Ensemble versus Deterministic Performance at the Kilometer Scale

M. P. MITTERMAIER AND G. CSIMA

Numerical Modelling, Weather Science, Met Office, Exeter, United Kingdom

(Manuscript received 7 September 2016, in final form 18 April 2017)

ABSTRACT

What is the benefit of a near-convection-resolving ensemble over a near-convection-resolving deterministic forecast? In this paper, a way in which ensemble and deterministic numerical weather prediction (NWP) systems can be compared is demonstrated using a probabilistic verification framework. Three years’ worth of raw forecasts from the Met Office Unified Model (UM) 12-member 2.2-km Met Office Global and Regional Ensemble Prediction System (MOGREPS-UK) ensemble and 1.5-km Met Office U.K. variable resolution (UKV) deterministic configuration were compared, utilizing a range of forecast neighborhood sizes centered on surface synoptic observing site locations. Six surface variables were evaluated: temperature, 10-m wind speed, visibility, cloud-base height, total cloud amount, and hourly precipitation. Deterministic forecasts benefit more from the application of neighborhoods, though ensemble forecast skill can also be improved. This confirms that while neighborhoods can enhance skill by sampling more of the forecast, a single deterministic model state in time cannot provide the variability, especially at the kilometer scale, where rapid error growth acts to limit local predictability. Ensembles are able to account for the uncertainty at larger, synoptic scales. The results also show that the rate of decrease in skill with lead time is greater for the deterministic UKV. MOGREPS-UK retains higher skill for longer. The concept of a skill differential is introduced to find the smallest neighborhood size at which the deterministic and ensemble scores are comparable. This was found to be $3 \times 3$ (6.6 km) for MOGREPS-UK and $11 \times 11$ (16.5 km) for UKV. Comparable scores are between 2% and 40% higher for MOGREPS-UK, depending on the variable. Naively, this would also suggest that an extra 10 km in spatial accuracy is gained by using a kilometer-scale ensemble.

1. Introduction

Kilometer-scale numerical weather prediction (NWP) forecasts are often praised for the detail they provide and the realism of the features they produce (e.g., Kain et al. 2006). The reasons as to why the detail in these kilometer-scale NWP forecasts may be misleading (and often not skillful) are also fairly well understood; for example, the aliasing of subgrid-scale features onto the grid means that objectively speaking there is often no material increase in forecast accuracy due to higher horizontal resolution (e.g., Mass et al. 2002; Roberts and Lean 2008; Roberts 2008; Mittermaier et al. 2013). This has led to the development of a plethora of new spatial verification methods, many of which formed part of a spatial methods intercomparison project, summarized in overview publications such as Gilleland et al. (2009) and Gilleland et al. (2010). Another reason is the loss of predictability. Rapid error growth at, or near, the grid scale has been seen as one of the main causes for forecast skill with lead time decreasing more quickly at high resolution. Lorenz (1969) showed that there is a limit to predictability due to the evolution of small-scale circulation features. However, more recently Durran and Gingrich (2014), and Nielsen and Schumacher (2016) have demonstrated that errors originating at the large scale tend to dominate forecast evolution, and that the large-scale errors are not merely a result of errors sourced at the grid scale, and cascaded from short- to longer wavelengths. Lorenz showed this too, though this aspect has generally been less publicized. Irrespective of the error sources, the loss of predictability and the impact on forecast skill remain among the main reasons why an ensemble approach is necessary at the kilometer scale, with a new verification approach to match this modeling requirement. For many national weather services, the value added by running a kilometer-scale ensemble...
over a deterministic kilometer-scale forecast must be quantifiable. One major reason is the need to justify the huge expense of running such ensemble systems.

Traditionally speaking, any comparison between an ensemble and a deterministic forecast system has been flawed. Mittermaier (2014) proposed a strategy that would enable such a comparison, by considering both in probabilistic and distributional ways. This strategy, which is now referred to as the High Resolution Assessment Framework (HiRA), introduces the concept that a probabilistic ethos is essential for kilometer-scale NWP. HiRA is based on the use of a square forecast neighborhood centered on observing sites for the purpose of assessing raw forecast skill. The assumption is that all forecast grid points in the neighborhood are equiprobable outcomes at the observing site. This is deemed necessary to counteract the growth of grid-scale error, which affects predictability. The so-called double penalty effect is a manifestation of this where the grid-scale error is a displacement error (which is often, but not always, a manifestation of a timing error). This displacement error means that an event that occurred at a given location, and was missed by the forecast at that location, occurred elsewhere where it was not observed to occur (the false alarm). Therefore, within a spatial (or temporal) context, the error is counted twice, as both a miss and a false alarm. The generation of a pseudoensemble from the grid points in a neighborhood surrounding an observing site is therefore able to introduce a form of spread that assists in providing a better forecast for a given location [and is a clear illustration of the commonalities between verification and postprocessing, e.g., Bouallégue and Theis (2014) and Dey et al. (2014)]. Postprocessing as a concept is multifaceted, and opinions on the subject, and what counts as postprocessing and what does not, can vary substantially. In this paper, the raw forecasts are not adjusted or corrected in any way. HiRA merely samples more of the forecast in the spatial dimension, which is something one may choose to do for the postprocessing of site-specific forecasts. Evaluation frameworks such as HiRA signpost the way in which raw forecasts should be utilized, to make the most of what they can provide.

In this paper, the strategy and framework proposed by Mittermaier (2014) is applied to just over three years’ worth of operational kilometer-scale ensemble and deterministic forecasts. The deterministic and ensemble configurations of the Met Office Unified Model (UM) are described in section 2. Section 3 describes the methodology, which includes a short review of HiRA, and how the method enables the comparison of an ensemble with deterministic forecasts. It also introduces the concept of a skill differential between successive neighborhood sizes, to identify comparable neighborhood sizes for the ensemble and deterministic forecasts. These comparable neighborhood sizes are compared to the contrasting concepts of “equalizing” on grid square size (which is preferred for comparing deterministic forecasts) or the number of ensemble members. Section 4 presents a selection of results from the modelers’ and the users’ point of view, demonstrating that the framework can accommodate both. Conclusions and further work are provided in section 5.

2. Model configurations

Ahead of the 2012 London Olympics, the Met Office introduced a 2.2-km 12-member Met Office Global and Regional Ensemble Prediction System (MOGREPS-UK) downscale ensemble across the United Kingdom (Tennant 2015; Lewis et al. 2015), primarily to showcase Met Office forecasting capabilities (Golding et al. 2014). The ensemble is based on the UM (Davies et al. 2005). The UM is a nonhydrostatic, semi-implicit, semi-Lagrangian, gridpoint model with 70 variably spaced vertical levels, which are terrain following near the surface. Parameterization schemes for short- and longwave radiation, as well as mixed-phase cloud, are included. There is also a one-dimensional boundary layer scheme and a nine-tile surface exchange scheme. Limited area configurations run on a rotated latitude–longitude projection to ensure that the grid shape is approximately square.

The ensemble complements the 1.5-km Met Office U.K. variable resolution (UKV) deterministic configuration of the UM, for short-range forecasting (Tang et al. 2013). This model has a variable horizontal resolution, with the inner domain a regular 0.0135° or 1.5-km grid. Only the inner domain is used for forecasting and verification purposes, and the domain is virtually identical for both configurations. Both of these configurations use the same underlying model physics, with small differences in time steps; UKV runs with a 50-s time step, whereas MOGREPS-UK runs with a 75-s time step. The UKV is one-way nested in the 17-km global model, and benefits from 3-hourly three-dimensional variational data assimilation (3DVAR) with additional latent heat nudging of radar rainfall rates and an incremental analysis update (IAU). MOGREPS-UK is dynamically downscaling the global MOGREPS-G (33 km) ensemble members. In February 2015, the dynamical core of all limited area models was upgraded to the “Even Newer Dynamics” (Wood et al. 2013). Both these configurations produce 36-h forecasts four times a day: 0300, 0900, 1500, and 2100 UTC.

At the Met Office, verification of U.K. models has always focused on six primary surface variables: 2-m
temperature, 10-m winds, total cloud amount, cloud-base height, visibility, and precipitation. As the resolution of the models has improved, the focus for precipitation shifted from 6-hourly to hourly accumulations.

3. The assessment strategy

This assessment strategy is based on understanding the local accuracy of forecasts at observing sites. As such, it updates the traditional method of verifying the nearest model grid point to an observing site, making it relevant and appropriate for kilometer-scale modeling.

In the United Kingdom, there are currently 166 official World Meteorological Organization (WMO) land surface synoptic observation (SYNOP) (LNDSYN) observing stations that send standard SYNOP messages via the WMO Global Telecommunication System (GTS). The WMO block identifier for U.K. and Republic of Ireland sites is 03. The locations of U.K. block 03 sites (excluding the Republic of Ireland) within the UKV and MOGREPS-UK domains are shown in Fig. 1. It is worth noting that HiRA could be calculated for marine sites such as buoys and platforms, but this has not been done so far.

The HiRA framework

Table 1 summarizes the scores that are used for each of the six forecast variables that are evaluated. For a comprehensive summary of verification scores, the reader is referred to Wilks (2006) or Jolliffe and Stephenson (2012). Table 1 describes the change in scores from the traditional deterministic method of evaluation (by extracting only the nearest model grid point and pairing it with the observation) to what is used in HiRA, where a neighborhood of model grid points centered on the observing location is used. The framework can still be used to evaluate “in the traditional sense,” when only the model grid point nearest to the observing location is used (though this really is not recommended for deterministic forecasts). For each variable the “deterministic” analog is provided in the final column of Table 1. The HiRA framework accommodates both distribution-based evaluation (which is a more model-oriented approach to verification) and a more user-relevant view, where a user is typically interested in whether a single specific threshold is exceeded, and whether the forecast models are getting better at making that decision. This is possible by computing the Brier score (BS; Brier 1950), or squared

![Fig. 1. Map showing the domain of the UKV and MOGREPS-UK models as well as the location of the WMO block 03 synoptic observing sites used by HiRA.](image-url)
probability error, where \( p_i \in [0, 1] \) is a sequence of \( i = 1, \ldots, N \) probabilities of exceedance, verified against whether the event occurred or not, where \( \alpha_i \in \{0, 1\} \):

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

Previously, “categorical” variables such as visibility, cloud-base height, cloud amount, and precipitation were evaluated for only three thresholds, using deterministic scores such as the equitable threat score (ETS). This was a Met Office choice. Over the years, there have been examples of experiential evidence provided by operational meteorologists that the improvements in skill for these thresholds were not replicated just above or below them, as well as where the forecasts were made worse. For example, there was one model upgrade where the visibility forecasts below 4000 m were improved, but forecasts of light haze (~5000 m) were made worse. This behavior did not become apparent after the model upgrade was made operational, but only when suitable conditions brought them to light. A more quantitative study into the characteristics of cloud cover forecasts showed that the behavior is not well described using only three thresholds in isolation, and a distribution approach is called for (Mittermaier 2012). Therefore, HiRA uses a larger number of thresholds evaluated together (as shown in Table 2) and computes a single ranked probability score (RPS; Epstein 1969) for each variable. In this case, forecast values and the observed value are binned into \( J \) categories, and the score is the sum of the squared differences over all categories in a cumulative distribution sense, where \( P_m \) and \( O_m \) are discrete cumulative distributions over \( J \) bins:

\[
RPS = \sum_{m=1}^{J} (P_m - O_m)^2. \tag{2}
\]

For variables such as temperature, the continuous ranked probability score (CRPS; Hersbach 2000) evaluates the entire distribution without binning and computes the integral of the squared differences between the forecast cumulative distribution \( P_f(x) \) and the observed Heaviside (step) function \( P_o(x) \):

\[
CRPS = \int_{-\infty}^{\infty} [P_f(x) - P_o(x)]^2 \, dx. \tag{3}
\]

The threshold sequences in Table 2 incorporate the range of most (if not all) possible values for cloud cover, 10-m wind speed, and hourly precipitation. For cloud-base height and visibility, the range of thresholds is restricted, as these variables can vary over several orders of magnitude and the values of interest are at the lower end. Some thought has also been given to possible user-relevant thresholds. Here, the sample size and the number of observed events (or base rate) become very influential, especially as the thresholds are pushed to the extremes, and some care is required when making the choice of threshold and in interpreting the results. For this paper we have restricted the thresholds to the following relatively modest values: temperature \(< 0^\circ C\), wind speeds \(\geq 17 \, m \, s^{-1}\), hourly precipitation \(\geq 4 \, mm \, h^{-1}\), cloud bases \(\leq 500 \, m\) (given 3 okta of cloud), and visibility \(\leq 1000 \, m\). It is worth noting that these thresholds need to be regionally specific, and in this instance they are specific to the U.K. climate and local user requirements.

The scores defined in Eqs. (1)–(3) are all “error” scores, in that a value of 0 indicates a perfect forecast. They can be converted to skill scores using a generic formula (Wilks 2006), where Sc can be any of the scores. The skill score is relative to a reference forecast. In this case, 24-h (observed) persistence is used. This is quite a challenging form of reference forecast, in that yesterday’s weather may be quite a good forecast for today. Mittermaier (2008) provides a more detailed analysis on the use of persistence as a reference. A perfect forecast has a skill score (SS) of 1:

\[
SS = 1 - \frac{Sc}{Sc_{ref}}. \tag{4}
\]

Finally, it is worth reiterating that HiRA, as a verification strategy, is still primarily focused on local forecast accuracy; even if a neighborhood is used, they are kept relatively small to ensure that the assumption of equi-probability is not invalidated.

4. Results

Thirty-six months’ worth of MOGREPS-UK and UKV forecasts, spanning the period October 2012–September 2015, have been evaluated using the HiRA
shown in Fig. 2, for each of the six variables as a function of the MOGREPS-UK control member and the UKV are presented to demonstrate that the combination of precipitation. The biggest gains for MOGREPS-UK are obtained by using a 3 × 3 neighborhood, compared with using a single grid point (i.e., no neighborhood). Note that MOGREPS-UK temperatures can only be improved up to a point, before the scores dip below those obtained at the grid scale (e.g., dotted lines). A similar pattern of behavior holds for wind speed. Therefore, it is possible for neighborhoods to be too large and become detrimental to local forecast skill, and it is conceivable that the scores for the other variables could also become worse for neighborhoods larger than those shown here, but these are unlikely to be computed when local forecast accuracy is the attribute of primary interest. With HiRA the neighborhoods are constrained to a maximum of 25 km, to ensure that the assumption of equiprobability is not invalidated. From Fig. 3, it is clear that most variables do not need neighborhoods that large. It is interesting to note that chapter 1 of the WMO guide to surface synoptic observations (WMO 2008) suggests that a synoptic observation should be representative of an area 100 km² around it. It does go on to concede that for some local “applications” it is more on the order of 100 km², or less. Representativeness considerations are aligned with the HiRA neighborhoods used for this study.

Another compelling feature of these results is that MOGREPS-UK maintains a higher level of skill for longer. The UKV shows a much steeper rate of decrease in skill with lead time. This suggests that using a larger neighborhood can account for some of the location (or timing) errors at the grid scale. However, the larger neighborhood may not be able to compensate for the fact that a single deterministic forecast represents only one realization, which may not match reality very well, and which will dominate the forecast error. An ensemble, on the other hand, provides multiple realizations,
and that alone provides more scope for counteracting forecast error. Using neighborhoods on top of multiple realizations adds another level of variability to help mitigate against the grid-scale component of the forecast error. The results show very clearly that the ensemble at the grid scale is already far superior in skill to the deterministic forecast. To conclude, the HiRA scores in Fig. 3 differentiate between the impact of error sources at the synoptic scale and those at, or near, the grid scale.

a. Making deterministic and ensemble scores comparable

What is the appropriate neighborhood size for comparing the 2.2-km 12-member MOGREPS-UK ensemble and the 1.5-km UKV scores to quantify the benefit of

![Fig. 2: Impact of neighborhood size on skill scores as a function of lead time for the MOGREPS-UK control member (MUKc) and UKV.](image)
the ensemble over a deterministic approach at the kilometer scale? Table 3 shows the range of neighborhood sizes currently available for consideration, along with their equivalent neighborhood lengths, and the number of pseudoensemble members.

From Table 3, it is clear that the ensemble provides 12 times the number of pseudoensemble members or number of grid points in the neighborhood. This acts as a constant inflation factor. Figure 3 shows that both UKV and MOGREPS-UK benefit from incremental increases in skill scores with increasing neighborhood size, but the size of the increments differs as the UKV benefits more.

To objectively derive a neighborhood size at which a fair comparison between these two models can be made, a skill differential $dSS/dNS$ is defined as the ratio of the difference in skill scores $dSS$ between successive neighborhood sizes and the neighborhood size difference $dNS$. This is calculated in kilometers, to account for
the fact that the models are not the same resolution. With the skill differential, a cutoff value $C_1$ can then be defined, below which the increment being added to the skill score by using a larger neighborhood can be considered negligible. The models are considered comparable at the respective neighborhood sizes at which $C_1$ is reached. A value of $C_1 = 0$ indicates the point where using a larger neighborhood would have a detrimental impact on skill. This would be a very strict criterion, and if the skill differential does not drop to below zero quickly, could lead to the use of very large neighborhoods for some variables (e.g., precipitation). This is shown below. A value of $C_1 = 0$ could also violate one of the initial assumptions of HiRA, which is that all the points in the forecast neighborhood are equiprobable outcomes at the observing site. As the methodology is still intent on measuring “local” accuracy, making the neighborhoods too large may make the association too tenuous for this to be believable, or realistic. Therefore, $C_1$ is best chosen as a small positive value.

A mean skill differential, calculated over the entire time series, and all lead times, is plotted for each of the variables and shown in Fig. 4. Also shown is the $C_1$ cutoff value ($C_1 = 0.0025$) used to establish which neighborhood sizes are comparable between the deterministic UKV and the MOGREPS-UK (MUK) ensemble. The sizes were identified as $11 \times 11$ (16.5 km) for the UKV and $3 \times 3$ (6.6 km) for MOGREPS-UK.

FIG. 4. The skill differential $dSS/dNS$ for each variable as a function of neighborhood size. Also shown is the cutoff value ($C_1 = 0.0025$) used to establish which neighborhood sizes are comparable between the deterministic UKV and the MOGREPS-UK (MUK) ensemble. The sizes were identified as $11 \times 11$ (16.5 km) for the UKV and $3 \times 3$ (6.6 km) for MOGREPS-UK.
skill scores for UKV and MOGREPS-UK based on these comparable neighborhood sizes (UKV, 11 × 11; MOGREPS-UK, 3 × 3). Both the monthly median scores (over all initialization times) and a 12-month running means of these median scores are plotted. Vertical lines indicate the model upgrade implementation dates when parallel suites (PSs) become operational suites (OSs). The impacts of the model change cycles, in terms of a response in the scores, are sometimes clearly visible; for example, for temperature OS34 led to a closing of the gap between MOGREPS-UK and UKV. Since OS37, the gap has opened up again. As both systems are sampling the same weather patterns in time, the temporal correlation in trends is high. For individual months, there are occasions when both the deterministic or ensemble monthly median scores are considerably better or worse than the running mean (e.g., September 2014 precipitation scores). By contrast, there are times when the way in which these model configurations are initialized can lead to diverging forecasts, and larger variations in skill, which are noticeable on a monthly time scale. Overall, there are few strong seasonal signals in the scores, with the exception of temperature and wind speed, which show higher scores in the colder seasons. The benefit in skill depends on the variable, but the running mean scores range between ~2% for temperature and ~40% for visibility. This highlights the need for ensembles for variables such as visibility. Fog formation is hugely sensitive to location, humidity, and temperature. With only a single forecast initialization, even negligible temperature deviations could determine whether the model will initiate fog formation or not. With an ensemble these very small (in magnitude) and local variations can be sampled more effectively, which will give a much better, probabilistic, estimate of fog formation. At best a deterministic forecast can provide a binary response at the grid scale. This is why the use of neighborhoods can improve the forecast, in cases where the model does produce fog, but not quite in the right place.

b. Performance for thresholds potentially related to hazards

Having established which neighborhoods are comparable, it is now worth focusing on the performance of the deterministic and ensemble forecasts for thresholds that are of interest to the user. Often these are associated with potential hazards. Figure 6 shows the model-oriented and the user-relevant results, comparing the 11 × 11 UKV neighborhood with the 3 × 3
MOGREPS-UK neighborhood. Figure 6 shows the skill scores—Brier skill score (BSS) and (continuous) ranked probability skill scores [(C)RPSS]—for the six abovementioned variables for a subset of months of the year. Often, the occurrence of events for user-relevant thresholds (listed in section 3) is highly dependent on the time of the year. For this reason some time constraints have been applied, to ensure that the signal is not unduly diluted by the lack of events. The relevant months and seasons, as well as the associated single thresholds, can be seen in headers of the subplots for the different variables. The symbol size denotes statistical significance at the 5% level, where larger symbols show that the differences are
significant. Differences in scores between the UKV and MOGREPS-UK simulations are significant at the 5% level, except for temperature at $t + 6\text{h}$.

Note how the single-threshold BSSs are generally lower than the (C)RPSS results, often much lower. This relates to the fact that the single user-relevant thresholds are often in the tail of the distribution and, thus, are harder to forecast correctly. Biases may also come into play more when using a single threshold. The exception is total cloud, which is a bounded quantity. Given that, U.K. climate totally overcast conditions are relatively common, which is reflected by the fact that the single-threshold and distribution-based scores are similar.

The rate of decrease in skill with lead time is generally similar between the single-threshold and distribution-based scores. Noteworthy exceptions include cloud-base height, where the BSS is higher at $t + 6\text{h}$ but shows a steeper rate of decrease compared with the RPSS. The hourly precipitation BSS is also essentially flat with lead time, and even goes negative beyond $t + 18\text{h}$, implying the forecasts are worse than persistence for a $11 \times 11$ UKV neighborhood. Clearly, the forecast neighborhood for hourly precipitation is not large enough for totals greater than 4 mm h$^{-1}$, because the scores are negative and, therefore, worse than persistence.

Seasonal subsetting of the time series shows that, for temperature and wind speed, the differences between the models’ distribution-based scores (RPSS and CRPSS) are smaller when comparable neighborhoods are used as compared those in Fig. 3. This hints at the greater predictability of synoptic weather patterns during the colder months and is reflected in the time series in Fig. 5.

5. Conclusions

The purpose of HiRA is to inform on the local accuracy of kilometer-scale NWP. It is a probabilistic verification framework, appropriate for kilometer-scale models, which in this study has been applied to raw deterministic and ensemble forecasts. HiRA facilitates the comparison between models at different resolutions, as well as comparing deterministic and ensemble forecasts.

Quantitative output from kilometer-scale models should not be taken at face value (i.e., at the grid scale) for the purposes of providing a “local” forecast for a specific location or observing site. It follows then that it should not be evaluated at the grid scale. When using convective-scale ensembles, the grid scale can be used, but only in an ensemble sense, where any derived probabilities are one step removed from the physical fields. The results presented in this paper are proof that using the grid scale only will not extract the best usable forecast information from kilometer-scale models. Even ensembles can benefit from neighborhoods, and overall they provide better forecasts for all six variables presented.

It could be argued that using neighborhoods on raw model output is a basic form of neighborhood post-processing, because the spatial sampling has been increased, and the same approach could be applied for producing site-specific postprocessed forecasts. We suggest that the results from HiRA can be used to inform on how to extract maximum value from raw forecasts for producing site-specific postprocessed forecasts.

More qualitatively speaking, there is useful information at, or near, the grid scale for discerning storm type and morphology, as well as features of interest such as precipitation gradients. Even so, a note of caution is required, as the numerical discretization of the model dictates that the effective resolution of an NWP model is several times the model grid resolution. The detail that is seen at the grid scale may not represent what users think it does.

Mittermaier (2014) clearly showed that this method of verification does not fabricate skill in a forecast when there is none to begin with. A neighborhood is needed to overcome the double-penalty effect, not to account for larger-scale spatial errors that are due to a poor synoptic-scale forecast. In this way, HiRA can differentiate between the benefit of the ensemble, which tries to counteract the larger-scale synoptic errors, and the errors that arise specifically at the kilometer scale, which grow rapidly and undermine local forecast accuracy. In this instance, this kind of spatial forecast assessment is fundamentally unlike other spatial methods [e.g., fractions skill score; Roberts and Lean (2008)], in that the focus is still very much on local accuracy at a particular location, and not about a pattern on a larger scale. Therefore, for verification at observing sites, the neighborhood should be large enough to minimize the double-penalty effect, without breaking the assumptions of equiprobability.

A skill differential has been used to quantify what neighborhood sizes are appropriate for comparing ensemble and deterministic forecast skill scores. This was used to determine a neighborhood size at which the skill being added is negligible or even detrimental (as defined by $e$). These comparable neighborhoods have been found to be $11 \times 11$ (16.5 km) for the UKV and $3 \times 3$ (6.6 km) for MOGREPS-UK. Over the 3-yr period of study, this translates to skill score differences of between 2% and 40% across the six variables considered and measures the true benefit in skill between a kilometer-scale ensemble and a deterministic forecast. Naively, it would also suggest that an extra 10 km of spatial accuracy is gained from a kilometer-scale ensemble.
There are several potential criticisms of the HiRA framework, which are the subject of ongoing work. Currently, square neighborhoods are used. These could be referred to as the “vanilla” or basic option. No attempt is made to reduce heterogeneity; that is, there could be a mixture of land and sea points within the neighborhood (if at the coast) and grid points with markedly different elevations, for example. At present, this heterogeneity appears to be beneficial in increasing the spread of the pseudoensemble. Testing is under way to explore the impact of “homogenizing” the neighborhoods. In homogenizing neighborhoods, another problem is created where even if the basic neighborhood definition is the same (e.g., a square), the number of forecast grid points that count toward the spatial fraction or distribution may be reduced. In aggregating scores, the granularity of the underlying distributions then becomes heterogeneous. The variation in the number of pseudoensemble members must be taken into account before aggregation. This could be achieved through the use of fair scoring rules (Ferro 2014). It is as yet unclear how big an effect this will have on the scores but is likely to affect smaller neighborhoods more; for example, for a $3 \times 3$ neighborhood, a pseudoensemble of 9 members could be reduced to 4 or 5 by excluding grid points that lie outside some predefined height tolerance. This has potentially more effect than if the same happened to a $17 \times 17$ neighborhood that provides 289 pseudoensemble members, of which 150 are removed. Understanding the skill of model configurations when forecasting extremes will also be explored through the use of threshold-weighted scores such as the threshold-weighted CRPS (twCRPS) proposed by Lerch et al. (2015). A weighting function is applied to the CRPS, so that only the tails of the distributions can be evaluated. There are therefore many opportunities for continued improvement of the assessment framework applied during this study.

Acknowledgments. The authors acknowledge the Met Office Public Weather Service for funding this work.

REFERENCES


