is the equivalent of assessing the proportion of correct forecasts (hits and correct non-events, $a + d$), but cannot discriminate between the misses and false alarms. If the forecasts are getting better, one would expect this proportion to get bigger, through moving misses into hits and/or false alarms into correct negatives. Some may consider it unsatisfactory in a pure deterministic sense, but the grid scale is not the focus of this strategy. It merely provides the baseline; therefore, legitimacy is sufficient.

For wind and temperature there is a clear desire to verify the distribution instead of focusing on a few thresholds, also in part due to the large seasonal variations for these variables. To match the concept of distributions the ranked probability score (RPS) is used for assessing the distribution of 10-m wind speed values in the neighborhood. Wind speed, like precipitation, is a skewed distribution, rather than Gaussian. For wind energy applications, the Weibull distribution is often used and has been shown to be applicable over land and sea (e.g., Hennessey 1977). To use the continuous ranked probability score (CRPS) with wind, the data would need to be transformed first to make the distribution more Gaussian. This makes assessment less intuitive. This distribution behavior and interpretation is more easily accommodated with the RPS through choosing a threshold progression such as the Beaufort force (BF) scale. This choice links wind speeds to categories that are important from a forecasters’ perspective. The boundaries in BF 0–9 used for this study are 2.06, 3.60, 5.65, 8.74, 11.31, 14.39, 17.48, 21.07, and 24.67 m s$^{-1}$.

**FIG. 2.** Results from the forecasters’ subjective assessments of UKV vs UK4 between May 2011 and April 2012. The thin bars show the actual number of times UKV or UK4 scored better. Thick bars show scores expressed as a percentage (J. Bornemann 2012, personal communication).
The RPS (see, e.g., Epstein 1969) is defined as
\[ \text{RPS} = \frac{1}{J} \sum_{m=1}^{J} \left[ \left( \sum_{k=1}^{m} p_k \right)^2 - \left( \sum_{k=1}^{m} o_k \right)^2 \right]. \] (7)

The RPS is a multicategory extension of the BS over \( J \) categories. It is a squared-error score with respect to the bin into which the observed event falls. This makes it sensitive to distance in terms of the number of categories separating the forecast and observed values. Murphy (1970) pointed out that it is particularly suited for ordered variables where being more than one category off would be significant from a forecaster’s perspective, and this should be reflected in the score. The key difference is that cumulative probabilities \( p_k \) and \( o_k \) (which make up \( P_m \) and \( O_m \)) are used to compute the score. The empirical distribution of the forecast values in the neighborhood is determined through the application of a range of thresholds that should encapsulate all forecast values.

Like the BS, the perfect score is 0 with the range being between 0 and 1. It measures the sum of the squared differences in cumulative probability space for a multicategory probability forecast and tends to penalize forecasts more severely when probabilities are further from the actual outcome. The RPS can also be normalized by dividing by \( J^2 - 1 \), although when a skill score is computed this normalization is irrelevant. When evaluating a deterministic forecast using the RPS, given that there is only one realization, only one of the \( k \) bins will be equal to 1, the rest will be 0. This means that \( P_m \) is also a Heaviside function. If the observation falls into the same \( k \)th bin, then the RPS = 0.

The CRPS is used for the 2-m temperatures in the forecast neighborhood. The CRPS (e.g., Matheson and Winkler 1976) is written as
\[ \text{CRPS} = \int_{-\infty}^{\infty} \left[ P_f(x) - P_o(x) \right]^2 \, dx \] (8)

and determines the differences between the forecast \( P_f \) and observed \( P_o \) (Heaviside) cumulative probability distributions. A perfect forecast has a CRPS of 0. For a deterministic forecast the CRPS reduces to the mean absolute error (MAE). The CRPS can also be decomposed into reliability, resolution, and uncertainty terms (Hersbach 2000).

Skill scores can be calculated for both the RPS and CRPS by defining a reference forecast, using the general skill score formula. To maintain the link to the underlying deterministic roots, one choice of reference is, just like for the BSS, the DMO score:
\[ \text{RPSS} = 1 - \frac{\text{RPS}}{\text{RPS}_{\text{ref}}} \] and \[ \text{CRPSS} = 1 - \frac{\text{CRPS}}{\text{CRPS}_{\text{ref}}}. \] (9) (10)

c. The verification framework

The proposed verification framework consists of a three-way comparison. The first comparison relates to
how the use of a neighborhood affects forecast skill relative to using DMO. The benefit or deficit relative to using DMO may provide evidence of the double-penalty effect where forecasts are essentially correct but not quite in the right place. This facilitates determining the intrinsic skillful scale of a model for a given variable. Next, for model development the interest is often whether model A is better than model B in a test and control sense. The test may be the same model configuration as the control, but with a physics change. On the other hand, it may be testing whether a model at one (higher) horizontal resolution is more skillful than a model with a different (coarser) horizontal resolution. In this case, the relative skillfulness is of interest. This includes the comparison of forecast neighborhoods of comparable size. In considering both DMO and neighborhoods, can it be shown whether the double-penalty effect is instrumental in not being able to show objectively that the DMO of a higher-resolution model is more skillful? Last, the skill of a DMO or a forecast neighborhood is calculated relative to a potentially skillful persistence forecast. Here, a 24-h persisted observation is used. Using persistence retains the concept of memory as explored by Mittermaier (2008). The skill scores measure the accuracy of the forecast compared to the observed events.

d. Testing for significant differences

Using hypothesis testing for comparing forecast systems or testing model upgrades to the same forecast system is still relatively rare. Hamill (1999) suggested a method for precipitation forecasts that focused on some of the key elements: spatial correlation, pairing of samples, and serial correlation. More recently, Mittermaier et al. (2013) also explored these ideas further using a Student's t test for dependent pairs to consider a time series of 5000 forecast-by-forecast scores. This required calculating an effective sample size to take into account serial dependence and provide the correct (reduced) degrees of freedom.

In recognition that an assumption of normality may be false for a smaller sample, the nonparametric Wilcoxon signed rank test for paired dependent samples was used here. Aggregated hourly forecast neighborhood-to-single-gridpoint score differences (BS, RPS, and CRPSS) for each of the forecast lead times are used. Significant differences are computed as “test – control.” If the test has more accurate forecasts, then the differences will be negative in this case. Two-tailed (inequality) confidence intervals were computed at the 5% level and are used to test the null hypothesis. If the interval contains zero, the differences are not significant. Significant differences are highlighted on the graphs in section 3 with the use of the letter “s.”

3. Results

Despite the use of only 1 month’s worth of forecasts, the study is considering \((130 \times 30)\) days = \(~4000\) samples per variable, although these are not independent. Note therefore that the results presented here do not attempt to provide conclusive proof that the 1.5-km UKV is more skillful than the 4-km UK4, but they may provide useful clues. Subsequent sections that dwell on interpretation merely attempt to illustrate how the framework can be used. For some event thresholds, there are probably too few cases in the sample date for any conclusions to be drawn. Given a long enough time series, it is hoped that this strategy can provide this conclusive proof, on monthly, seasonal, and annual time scales.

For this study the control forecasts are either: (i) UKV or UK4 DMO 0300 UTC forecasts, (ii) 4-km UK4 0300 UTC forecast neighborhoods, or (iii) persistence taken to be the 24-h persisted synoptic observation. The CRPSS (temperature), RPSS (wind speed), and BSS (categorical variables) are computed relative to each of these control forecast options.

a. Skill of forecast neighborhood over a single grid point of the same model

Before comparing two models with different horizontal resolutions, the impact of using neighborhoods with respect to the native resolution is quantified. The skill scores of the forecast neighborhoods relative to DMO are shown in Fig. 4. Each of the six variables is discussed in turn. For temperature, there is a clear benefit through the use of a forecast distribution. It is important to stress here that this approach is not a form of smoothing. This methodology provides better sampling through the use of a distribution of values rather than a single value. It is also clear that for temperature there appears to be an optimal neighborhood size of absolute benefit, beyond which a bigger neighborhood yields less of a benefit. Overall, the biggest benefit is in the use of a neighborhood, with less benefit between the different neighborhoods shown here.

For 10-m wind speed, Fig. 4 shows the biggest increase in skill scores is again through the use of a neighborhood over just using DMO. The benefit continues to increase for all three neighborhood sizes considered. This can again be attributed to the use of a distribution and a categorical approach, which is less sensitive to grid-scale noise. There is previous evidence of improvements in skill scores (using the RMSVESS) of wind forecasts.
FIG. 4. Cumulative UKV neighborhood skill score benefits–deficits for December 2011 over 36 h computed relative to the single nearest (or bilinearly interpolated) grid point (DMO) from UKV for each of the six in Fig. 1. A positive value implies the neighborhood adds benefit over DMO. The letter “s” denotes the statistical significance of the neighborhood-to-single-gridpoint score difference at the 5% level. See section 2d.
through upscaling or smoothing, as shown in Fig. 5. Figure 5 shows results from a 13-month upscaling trial that suggest that upscaling the UKV to 4.5 km (or average of 3 × 3 grid squares around observing site) does not make the UKV better than the UK4 based on the RMSVESS. On the other hand, comparing an upscaled UKV against UKV DMO does suggest an improvement in skill, so there is some benefit to be gained by upscaling (smoothing). Visual inspection of UKV wind speed forecasts suggests there is too much grid-scale variability, in terms of wind direction (probably due to the higher-resolution model orography), which acts like noise, and can be particularly detrimental to rmse-based metrics.

Cloud (TCA) skill scores in Fig. 4 also show a clear benefit from deriving a spatial cloud cover fraction to counteract grid-scale problems related to model bias. The benefit continues to increase almost linearly with the neighborhood sizes shown, with no sign of leveling off. This would suggest these forecasts are potentially inherently poor. This is further substantiated by the comparison to the UK4 and persistence, as shown below (Figs. 6 and 8).

In Fig. 4 the use of neighborhoods improves the UKV CBH scores substantially compared to DMO, especially for the low cloud bases, which occur less frequently. CBH is a diagnosed quantity, preconditioned on a minimum TCA of at least 2.5 oktas. Therefore, TCA skill, or lack thereof, also has an impact. As a result, CBH is inherently noisy and possibly has even less spatial coherency than precipitation. It is worth noting that cloud variables do have skill; see, for example, Illingworth et al. (2007, and references therein) and Hogan et al. (2009) for a European model intercomparison perspective.

Mittermaier (2012) compares three different operational model configurations of the MetUM. The BS behavior for rarer events also needs some explaining. Typically as the frequency of observed occurrence decreases, the uncertainty term of the BS also decreases. For rare events the probabilities will tend to be small and the squared difference tends to 0. So instead of the probability at an observing site being 0, a neighborhood conceivably introduces a non-0 probability, which when an event is observed makes the scores better. However, this can also be counterproductive. If a model has a strong overforecasting bias for rarer events, introducing a non-0 neighborhood probability will make the BS worse if the event was not observed.

For visibility, Fig. 4 shows there is a clear benefit in the use of a neighborhood for all thresholds considered. Again the biggest increase in skill scores is through the use of a neighborhood over simply just DMO.

Hourly precipitation is spatially discontinuous, and given Fig. 4, responds positively to the use of a neighborhood. In addition the 4 mm h⁻¹ threshold is consistently the 95th percentile or higher over the United Kingdom, so that is rather extreme with few events (<200 in 4000), even in winter. Therefore, a ~25-km neighborhood may yet be too small. Nevertheless, using a neighborhood shows a large benefit for the 4 mm h⁻¹ threshold, with statistically significant increases for most thresholds.

Making this type of comparison provides a measure of the impact of the double-penalty effect, potentially
FIG. 6. As in Fig. 4, but for UKV DMO and neighborhood skill score computed relative to UK4 DMO and neighborhoods. A positive value implies the neighborhood shows a benefit over DMO or a neighborhood.
providing the means of identifying forecast variables that would be more skillful at an observing site when a forecast neighborhood is used instead of DMO to provide a site-specific forecast. This then links the process of verification to postprocessing, and is best illustrated by precipitation and TCA, which show a uniform upward step change in skill with increasing neighborhood size for all thresholds. Undoubtedly, bias effects are also embedded in these results, although experience with other neighborhood metrics such as the fractions skill score (FSS; Roberts and Lean 2008) show that bias is most clearly manifested through low scores over large neighborhoods approaching the domain size.

The temperature (and potentially wind) benefit in Fig. 4 seems to reflect the trade-off between small-scale variability and topographical constraint. With increasing neighborhood sizes potentially more topographically unrepresentative model grid points are included (e.g., sea instead of land, valley instead of hill). Including these variations acts to increase the spread of the values considered to be associated with a particular observing site, and from a probabilistic viewpoint this may be a good thing. On the other hand, these values may diminish the benefit of the neighborhood. Therefore, masking potentially unrepresentative grid points (in terms of a given observing site’s characteristics) from a neighborhood could be essential for considering very large neighborhoods.

The relative benefit of using the smallest neighborhood over DMO is ~15% for variables such as wind and temperature. For the other variables the benefits are less consistent, with increases on the order of 5%–35%.

b. Skill of one model relative to another

This comparison attempts to assess model configurations in a relative sense and is of great interest to operational NWP communities intent on testing whether model A is better than model B. Typically, model B would be an upgrade package to model A. The expectation would be for the benefits–deficits to be much smaller, especially if the model configurations are similar, when it becomes more difficult to detect a signal from the noise. One future test of this strategy is to determine whether this framework can provide better guidance.

In Fig. 6 four comparisons are made: 1) UKV DMO to UK4 DMO (mimicking what is done routinely using the metrics described in section 2a), 2) a 3 × 3 UKV (~4.5 km) neighborhood compared to UK4 DMO, and 3) and 4) a comparison of two equivalent UKV and UK4 neighborhoods (~12 and ~25 km).

From Fig. 6, the UKV DMO temperature is more skillful than the UK4, as was shown using the RMSESS in Fig. 1. There is a relatively large positive benefit for the UKV when even the smallest neighborhood is used when compared to UK4 DMO. The larger UKV neighborhoods still show more benefit over the equivalent UK4 neighborhoods but the margin is reduced.

An interesting result is that for 10-m wind the UKV DMO is more skillful than the UK4 DMO when RPS is used, as seen in Fig. 6. This is in contrast to the skill deficit found in assessing the vector wind using an RMSVESS, as seen in Fig. 1. This result suggests that the UKV DMO places the wind speed in the correct category more frequently than the UK4. Yet again, there is a large positive benefit when a neighborhood is introduced. When comparing neighborhoods of equivalent size, Fig. 6 also suggests that the UKV still enjoys a benefit over the UK4 but with a reduced margin.

For TCA, Fig. 6 shows that UKV DMO has an overall deficit compared to UK4 DMO. Using a 4.5-km neighborhood, the UKV deficit is turned into a benefit for all cloud categories, but for larger neighborhoods the UKV deficit returns against equivalent UK4 neighborhoods. This suggests that the UKV deficit in relative skill is due to more than just the double-penalty effect, and this remains an active area of model development.

Figure 6 shows that UKV CBH DMO scores are only better than the UK4 DMO for the lowest cloud bases. Comparing the benefit over similar-sized neighborhoods, Fig. 6 shows a similar signal although the benefit for the lowest cloud bases diminishes with increasing neighborhood size, and the deficit for the other categories increases.

The BSS for visibility shows that just like in Fig. 1 UKV DMO is relatively more skillful than the UK4 DMO. Further relative gains are possible through the use of a forecast neighborhood. When comparing equivalent neighborhoods, Fig. 6 suggests that increasing the neighborhood size yields diminishing benefits.

The aggregated ETS difference for hourly precipitation in December 2011 was one of the few months in the time series when it was positive for the UKV DMO (see Fig. 1, which shows the aggregated difference over all three thresholds as previously defined). In Fig. 6, two of the thresholds are showing a positive benefit over UK4 DMO with the 4 mm h⁻¹ threshold statistically significant when the BSS is used instead of the ETS. A small (~4.5 km) UKV neighborhood yields substantial benefit over UK4 DMO, including the overturning of the 0.5 mm h⁻¹ deficit. When using equivalent UK4 neighborhoods, Fig. 6 would seem to indicate that the lowest threshold varies between being a statistically significant benefit or deficit, while the UKV benefit for the largest threshold remains clear. This at least demonstrates the benefit of higher resolution for higher thresholds, but also suggests that there are possible
differences in the spatial bias of low rainfall amounts between the models.

So far only the cumulative benefit–deficit over the entire forecast length has been shown, and this is a key summary tool. However, the hour-by-hour comparison of performance for a given neighborhood size can be equally important, if specific forecast issues are of interest (e.g., the impact of data assimilation in the 0–6-h window). Figure 7 shows the hourly benefits–deficits of the ~12-km UKV to the 12-km UK4 neighborhoods as a function of lead time. Note that the cumulative benefit–deficit from Fig. 7 is summarized by the third bar in Fig. 6. Figure 7 shows the UKV benefit for temperature and wind remaining relatively uniform with lead time. Figure 7 also indicates that the rate of change in skill between the UKV and UK4 remains fairly constant. There is a slight decrease in the UKV skill at the longest lead times, suggesting that the UKV skill scores drop off more quickly relative to the UK4 scores. For TCA, the UKV has a deficit at almost all lead times with a downward trend with lead time. The lowest cloud bases show considerable benefit for the UKV up to t + 25 h. Visibility shows a UKV benefit for almost all lead times. For hourly precipitation the 4-mm threshold shows the most benefit at all lead times. Overall, the behavior is less consistent for the categorical variables.

Thus, it can be concluded that neighborhoods have an impact when comparing models of different resolutions, and it is insightful to compare to a single grid point and a neighborhood. The skill of an inherently good forecast (but for the double-penalty effect) can be further enhanced through the use of a neighborhood. On the other hand, using a neighborhood will still expose inherently poor forecasts (e.g., TCA in this study), which cannot necessarily be made to look better than another model, using a SO-NF methodology.

The differences in the cloud variables between the UK4 and UKV methods are particularly interesting, as cloud has been the focus of several studies recently. Crocker and Mittermaier (2013) investigated the spatial cloud cover biases of these two models using a satellite cloud mask to determine that the UKV is more cloudy than the UK4. Synoptic observations of cloud are also flawed, as highlighted by Mittermaier (2012). As both models are compared against the same observations, the differences seen here are not due to observations but are model configuration related, even though they are very similar. This may include the initial conditions and the way in which the diagnostics are formulated. The UK4 approach uses a convection scheme that can produce a small amount of precipitation and can also produce convective cloud. The UKV must produce this convective precipitation and cloud dynamically. At each model level the total cloud amount is calculated, which includes any contributions from the convective parameterization. TCA and CBH are diagnosed from the total cloud on each model level. As the grid spacing decreases, the distribution of TCA in model configurations is changing. One hypothesis is that the classic view of a model grid-box value as a representation of the average of a distribution of subgrid-scale values is being sharpened to the extent that, for TCA at 1.5-km resolution, the outcome approaches a binary response. This leads to potentially substantial representativeness errors against synoptic observations, which are either hemispheric (with typical visible horizons of 40 km or more) or time means, representing areas much larger than a 1.5-km grid box. It is also worth noting here that MetUM TCA and CBH are instantaneous time-step values. When using DMO from high-resolution models, these large representativeness errors are further exacerbated by the double-penalty effect with small-scale detail not being in the right place.

### c. Skill relative to persistence

Finally, results for the most conventional comparison, actual forecast skill against a reference forecast, are shown in Fig. 8. A persisted observed state can often be a difficult reference forecast to beat, given the memory that is inherently present in weather patterns (Mittermaier 2008).

For UKV temperature Fig. 8 shows there is a marginal improvement over persistence when a neighborhood is used. The change in score as a function of neighborhood size is smaller than the difference in DMO to the smallest neighborhood. Somewhere between 4.5 and 25.5 km, an optimal neighborhood exists where the skill score is maximized. UKV scores relative to persistence are marginally higher, with the use of the smallest neighborhood introducing a 7% increase.

Compared to persistence, the use of a 4.5-km neighborhood increases the UKV 10-m wind speed skill score by −0.1 (or ~20%) over DMO (Fig. 8). Subsequent increases are smaller. Contrary to the RMSESS (Fig. 1), the UKV and UK4 wind scores (not shown) are the same without neighborhoods, but higher for the UKV when neighborhoods are used.

Figure 8 shows that for TCA there is also an improvement in skill scores when neighborhoods are used, and this improvement increases with increasing neighborhood size. This reflects the lack of skill at the grid scale, which is probably related to displacement errors. It may also be due to the impact of spatial biases, where considering a spatial fraction counteracts an overforecast bias, especially for more overcast conditions. The biggest improvement is for the mostly cloudy category of ≥6.5 oktas, where using a 25.5-km neighborhood yields
Fig. 7. Hour-by-hour benefits–deficits as a function of forecast lead time of ~12-km UKV neighborhoods over equivalent 12-km UK4 neighborhoods for December 2011 for each of the six variables in Fig. 1. A positive value implies the UKV neighborhood shows a benefit over the UK4 neighborhood. As before, the s denotes the statistical significance of the neighborhood-to-single-gridpoint score differences at the 5% level.
FIG. 8. As in Fig. 6, but for values computed relative to persistence (24-h persisted observation). A positive value indicates the forecast is better than persistence.
a 0.1 increase in BSS. Given the relative skill between the UK4 and UKV methods, it suggests that while the UK4 and UKV approaches both benefit from this framework, UK4 appears to benefit fractionally more (not shown). Neighborhoods yield between 8% and 12% improvement in scores, over DMO against persistence.

The UKV CBH scores against persistence in Fig. 8 for the lowest cloud bases are more than doubled in magnitude through the use of neighborhoods, while the deficits for the others are reduced. Given some of the deficiencies in UKV TCA, it is hypothesized that there are other contributors to poor skill that are as yet not fully understood and are the subject of further investigation. Overall, Fig. 1 suggested a fairly well-mixed month-to-month signal for CBH based on the aggregated ETS differences. Using the BSS still supports the UK4 forecasts as being slightly better overall (not shown), although the UKV forecasts of the lowest cloud bases remain superior to those from UK4.

The aggregated ETS difference for visibility in Fig. 1 during December 2011 showed the UKV DMO to be better than the UK4 DMO. Using the BSS, Fig. 8 shows the UKV DMO is only skillful against persistence for the largest, least severe threshold. Using a neighborhood introduces some a benefit for the UKV 1000-m threshold. This suggests an even larger neighborhood is required. A benefit of ~25% is achievable over persistence through the use of neighborhoods.

At this point, even though the method (rather than the results) is the main objective of this study, it is important to put some of the poor CBH and visibility scores against persistence into context. The expectation is that an NWP forecast should beat persistence. However, in the United Kingdom at least, especially in the winter, persistent low-cloud and reduced-visibility (fog) events are not uncommon. Under these circumstances persistence is often a very skillful forecast, and even if the NWP forecast is as good, the rewards in terms of attributable skill are small. If the model does not predict the onset, duration, intensity, or demise of these events as well, it can lead to poor scores compared to persistence. This is why persistence is considered the toughest reference forecast of all. On the other hand, if forecasts are better than persistence, it is a strong endorsement of forecast skill.

During December 2011, Fig. 1 suggested the UKV DMO hourly precipitation was more skillful than UK4 DMO based on the aggregated ETS (over the three thresholds). Figure 8 shows the UKV BSS results for the 4-mm threshold are worse than persistence for DMO and the 4.5-km neighborhoods. Only the DMO deficit is statistically significant. Larger neighborhoods reduce, and then overturn, the deficit. These results suggest that a neighborhood larger than 25.5 km is required to enhance the benefit further. This would certainly be true when considering more convective times of year such as the spring and summer. Mittermaier et al. (2013) already demonstrated the skillful spatial scale for the UK4 is, on average, of the order of 60 km when considering the upper 10% percentile threshold for 6-h precipitation totals against radar accumulations. The expectation is that for shorter (hourly) accumulations, this length scale would increase (get worse) because any timing or location errors would become more pronounced. Within this framework, using rain gauges instead of radar and the use of neighborhoods yields increases of 30%–40% in scores relative to persistence.

4. Conclusions

In this study a strategy has been presented that enables testing of near-convection-resolving model forecast skill relative to neighborhood size, forecasts from another (convection permitting) model, and a reference forecast. Incorporated in this process is statistical significance testing of the score differences to add credibility to the results and assist in decision making. The benefits of this strategy include the following:

- providing a different perspective of deterministic forecast skill through using probabilistic metrics, and treating the near-convection-resolving deterministic model forecasts probabilistically through the use of spatial fractions (probabilities); the basic probabilistic metrics can also be decomposed to provide further diagnostic scores if desired;
- providing a readiness for comparison to convective-scale ensembles, potentially providing useful guidance on the pre- and postprocessing of individual ensemble members to spatial probabilities to maximize ensemble skill (e.g., Ben Bouallégue et al. 2011); and
- counteracting the spatial and temporal representativeness errors of both the single-site observations and forecasts on verification metrics.

In section 1, several questions were posed that the study aimed to answer. From the results, we may conclude the following:

(i) Any neighborhood has, for the most part, a positive impact. The general behavior of the scores as a function of neighborhood size can provide clues as to whether the lack of skill seen at the grid scale is to do with the double-penalty or representativeness mismatches, or whether it is simply due to a model’s inability to forecast certain event thresholds (e.g., low cloud bases, visibility, cloud amounts). Using neighborhoods is generally beneficial for more
extreme thresholds. Using a neighborhood can mitigate against the double-penalty effect and representativeness mismatches, by introducing or enhancing forecast skill, but a basic level of model skill must be present to do so. A poor, relatively unskillful forecast cannot be made to appear more skillful very easily. The benefit is variable dependent; for example, for this forecast sample temperature and wind results suggest ~15% can be gained (for this forecast sample) over DMO, and can be ascribed to these effects. For the noncontinuous variables such as TCA, 30%–50% or more is possible.

(ii) While the strategy will have varied applications, the limited results presented here suggest that the framework should provide the means to show that a near-convection-resolving model with a neighborhood is more skillful than a convection-permitting model single grid point or neighborhood. Initial results suggest that this is variable dependent. Failure to show skill relative to another (coarser) model using a neighborhood could suggest a more fundamental model deficiency. UKV TCA and CBH forecasts shown here seem inherently less skillful, and the difference is not simply due to the double-penalty effect.

(iii) Forecasts ought to be more skillful than persistence. New model systems or model upgrades should also be more skillful than their predecessors; otherwise, why introduce them? Here, for this small sample of forecasts, it has been shown that for the majority of variables UKV and UK4 (not shown) neighborhoods are significantly more skillful than persistence and also significantly better than UKV DMO and the UK4 DMO or neighborhoods.

Traditionally, the focus of studies has been on the improvements over a reference forecast and other model configurations. Here, an additional element, the forecast neighborhood, is introduced to measure the intrinsic skill in a model configuration. This study attempts to illustrate a strategy based on a small sample of forecasts. While many of the signals, such as the impact of neighborhoods, should hold true for other models, seasons, and scenarios, results shown here should not be generalized too far, and were not the main focus of the study in an absolute sense. Even for the MetUM, it would be unwise to generalize results to other months for the same model until a similar pattern was established through further analysis of much larger datasets. That is the subject of ongoing work.

There are some notable similarities between the traditional metrics plotted in Fig. 1 and the new framework applied to DMO. This should be reassuring. The results do change with the introduction of neighborhoods. Are the results provided by the new framework in closer agreement with forecaster assessment? It is difficult to draw conclusions here as the forecaster assessment is for a much longer period, not just for December 2011. There is the strong endorsement of convective precipitation, which is mirrored in the UKV 4-mm precipitation benefit. There is also a positive signal from the low cloud bases and visibility. Since the survey ended, the UKV has been updated further to improve the cloud overforecast bias and, hence, improve the forecast temperatures. Other aspects such as the wind gust parameterization are being redeveloped specifically for convection-resolving NWP.

In conclusion, two other points are worth making. The first is the difference between the concepts of native grid length or scale and the model resolution. Often these terms are used interchangeably, but they are not the same thing. Fundamentally, when comparing gridded forecasts, the native grid resolution should never be used for generating forecast products or verification. The native grid spacing determines the size of the feature that can be resolved. This aspect of the numerics applies to all models. For kilometer-scale NWP this will be at least 5 times the native grid (Cullen and Brown 2009). This tends to be a fairly unpopular argument from the user perspective, as it is difficult to explain why after all the computational expense of producing, say, a 4-km forecast, these should be presented to the end user as gridded products at no less than 25-km-averaging scale. Neighborhood verification methods have gone a long way in substantiating these fundamental dynamical and numerical constraints for kilometer-scale NWP. The issue is somewhat different for creating site-specific forecasts or verifying forecasts at a point. While a large grid box is an average of the subgrid scale, here it has been argued that to make optimal use of kilometer-scale NWP at the grid scale, the only alternative to counter the averaging length is to consider a forecast quasi-probabilistically.

The second and final comment relates to the issue of evaluating the bias in this new strategy. If assessing the kilometer-scale NWP skill using DMO is flawed, then so is the assessment of the bias using DMO. The concept of reliability is the probabilistic analog but the application of this concept to spatial probabilities for assessing model bias remains an area of ongoing work.

Acknowledgments. I would like to thank Nigel Roberts in particular for the constructive interaction on this topic.
REFERENCES


