A Technique for Verification of Convection-Permitting NWP Model Deterministic Forecasts of Lightning Activity

JONATHAN M. WILKINSON
Met Office, Exeter, United Kingdom

(Manuscript received 1 June 2016, in final form 5 October 2016)

ABSTRACT

This manuscript introduces a new technique for evaluating lightning forecasts from convection-permitting models. In recent years, numerical weather prediction models at the convection-permitting scales (horizontal grid resolutions of 1–5 km) have been able to produce realistic-looking forecasts of lightning activity when compared with observations. However, it is challenging to assess what value these forecasts add above standard large-scale indices. Examining this problem, it is found that existing skill scores and neighborhood verification methods are unable to cope with both the double-penalty effect and the model’s variable frequency bias. A displacement distance and a quasi-symmetric distance score are introduced based on the distance between the model and the observations, the latter showing any improvement the forecast has over a completely “hedged” forecast. This can be combined with a domain-improved contingency table and comparisons between modeled and observed lightning flashes to evaluate the forecast performance in three important dimensions: coverage, distance, and intensity. The verification metric is illustrated with a single case, which shows that the convective-scale U.K. variable resolution model (UKV) delivers improved forecasts compared with the large-scale indices in both coverage and distance. Additionally, a month-long analysis is performed, which reveals that the coverage of lightning is in good agreement with the observations; lightning is displaced by the model by a distance on the order of 50–75 km, but the model overpredicts the lightning intensity by at least a factor of 6 after observational detection efficiencies have been considered.

1. Introduction

Lightning is hazardous to human health (e.g., Elsom 2001; Ashley and Gilson 2009), can cause issues for aviation safety (e.g., Mäkelä et al. 2013; Wilkinson et al. 2013b), and causes property damage (e.g., Curran et al. 2000). Accurate forecasts of when and where lightning is likely to occur are useful in minimizing the risks to human life and property. Traditionally, lightning events have been forecast by using large-scale instability indices, which have been evaluated by several studies in the literature (e.g., Haklander and van Delden 2003; Vujović et al. 2015). Such indices rely on atmospheric variables that vary only slowly in time and space, leading to lightning being forecast over large geographical areas. In addition, these indices are not based on the cloud microphysics that creates charge separation and lightning.

Because of advances in supercomputer processing capability, the horizontal grid scales of many operational numerical weather prediction (NWP) models have been reduced in recent years, with many centers now running convection-permitting models with horizontal grid lengths of 1–5 km. At the Met Office, a U.K. variable resolution model (hereafter UKV) has been operating since 2009 (Tang et al. 2013), with a horizontal resolution of 1.5 km in its central domain and lower resolution closer toward the boundaries. Such resolutions allow thunderstorms to be partially resolved.

Meanwhile, a number of modeling studies have examined attempts to explicitly represent lightning flash rate in terms of parameterizations in convection-permitting models. A number of different approaches to modeling the flash rate have been developed. McCaul et al. (2009) used a statistical relation, linking lightning flash rate to the graupel flux at the −15°C level and the ice in the column. Dahl et al. (2011a,b) modeled the thunderstorm dipole as a capacitor. Yair et al. (2010) developed a lightning potential index (LPI) based on the potential for charge generation and separation due to cloud microphysics. Lynn et al. (2012) used the LPI to generate the potential electrical energy of the storms and produced cloud-to-ground and intracloud flash rates. Barthe and Pinty (2007a,b) and Fierro et al. (2013)
explicitly modeled the charging physics of the cloud, with lightning flash discharge schemes. In addition, some model parameterizations relate flash rate to cloud-top height, following Price and Rind (1992). All of these schemes produced localized predictions of lightning occurrence, in contrast to the large-scale indices.

Most of the studies using convection-permitting models studies include some verification of the lightning flash rate. For example, McCaul et al. (2009) evaluated their algorithm against lightning mapping array data. However, comparisons of observed and modeled lightning locations (which is dependent on the whole of the model, not just the lightning parameterization) are often subjective in nature. Exceptions to this include the work of Lynn et al. (2012, 2015), who produced some simple verifications using the probability of detection (POD; aka hit rate) and probability of false detection (POFD; aka false alarm rate). Lynn et al. (2012) found that these scores improved when forecasts were regridded from their native 4-km grid to 12- and 36-km grids. Lynn et al. (2015) examined equitable threat scores (ETSs) for lightning forecasts using a neighborhood approach (e.g., Roberts and Lean 2008; Ebert 2009; Clark et al. 2010) and a range of neighborhoods from 12 to 96 km. Lynn et al. (2015) found that the ETS increased with increasing neighborhood size. However, as far as can be ascertained from the literature, no measure has been able to show what value such forecasts have above large-scale indices.

In this manuscript, a new technique for determining the accuracy of lightning forecasts is developed and illustrated using UKV model lightning forecasts with the McCaul et al. (2009) parameterization, chosen because of its ease of inclusion in the UKV model and past successes in forecasting lightning around the United Kingdom: while the McCaul et al. (2009) parameterization was developed for severe storms, Wilkinson and Bornemann (2014) found it was a useful predictor of lightning risk for a high-profile U.K.-based case study with lower flash rates. For comparison, a large-scale stability index relevant to the United Kingdom (Boyd 1960) is also examined, to see if the McCaul et al. (2009) parameterization shows improvement over the large-scale index.

Existing techniques are reviewed in section 2, illustrating why these techniques were not suitable for verification of the UKV data. The derivation of a new coverage–distance approach is discussed in section 3 and applied to the UKV data in section 4. A summary discussion is provided in section 5.

2. Issues with existing forecast verification techniques

a. Evaluation at the grid scale

Traditionally, binary forecasts of events such as lightning have been classified as a contingency table, as shown in Table 1. There are four possible options dependent on whether the model forecasts an event or not and whether an event is observed or not.

Table 1. Basic contingency table (as used for traditional verification of binary events).

<table>
<thead>
<tr>
<th>Lightning forecast</th>
<th>Lightning observed</th>
<th>Lightning not observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits (a)</td>
<td>False alarms (b)</td>
<td></td>
</tr>
<tr>
<td>Misses (c)</td>
<td>Correct rejections (d)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 shows forecasts and observations of lightning activity for the period 2300–2359 UTC 17 July 2014. Observations of lightning occurrence are taken from the Met Office’s Arrival Time Difference Network (ATDnet) system, which has 11 stations across the United Kingdom and Europe and is described in more detail by Gaffard et al. (2008). More recently, Enno et al. (2016) reported detection efficiencies for ATDnet of 89% for cloud-to-ground lightning and 24% for intracloud lightning. While better lightning detectors exist elsewhere at the time of writing, there are no total lightning detectors used operationally that cover the whole of the United Kingdom, so it is not possible to verify the model performance with these systems. To avoid issues affecting the verification technique proposed, any hours where fewer than 20 ATDnet observations are recorded are removed from the study.

Lightning forecasts are generated by the UKV model, a configuration of the Met Office Unified Model (MetUM). The forecast has been initialized from the Global MetUM at 1200 UTC 17 July 2014, ahead of the lightning event to allow convection in the model to spin up ahead of the lightning event taking place. The UKV model data are also used to generate two different forecasts of lightning. One is based on the McCaul et al. (2009) parameterization in the UKV model. As such lightning forecasts will be both dependent on the McCaul et al. (2009) parameterization and the rest of the UKV model; these forecasts are hereafter labeled UKV-McCaul. For comparison, the other measure used is the Boyd index (Boyd 1963), which is a simple measure of instability, dependent on the thickness of the 1000–700-hPa layer and the temperature at 700 hPa. Further details of the UKV model are presented in appendix A.

A mask can be applied to each forecast and observation. For the ATDnet data, this simply involves taking the latitude/longitude of each lightning observation and generating a mask at the model grid scale. For the Boyd index, values greater than or equal to 94 are assumed to be at risk of lightning (see appendix A).
while for the UKV-McCaul scheme, each point where the model predicts one or more lightning flashes in that time window is assumed to be a forecast of lightning.

Examining Fig. 1, it can be seen that the Boyden index overestimates the area of the model domain that is at risk of lightning. This is because the temperature and thickness properties input into the index vary only slowly with time and horizontal distance. The UKV-McCaul forecast has a smaller proportion of the domain at risk than the Boyden index; this finding is similar in nature to that of McCaul et al. (2009), who found that their forecast lightning locations were much lower than the area where CAPE was positive. However, it can be seen that the proportion of the domain where lightning is forecast is still greater than the proportion of the domain where lightning is observed. As an additional test, using the Boyden index with a slightly higher threshold of 97 produces a lower proportion of the domain with lightning risk but still carries a high number of false alarms.

The location of the lightning forecast by the UKV-McCaul forecast is more specific than the Boyden index, but the UKV-McCaul system has misplaced the location

---

**Fig. 1.** Forecast and observed lightning for 2300–2359 UTC 17 Jul 2014. (a) Number of predicted lightning flashes by the UKV model using the McCaul et al. (2009) scheme. (b) Scatterplot of observed lightning locations as recorded by ATDnet between 2300 and 2359 UTC 17 Jul 2014. (c) Boyden index valid for 2300–2359 UTC produced from the same model data as used to produce (a). (d) Lightning risk/no-risk mask for the data in (a). (e) As in (d), but for the ATDnet data mapped onto the UKV model grid. (f) As in (d), but with a gray mask for Boyden index ≥ 94 and a black mask for Boyden index ≥ 97. To avoid confusion with the data points, the coastlines are not plotted in (d)–(f). For the purposes of visualization only, the mask plotted in (c) has been increased to a $3 \times 3$ square centered on the lightning observation; without this modification, the data would be invisible to the naked eye. The unmodified mask is used in all further calculations in this manuscript.
of lightning in the southwest of the domain by around 100 km. In addition, the lightning over south-central England is poorly represented in the model forecast for this time and date, with hardly any predicted lightning being recorded. Figure 2 shows the accumulated precipitation from the model forecast and derived from the U.K. weather radar for the same time period as Fig. 1. This shows the same misplacement in the precipitation location in the southwest of the domain. In the observations, the precipitation was in the form of an arc, which the UKV model has managed to represent to some degree, but the model has positioned the arc too far to the west-northwest and underestimated the intensity of the precipitation in the eastern flank of the arc. This suggests that the whole cloud system in this area was forecast weaker than observed, leading to an incorrect forecast of lightning activity.

A contingency table based on Table 1 can be drawn up for the data in Fig. 1, and this is presented as Table 2, which also includes the values of four skill scores derived from the contingency table. They are POD and POFD, as described, for example, in Wilks (1995), along with the symmetric extreme dependency score (SEDS) from Hogan et al. (2009) and the symmetric extremal dependence index (SEDI) from Ferro and Stephenson (2011). Both scores are chosen for their ability to judge a forecast in terms of a single score, whereas for POD/POFD both are required together to make an evaluation. In addition, SEDI and SEDS have useful properties of equitability and are suitable for forecasting rare events, such as lightning.

The POFD for the UKV-McCaul forecast is smaller than the two Boyden indices, but because the forecast shown in Fig. 1 mismatches the location of the lightning slightly, the number of hits on the grid scale is small. This illustrates that a double-penalty effect occurs, where a forecast that is misplaced by as little as a single grid box is counted as both a miss and a false alarm in the contingency table, whereas at a coarser model resolution, it would have been counted as a hit. This effect means that the SEDS and SEDI skill scores for the Boyden index above 97 show better skill than does the UKV-McCaul forecast, despite the UKV-McCaul forecast having a smaller bias and appearing to be subjectively better. Therefore, the implication of the double-penalty effect is that the verification at the grid scale is not valid for the UKV-McCaul forecast. Thus, some form of neighborhood processing approach is required to verify the lightning forecasts.

b. Existing neighborhood-based approaches

To avoid issues with the double-penalty effect, Roberts and Lean (2008) introduce the concept of a fractions skill score (FSS) to evaluate precipitation forecasts. This involves establishing a square “neighborhood” of increasing size from the grid scale out to the whole domain size. Roberts and Lean (2008) show that for good precipitation forecasts, the FSS increases rapidly for increasing neighborhood size and asymptotes toward its highest value when the neighborhood size reaches the domain size. However, the FSS asymptotes to 1 for an unbiased forecast; more specifically Mittermaier and Roberts (2010) show that forecasts with high frequency bias give low (poor) values of FSS. Unfortunately, because of the significant frequency bias reported in Table 2, the FSS for the UKV-McCaul forecast in this case asymptotes to 0.14 (not shown), and therefore this metric also makes the forecast produced by UKV-McCaul appear to be poor. Mittermaier and Roberts (2010) used percentile thresholds to try and
account for the bias in forecasts of precipitation coverage, but this would be difficult to do with a lightning/no-lightning coverage mask, which is inherently binary. It is particularly problematic in situations where there are scattered storms, producing lower lightning amounts (common over the United Kingdom), where the observations of lightning flash rates are more sporadic in nature and not smooth fields.

While it is possible to change the model parameterization so that the bias is close to 1, generating forecast statistics for past cases will not be possible without re-running every single forecast, which is computationally expensive. In addition, it was found that the bias varies significantly from hour to hour. This is illustrated in Fig. 3, which shows the bias for the model run shown in Fig. 1; the bias here can clearly be seen to vary between values of less than 1 and 40 in the space of a few hours. Part of this bias may be due to errors in the model phase (e.g., producing convection too early or too late), but even if this was taken into account by examining the bias over a longer time period, some variability in the bias would still exist. If the parameterization was scaled to match the mean bias, there would still be a significant number of hours when the bias was greater than 1, leading to poor FSS values. If the bias was corrected for each individual forecast, the model would no longer be independent of the observations, leading to suggestions that the verification is biased in favor of the model. Thus, an alternative metric that is able to cope with a bias value that is not close to 1 is desirable.

It is possible to adopt the neighborhood approach of Clark et al. (2010) and Lynn et al. (2015), with a hit being recorded where lightning is both observed and forecast within the neighborhood. However, doing this would ignore any bias there may be in the neighborhood: multiple model points where lightning is forecast could count a single observation of lightning as a match, translating into spurious multiple hits. Such verification may disguise issues with the model or its lightning parameterization scheme, such as the lightning forecast being spread out over too large an area. In addition, the number of hits rises steadily as the neighborhood size increases. At the domain size, the number of hits would total the number of points in the domain, with the number of misses, false alarms, and correct rejections being equal to zero. If there is any bias within the domain, this information would be lost by using the neighborhood method.

For precipitation forecasts, Wernli et al. (2008) developed a structure, amplitude, and location (SAL) verification methodology. However, this methodology was found difficult to adapt directly to the lightning data, particularly for lightning observed around the United Kingdom, which can be very sporadic in nature and difficult to verify with the structure component of the Wernli et al. (2008) score. However, it is possible to take the principles of Wernli et al. (2008) and adapt these for lightning forecasts. This method is based on coverage (important for comparisons for large-scale indices), distance between the observed and modeled features, and the intensity of the lightning produced and is introduced and described in the following sections.

3. A new coverage–distance–intensity (CDI) verification method

a. Derivation of a domain-maximum contingency table

An alternative approach to the methods described in section 2b would look to keep the same bias with increasing neighborhood size. For example, if in a
neighborhood there are 10 points that forecast lightning and 2 points where lightning is observed (i.e., a bias of 5), only 2 hits should be allowed for that neighborhood, and the remaining 8 points in the neighborhood should be counted as false alarms. This can be done by matching each observation of lightning to its nearest point in the model dataset and considering these to be a forecast–observation pair. The distance between the two points in the pair is $D_{of}$. As the neighborhood size $D_{neigh}$ increases, any forecast–observation pairs that are within the same neighborhood (i.e., $D_{of} \leq D_{neigh}$) but not coexistent are changed from being a false alarm and a miss to a hit and a correct rejection, as if they were collocated at the same point. Thus, the double-penalty effect discussed earlier is canceled with increased neighborhood size.

To illustrate this, Fig. 4 shows the results of this process with increasing neighborhood size for the model and observation data presented in Fig. 1. In Fig. 4a it is seen that the number of hits increases rapidly with increasing neighborhood size as the double-penalty effect is canceled; thus, the hit rate is a function of neighborhood size. It should also be noted that there is little change once the neighborhood size goes beyond 60 km. The curves in Figs. 4a–d reach their asymptotes at 87 km. Assuming the number of model points with lightning in the model is $n_{fl}$, the number of points with lightning in the observations mask is $n_{ol}$, and the number of points in the domain is $n_{dom}$, then the hits, misses, and false alarms will always asymptote to

$$
a_{dom} = \text{MIN}[n_{fl}, n_{ol}]
$$

$$
b_{dom} = \text{MAX}[n_{fl} - n_{ol}, 0]
$$

$$
c_{dom} = \text{MAX}[n_{ol} - n_{fl}, 0]
$$

$$
d_{dom} = n_{dom} - (a_{dom} + b_{dom} + c_{dom}).
$$

(1)

The bias at the domain $B_{dom}$ is the same as that at the grid scale $B_g$ and can be expressed as

$$
B_{dom} = B_g = \frac{n_{fl}}{n_{ol}}.
$$

(2)

As $n_{fl}$ and $n_{ol}$ are known from the raw grid-scale model and observational data, the values of the asymptotes can be predicted without determining the distances $D_{of}$, and Eq. (1) is thus independent of the distance between the points. This has two important implications: first, the
values derived by Eq. (1) are purely a measure of the skill of the forecast in covering the correct number of points in the mask; any skill scores derived from Eq. (1) would be the same if the model had correctly forecast the location of lightning or if it was at the opposite end of the domain. This means that an additional distance metric is required; this will be derived in section 3d. Second, the worst performance that the model can produce is to produce a completely “hedged” forecast. The definition of hedging is taken to be that used by Marzban (1998) and Brill (2009), where a model or forecaster produces forecasts with a high-frequency bias, greater than 1, in order to maximize their success of correctly forecasting an event. The UKV forecast would be completely hedged if it forecast lightning at every single point in the domain.

b. Producing a coverage skill score

It is possible to simply use the frequency bias as a measure of the skill of the model in getting the correct coverage of the forecast domain. However, this is likely to vary over several orders of magnitude. The maximum bias for the UKV domain is the number of points in the domain ($n_{\text{dom}}$; 503 820). The bias will vary for the same forecast, but with different amounts of observed lightning. It will also be dependent on the domain size, and it will be tricky to compare the skill of two forecasts produced on different model domains. Therefore, it is more appropriate to show if the forecast is better than a completely hedged projection. If the forecast is better, it would be good to define some skill score, so that the completely hedged forecast (defined as $a_{\text{dom}} = n_{\text{ol}}$, $b_{\text{dom}} = n_{\text{dom}} - n_{\text{ol}}$, and $c_{\text{dom}} = d_{\text{dom}} = 0$) will score 0. A perfect forecast (defined as $a_{\text{dom}} = n_{\text{ol}}$, $b_{\text{dom}} = c_{\text{dom}} = 0$, and $d_{\text{dom}} = n_{\text{dom}} - n_{\text{ol}}$) should score 1. This is the first property that is desirable from the skill score:

1) Is equitable—The score will produce 1 for perfect coverage and 0 for a completely hedged forecast that forecasts lightning on every model grid point. Using the data obtained from Eq. (1), several skill scores can be used to determine the skill that the UKV model has in forecasting the coverage of lightning. However, it is important to ensure that the score produces sensible values and is able to distinguish between good and poor forecasts of lightning coverage. To do this, two further desirable properties of a good skill score have been determined. These are as follow:

2) Produces real numbers—The score will produce real numbers for a variety of values in Eq. (1).

3) Varies slowly between the score limits with increasing bias—As the bias increases, the value produced by the score should change slowly, irrespective of the number of observations in the model domain. The properties of some skill scores mean that they decrease to values close to zero very quickly as the number of observations increases, which should be avoided.

Property 2 is hopefully obvious: it is important that the skill scores should not produce nonreal numbers. The reasoning behind property 3 is less obvious, but it is important if forecasts of significantly different biases are to receive significantly different scores. This is best illustrated by examining Fig. 5, which shows the variation of three skill scores [SEDS, critical success index (CSI), and $1 - \text{POFD}$] for increasing numbers of points expected in the lightning observations mask. It can be seen that for all cases, there is little change in the value of CSI when the bias increases above 10; thus, a forecast with a bias of 50 and a bias of 100 will have very similar scores, despite the latter being much poorer in terms of coverage. For Fig. 5a, where there are only 50 observations of lightning, there is very little change in the $1 - \text{POFD}$ measure with increasing bias. This is because even for a bias of 200 (10 000 points in the forecast mask), the score is dominated by the correct rejections ($d_{\text{dom}}$: 493 820 points). It is only when there are many fewer correct rejections in Fig. 5d that the score starts to decrease as bias increases. The SEDS skill score is the only one of the three measures that produces a consistently different score for frequency biases of 50, 100, and 150 for the expected numbers of observations.

Table 3 examines these skill scores against the three desirable properties described above. It can be seen that SEDI will not produce a real number in the case that POD or POFD equals either 0 or 1, meaning that in Eq. (1), it will produce invalid results. Out of the remaining scores, the CSI measure will not produce a score of 0 for the completely hedged case. The POD score will produce a value of 1 in any case where $n_\beta \geq n_\alpha$ and so will produce a value of 1 for the completely hedged case. Only $1 - \text{POFD}$ and SEDS will produce 1 for the perfect forecast and 0 for the completely hedged case. However, as seen in Fig. 5, $1 - \text{POFD}$ is highly sensitive to the number of observations and, in the case of low observations, will not be able to distinguish between a forecast with no bias and one with a much larger bias. Therefore, the SEDS score will be used as a measure of the skill of the model in coverage.

c. Displacement distance

As noted in section 3a, any skill score based on the values of Eq. (1) has little value unless it is accompanied by some measure of the spacing between the model and the observations. Intuitively, one may choose either the asymptotes (e.g., 87 km in Fig. 4) or to calculate the peak SEDS score from the values in Eq. (1) and then use this
to define a horizontal distance scale (e.g., distance at which the model reaches 90% of the peak SEDS). However, this is also problematic for two reasons. First, it is easy to “hedge.” If the UKV were to forecast lightning at every grid point, the distance at which the peak SEDS score occurs will be zero. This could lead to the model appearing to be skillful when in fact it is not. The second issue with this measure will occur when lightning is forecast both close to and far away from the observations in the same domain. For example, consider a model forecast of lightning at two locations (A and B) within the same domain, which are spaced far apart, as shown in Fig. 6. Location A is close to some observations of lightning, while location B is much farther away from both location A and the observations. In this case, as a result of location A and the observations being close together, 90% of the peak SEDS distance will be roughly 90% of the distance between the observations and location A. Location B will not be penalized if this distance metric is used, thus giving the impression that the model forecasts lightning closer to the observations than in reality. It is therefore preferable to derive some form of distance between the model and the observations, valid for all points. This is done using a horizontal displacement distance $D_{dis}$, defined as

$$D_{dis} = \frac{D_{o \rightarrow f} + D_{f \rightarrow o}}{2}.$$  

FIG. 5. Skill scores as a function of frequency bias for (a) 50, (b) 100, (c) 500, and (d) 1000 observations. The total number of points in the mask is 503,820.

<table>
<thead>
<tr>
<th>Skill score</th>
<th>Valid for Eq. (1)</th>
<th>Perfect score</th>
<th>Hedged score</th>
<th>Good $\Delta$Score $\Delta$Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>Yes</td>
<td>1.0</td>
<td>Consistently 1.0</td>
<td>No</td>
</tr>
<tr>
<td>1 – POFD</td>
<td>Yes</td>
<td>1.0</td>
<td>0.0</td>
<td>Only if $n_{o}$ is large</td>
</tr>
<tr>
<td>CSI</td>
<td>Yes</td>
<td>1.0</td>
<td>$\frac{n_{o}}{n_{dom}} (&gt;0)$</td>
<td>No</td>
</tr>
<tr>
<td>SEDI</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SEDS</td>
<td>Yes</td>
<td>1.0</td>
<td>0.0</td>
<td>Yes</td>
</tr>
</tbody>
</table>
distances. Similarly, \( D_{f\rightarrow o} \) is the mean of the array of distances from each model point to its nearest observation. We can express \( D_{\text{dis}} \) as a unit of distance, either in standard units (e.g., m, km, mi) or in terms of the number of model grid squares.

Figure 7 shows a schematic example of this in practice. A simple domain is set up in Fig. 7a with forecast points in gray and the observations on the forecast grid in black. In Fig. 7b, \( D_{o\rightarrow f} \) for the three observation points is calculated. Each observation is mapped to its nearest neighbor where lightning is forecast. The value of \( D_{o\rightarrow f} \) for each point is the straight-line distance between the center of the model grid where lightning occurs and the center of the model grid where lightning is forecast and is expressed here in terms of the model grid spacing, so values of \( D_{o\rightarrow f} \) for these three points are \( \sqrt{17} \), \( \sqrt{2} \), and \( \sqrt{10} \) respectively. The mean of these values, \( D_{o\rightarrow f} \), is 2.90 grid lengths. Similarly in Fig. 7c, values for \( D_{f\rightarrow o} \) are \( \sqrt{2} \), \( \sqrt{8} \), \( \sqrt{17} \), and 5. The mean, \( D_{f\rightarrow o} \), is 3.34 grid lengths. Therefore, \( D_{\text{dis}} \) is 3.12 grid lengths. If this was the UKV model with 1.5-km grid spacing, \( D_{\text{dis}} \) would therefore equal 4.68 km.

It should be noted that individual lightning strikes may be detected some distance from the main convective core of the storm. This is especially true for mesoscale convective systems and intense supercell storm activity. However, the evaluation of distance is done using a lightning mask, which itself is based on a verification period of at least an hour where numerous lightning strikes are recorded, many of which are nearer to the core. As a result, the displacement distance should not be adversely affected. For weaker and smaller storms with lower flash rates, the lightning locations recorded in the observations should be nearer the storm core.

Looking at Fig. 7b, it should be noted that in determining the value of Eq. (3), two (or more) forecast points in the mask are allowed to consider a single observation point in the mask as their nearest neighbor (and vice versa in Fig. 7c); this is opposite the behavior used in section 3a. This does not cause an issue, because every model and observation point is included in the calculation of the displacement distance \( D_{\text{dis}} \). If there are many forecast points that are far away from the observations (as in the schematic in Fig. 6), then \( D_{\text{dis}} \) will increase to reflect this. Similarly, if all forecast and observation points are close together, then \( D_{\text{dis}} \) will be small. It is important to note the denominator in Eq. (3): if instead a weighted average based on the number of model and observation points was used, a forecast with high bias that overpredicts lightning occurrence would produce a \( D_{\text{dis}} \) value very close to \( D_{f\rightarrow o} \). It is better to equally weight the \( D_{f\rightarrow o} \) and \( D_{o\rightarrow f} \) to get a measure of the distance, while letting the SEDS_{dom} score highlight whether the forecast is biased or not: it is not possible by design for SEDS_{dom} to achieve a perfect score if the forecast has any bias.

d. A quasi-symmetric distance score

The displacement distance is a useful measure for indicating the spacing between the observed and forecast features in the domain. However, it does not indicate whether or not the forecast is better than a random or hedged forecast. It would be useful to derive a distance-based skill score, which has similar properties to SEDS; that is, a value of 1 would indicate a perfect forecast, with a hedged forecast scoring zero and the skill score having values in the range from -1 to 1.

To do this, it is necessary to define a distance \( D_{h} \) that would be the \( D_{\text{dis}} \) value of a completely hedged forecast. This can be obtained by inputting into Eq. (3) the observations mask along with a forecast mask that forecasts lightning over 100% of the domain (i.e., a completely hedged forecast). By definition, this hedged forecast will also always have a SEDS_{dom} score of exactly 0. If a model forecast has a \( D_{\text{dis}} \) value greater than \( D_{h} \), it would have worse positioning skill than forecasting lightning everywhere in the domain.

It is also possible to produce a distance \( D_{m} \), which is the farthest distance possible between the model and observations. This is done by working out the mean distance between each of the four corners of the domain and each point in the observations, producing an array of four points, one for each corner, where the greatest of
these distances is $D_m$. Both $D_h$ and $D_m$ will vary from case to case, dependent on the position of the observations within the model domain and also the size of the model domain. They will be largest when the observations are close to the corner of a large domain. The values of $D_h$ and $D_m$ will diminish as the observations move toward the center of the domain.

Using $D_h$ and $D_m$, a linear-logarithmic approach can be used to construct a quasi-symmetric distance score (QSDS) as

$$QSDS = \begin{cases} 
1, & D_{\text{dis}} = 0 \\
1 - \frac{1}{2} \left[ \frac{D_{\text{dis}}}{D_h} + \frac{\ln(D_{\text{dis}})}{\ln(D_h)} \right], & D_h > D_{\text{dis}} > 0 \\
0, & D_h = D_{\text{dis}} \\
\frac{1}{2} \left[ \frac{D_{\text{dis}} - D_m}{D_h - D_m} + \frac{\ln(D_m - D_{\text{dis}})}{\ln(D_m - D_h)} \right] - 1, & D_m > D_{\text{dis}} > D_h \\
-1, & D_{\text{dis}} = D_m
\end{cases}$$

\[(4)\]

The linear-logarithmic approach is a compromise between a linear function and a logarithmic one. For the data in Fig. 1, the value of $D_h$ is 174.4 km. If the linear-only portion is chosen for this data, the point at which QSDS reaches 0.5 is equal to 87.2 km ($0.5D_h$), which was felt to be too generous to the model. If the logarithmic portion of Eq. (4) is only used, then QSDS reaches 0.5 at $\sqrt{D_h}$, which equates to 13.2 km for the data in Fig. 1, meaning that it would be very difficult for the model to obtain a QSDS score above 0.5. For the UKV model grid spacing of 1.5 km, the QSDS score will rapidly transit from 0.5 to 1.0 in the space of just nine grid boxes, as $D_{\text{dis}}$ tends to zero. However, using the full form in Eq. (4), the QSDS will reach 0.5 at a value of $D_{\text{dis}} = 45.4$ km, which is stringent but also means that there is not a rapid transition from 0.5 to 1 in the space of a few grid boxes.

It should be noted that QSDS will produce sensible values while the value of $D_h$ is less than 1500 km. If the value of $D_h$ exceeds this figure, then credit may be given to significantly displaced forecasts during the verification process. However, this situation is only likely to occur with a very large model domain and sporadic amounts of lightning activity. It is also worth stressing that the QSDS score will be representative of the entire domain over which it is calculated and will therefore be a score based on contributions from all of the storms in this domain. Therefore, as with all verification metrics, sensible interpretation is required. Should the value of $D_h$ regularly exceed 1500 km, the model domain could be divided into a series of subdomains for verification over more specific regions. This behavior may also be desirable to help determine how the lightning forecasts
verify in a specific area, for example a few states selected from the entire continental United States domain.

Figure 8a shows the QSDS curve from Eq. (4) plotted using parameters derived from the data in Fig. 1. The curve is constrained to the range $[-1, 1]$, with the value of $D_m$ calculated as 933 km. There is a noticeable change in the gradient of the graph when it reaches a $D_h$ value of 174.4 km. This is an unfortunate feature owing to the fact that it is not possible to preserve the symmetry of the QSDS curve unless $D_h = 0.5D_m$; this situation is illustrated in Fig. 8b. If symmetry is desired, it is possible to set $D_m$ to be double the distance of $D_h$ and set any forecast $D_{dis}$ value that exceeds $D_m$ to $-1$. However, for the UKV model, it was decided not to do this, so that a forecast with QSDS of $-1$ represents the worst possible separation between the model and the observations.

**e. Intensity of lightning forecast**

Based on the forecast flash rate in each grid column, the model discharges a number of integer flashes on each time step. Noninteger remainders are kept with the storm system and passed to the following time step, meaning that all lightning activity is included when the flash rate is less than the inverse of the time step length. If the total number of forecast flashes recorded in the domain during the hour-long verification period is $T_F$ and the total number of flashes observed by ATDnet during the same period is $T_O$, then the flash bias (FB) can be expressed as

$$FB = \frac{T_F}{T_O}.$$  \hspace{1cm} (5)

However, the flash bias will take values in the range $[0, \infty]$. To produce a bounded score, the equation for the amplitude component of Wernli et al. (2008) can be modified so that like QSDS, it varies between $-1$ and $1$. This intensity score $I$ is expressed as

$$I = \frac{T_F - T_O}{T_F + T_O},$$  \hspace{1cm} (6)

with 0 being a perfect score. A score of close to 1 indicates significant overforecasting, while a score close to $-1$ indicates significant underforecasting of the lightning intensity.

However, for such comparisons, the fact that lightning detectors do not observe all lightning activity must be noted. If the model was to correctly forecast every lightning strike that occurred, but the lightning detector only recorded a percentage of the strikes, the model would have a true bias of 1, but there would be an apparent bias due to sensor deficiencies. For simultaneous ATDnet comparisons with a lightning mapping array, Enno et al. (2016) observe that ATDnet records 89% of...
cloud-to-ground lightning and 24% of intracloud lightning activity. This information can be used to construct a curve of apparent bias and intensity plotted against the ratio of intracloud to cloud-to-ground lightning activity, as shown in Fig. 9. If the ratio of intracloud to cloud-to-ground lightning is known, then the maximum apparent bias and intensity can be predicted; otherwise, the maximum apparent bias is the asymptote (where all lightning activity is assumed to be intracloud). Should the model have a bias above the maximum apparent bias, it is clearly overforecasting lightning activity. If the value is below the maximum apparent bias, then it is not clear whether the difference is due to the model or the detection efficiency of the lightning sensors.

f. When no lightning is forecast or observed in the model domain

If lightning is not forecast or observed in the domain, the number of hits will be zero and the SEDS skill score will produce an invalid value (natural log of zero). Similarly, \( D_{\text{dis}} \) will be undefined and QSDS will not be defined. Examining archived UKV model data and ATDnet data, this situation does occur, but rarely and in marginal lightning situations. For larger domains than the UKV, a model domain without forecast lightning or an observation-free domain would be rarer, especially for regions where the lightning flash density is higher than across the United Kingdom. Rather than ignore such forecasts, a better strategy would be to maintain a record of hours when lightning was forecast and not observed (“whole domain false alarms”) and hours when lightning is observed but not forecast (“whole domain misses”) and reporting these separately as well as the SEDS, QSDS, and \( I \) scores. Another alternative in marginal situations would be to increase the sampling period from a 1- to a 3-h window. This could help evaluate model forecasts that are slightly earlier or slightly later than the observations.

4. Results for UKV lightning forecasts

a. Case of 17 July 2014

The UKV model forecast in Fig. 1 was initialized at 1200 UTC 17 July 2014 and ran for 21 h until 0900 UTC 18 July 2014. As a result of initialization from the global model, it took a few hours before the UKV produced convection that contained lightning. After this time, lightning was observed and forecast in the domain for each hour of the forecast. For each hour from 1500–1559 UTC 17 July 2014 until 0800–0859 18 July 2014, the forecast data from the UKV-McCaul forecast and the Boyden index were analyzed. In the case of the Boyden
index, the two thresholds (≥94 and ≥97) used previously were applied. As in Fig. 1, a binary mask was applied to the data by taking each forecast point where there was either a single lightning flash produced by the McCaul parameterization as a forecast of lightning. Similarly, two forecasts of the Boyden index were generated, one for each threshold with any point in the domain with a value above the threshold deemed to be a forecast of lightning.

From this information, values $a_{\text{dom}}$, $b_{\text{dom}}$, $c_{\text{dom}}$, and $d_{\text{dom}}$ are calculated using Eq. (1) for each of the forecasts and SEDS$_{\text{dom}}$ determined. Similarly, using Eq. (3) and the method in sections 3c and 3d, the values of $D_{\text{dis}}$ in kilometers and QSDS are determined. Values of these properties for the example illustrated in Fig. 1 are provided in Table 4. As is shown in Fig. 10, SEDS$_{\text{dom}}$ can be plotted against either $D_{\text{dis}}$ or QSDS.

Looking at Fig. 10a, it is seen that all of the Boyden index forecasts for values above 94 are clustered with SEDS$_{\text{dom}}$ values less than or equal to 0.15, as well as $D_{\text{dis}}$ values of 50–150 km. This is a feature of the Boyden index; for as can be seen in Fig. 1, the Boyden index covers a large part of the domain and by design it will naturally produce a hedged forecast.

The grouping of $D_{\text{dis}}$ is an indication that the forecast is slowly varying in time and space, as the values suggest the model has captured most, if not all, of the places where lightning occurred, but has hedged the forecast somewhat by covering a large area around the observations. By changing to a Boyden index with the higher threshold, fewer points are forecast to be at risk of lightning, which has increased the SEDS$_{\text{dom}}$ score. However, the displacement distances have been allowed to have a much greater variability. For some forecasts, like that shown in Fig. 1 (circled in the plot), the displacement distance is much lower, meaning that the Boyden index with the 97 threshold has the features closer than the 94 threshold. However, three points using the 97 threshold have displacement distances that are greater than any of those with the 94 threshold. Six points in the 97 threshold group have displacement distances greater than the mean displacement distance produced by the 94 threshold. This result implies that while the number of points forecast as being at risk of lightning is smaller, the points can also be spaced farther away from where the lightning is observed. This result also suggests that moving from the 94 threshold [as originally suggested by Boyden (1963)] to the higher threshold does not improve the quality of the forecast.

Finally, the UKV-McCaul forecasts have a better SEDS$_{\text{dom}}$ score, as a result of lower points being highlighted as at risk of lightning. For two points, the SEDS$_{\text{dom}}$ score is very close to 1, indicating that the parameterization has produced forecasts of lightning over a similar-sized area to where it is observed, which is encouraging. The remaining points lie in the 0.5–0.9 range, suggesting that while better than the Boyden index, the parameterization is still overforecasting lightning. The $D_{\text{dis}}$ values are variable, but range between 25 and 125 km, with 17 out of 18 h (94%) being under 100 km and 12 out of 18 h of data (67%) being under 50 km. This suggests that the UKV parameterization produced features that were reasonably close in terms of their location to the observations. However, the forecasts rarely align with the exact position where they are observed in the model domain.

When examining Fig. 10, it should be noted that the SEDS$_{\text{dom}}$ score will be constrained in the limit [0, 1] for values derived using the variables in Eq. (1). Meanwhile, $D_{\text{dis}}$ can take values in the range [0, ∞]; the upper limit will ultimately be bounded by the domain size, but this will vary between models with different domain sizes. The values that QSDS takes are bounded by Eq. (4) to lie in the range [−1, 1]. Thus, Fig. 10b indicates a perfect forecast both in terms of distance and coverage at the point (1, 1) at the top-right corner of the plot.

In Fig. 10b, it is seen that the Boyden index with a threshold of 94 has QSDS values from 0.05 to 0.37 over the forecast period, meaning that it is performing slightly better than a completely hedged forecast in distance terms. The Boyden index with the threshold value of 97 has QSDS values ranging from 0.2 to 0.6 for 15 of 18 h of the forecast, which again suggests it is better than random in terms of distance for these hours. For 3 h, the index using the 97 threshold is below 0.1 and in two of those occasions the QSDS value is actually worse than random, indicating that the lightning forecasts are
spaced farther away than a random value would be. The Boyden index forecast from Fig. 1 (circled in Fig. 10b) performs better than the UKV-McCaul forecast. This agrees well with what can be seen by looking back at Figs. 1d–f, where the Boyden index with the 97 threshold appears to be closer to the observed lightning locations in the south-central and southeast areas of the domain, while the UKV-McCaul forecast has put the lighting too far to the west. However, the SEDS value for the Boyden forecast is lower than the UKV-McCaul forecast.

On the whole, the UKV-McCaul forecasts produce lightning features that are better than hedged forecasts in terms of distance (QSDS values range from 0.26 to 0.65). Of the values shown in Fig. 10b, 14 of 18 lie in the area where QSDS and SEDS_{dom} are both above 0.4. The mean and standard deviation of the values in Fig. 10 are presented in Table 5, where it can be seen that the UKV-McCaul forecast has a mean SEDS_{dom} of 0.721 and a mean QSDS of 0.479. Both of these values are higher than the comparable values for the Boyden index with either threshold chosen. The displacement distance for the UKV-McCaul forecasts is on average 51.94 km for this case, which is lower than the two values generated by the Boyden index.

The UKV-McCaul results can also be examined via the intensity score, but as the Boyden index does not forecast flash rate explicitly, it cannot be included in this analysis. Figure 11 shows the relationship between the three components of distance, coverage, and intensity for the same UKV-McCaul data displayed in Fig. 10. Looking at Figs. 11a and 11c, it can be seen that 9 out of 15 forecast hours lie above the asymptote in Fig. 9 and therefore are significantly overestimating the intensity of the lightning produced, even when the detector error is taken into consideration. This is also reflected in the intensity score in Figs. 11b and 11d, where 6 out of 15 points have an intensity score above 0.9. Most flash biases that cannot be attributed to detector error are a factor of 10–100 larger (so at least a factor of 2.5–25 even if detector error is considered). This suggests that the combination of the McCaul et al. (2009) parameterization with the UKV model microphysics is overdoing the lightning flash rate, although it is unclear whether this is due to the applicability of the McCaul et al. (2009) parameterization to lower flash rate storms found around the United Kingdom, or whether some error in the UKV model (e.g., too large an ice water content) is causing the bias; further work is required to investigate this possibility properly.

b. June 2016

In addition to the single case illustrated above, the UKV model forecasts for the month of June 2016 can be analyzed to ensure that the verification technique works over a range of data and not just for a single case with intense lightning. Data were available from four model runs each day, starting at 0300, 0900, 1500, and 2100 UTC. Each forecast was run out for 36 h, and data from 3 h into each model run were considered, to avoid any issues with spinup. This produces a total of 3960 hour-long samples of model forecasts and observed lightning data, of which 1005 samples had both model lightning forecasts as well as lightning observed by ATDnet during that hour.

Summary results from the 1005 hour-long samples in June 2016 are plotted in Fig. 12. The verification of the
UKV in terms of distance is shown in Fig. 12a for QSDS and Fig. 12b for $D_{\text{dis}}$. Values of QSDS are mostly positive, indicating that the UKV is forecasting lightning location better than a hedged forecast. The peak of the QSDS distribution is around 0.6. The $D_{\text{dis}}$ distribution has a peak around 25 km, but with a strong positive skewness. The mean value of $D_{\text{dis}}$ is 73.3 km, with occasional values as high as 600 km, which probably account for the negative values of QSDS. On examining the data, it was found that 81.4% of $D_{\text{dis}}$ values were less than 100 km and 55.2% were less than 50 km.

Figure 12c shows the coverage score, SEDS$_{\text{dom}}$, with values ranging from 0.4 to 1.0, with a peak around 0.7, so the model is able to reproduce fairly accurately the amount of the domain that has lightning forecast. It is interesting to examine the intensity score in Fig. 12d, which peaks around 0.9, suggesting that nearly all forecasts overestimate the actual number of UKV lightning discharges. Figures 12e and 12f show scatterplots of flash bias and intensity score against SEDS$_{\text{dom}}$, respectively. From the data in Fig. 12f, it is reported that 82% of forecast hours have a flash intensity score (and bias) greater than the maximum value possible as a result of detection errors within the ATDnet system. Similarly, the mean flash bias is recorded in Fig. 12e to be 25.56. Dividing this by the asymptote in Fig. 9a shows that the UKV-McCaul simulation overforecasts the intensity of lightning by a factor of 6.13 on average. As mentioned earlier in section 4a, it is not clear as to whether this is because of the McCaul et al. (2009) parameterization itself or another error in the UKV model system.

Finally, it is interesting to see the relationship between the modeled flash bias and intensity score when plotted against the SEDS$_{\text{dom}}$ score for coverage, as shown in

![Figure 11](image-url)
When the flash bias and intensity scores are high, suggesting significant overforecasting, the corresponding SEDS\textsubscript{dom} score is low, suggesting poorer coverage. This suggests that intense storms in the UKV model are too large both in terms of their flash rate and horizontal extent where lightning is forecast. Particularly in Fig. 12f, there appears to be a fairly narrow band of values that the model forecast can take, and it seems like the model is incapable of producing storm systems that have the structure correct, but with an intensity that is too large, or producing the correct intensity, but with a low SEDS\textsubscript{dom} (structure) score. This needs to be examined in more detail in a future study.

5. Discussion and applications

This manuscript has introduced a new method for verifying lightning forecasts in terms of three properties: coverage of the domain, distance to the observations, and intensity of the lightning forecast. The technique is illustrated using a single case and a month-long sample. While these results give an indication of the UKV-McCaul scheme’s performance, this study is not intended to form a long-term evaluation of the model, which is expected to follow in a future paper.

A new distance skill score, the quasi-symmetric distance score, is introduced. This will illustrate whether the forecast performs better than a completely hedged forecast of lightning (100% domain coverage). When plotted against the skill scores for coverage [SEDS\textsubscript{dom} after Hogan et al. (2009)] and the score for lightning intensity, it is possible to see graphically whether a lightning forecast is better than random in terms of both distance and coverage and how biased the forecast is. Perfect forecasts are defined as occurring when QS\textsubscript{DS} and SEDS\textsubscript{dom} are equal to 1.0 and the intensity score \( I \) is equal to zero.

The UKV-McCaul forecasts perform better than the large-scale Boyden index for a single event in terms of both the coverage and the distance between features,
with the UKV-McCaul forecasts having a mean $D_{\text{dis}}$ of 51.94 km for this case. The fact that the UKV-McCaul forecasts perform better is not surprising; the large-scale indices tend to “hedge” forecasts by overpredicting the area at risk of lightning.

Using a month-long dataset, the UKV-McCaul forecasts are shown to perform reasonably well in terms of both distance and coverage, with $D_{\text{dis}}$ averaging 73.2 km, with 81.3% of $D_{\text{dis}}$ values being less than 100 km. However, the UKV-McCaul simulations overforecast lightning intensity by at least a factor of 6 when observational detection efficiencies are considered. By plotting the scores graphically, as in Fig. 11, it is possible to see graphically the coverage, distance, and intensity scores in a single plot. Using this technique, it is now possible to evaluate model systems coupled with different parameterizations in the literature such as McCaul et al. (2009), Dahl et al. (2011a,b), Lynn et al. (2012), Fierro et al. (2013), and others to compare their relative performance. With a verification technique now established, further study can evaluate the performance of the UKV model over a longer time period with additional statistical analysis that has not been possible in this manuscript.

Acknowledgments. Thanks go to the following Met Office staff for useful discussions and comments on earlier versions of this manuscript: Paul Field, Philip Gill, Nigel Roberts, Rebecca Stretton, and Simon Vosper. The author would also like to thank Robin Hogan (ECMWF/University of Reading) for assistance with the error bars on the various skill scores in the manuscript and also the three anonymous reviewers for their helpful suggestions, which led to improvements in the manuscript.

APPENDIX A

Model Data

As part of the operational NWP suite, the Met Office runs its Unified Model (MetUM) as a U.K. variable resolution (UKV) model configuration over the British Isles. This configuration is mostly as described in Tang et al. (2013). Boundary conditions are taken from the operational global model. The microphysics in the MetUM is a single-moment bulk scheme based on Wilson and Ballard (1999), with the addition of prognostic rain and graupel (Forbes and Halliwell 2003) and a simple aerosol–cloud interaction to predict the cloud drop number (Wilkinson et al. 2013a). The ice particle size distribution has changed from the formulation given by Wilson and Ballard (1999) to a generic distribution developed by Field et al. (2007). The lightning parameterization used is McCaul et al. (2009), which produces a flash rate $F$ as

\begin{equation}
F = 0.95F_1 + 0.05F_2, \quad (A1)
\end{equation}

\begin{equation}
F_1 = 0.042[w_{q_g}(-15^\circ C)], \quad (A2)
\end{equation}

\begin{equation}
F_2 = 0.2 \left\{ \int [\rho(q_g + q_s + q_i)]dz \right\}. \quad (A3)
\end{equation}

Here, $w_{q_g}(-15^\circ C)$ represents the flux of graupel (vertical velocity multiplied by graupel mixing ratio) at the $-15^\circ C$ level in the storm and $\int [\rho(q_g + q_s + q_i)]dz$ is the total column ice water path (integral of air density multiplied by the sum of graupel, snow, and ice mixing ratios). Coefficients are taken as expressed in McCaul et al. (2009), with no attempt to tune the flash rates to the ATDnet data. Lightning flashes are discharged on each model time step where the product of the flash rate and the time step length (currently 50 s) is greater than 1. Lightning is deemed to be forecast in any grid column where one or more flashes is accumulated during the hour of interest. The Boyden index is defined by Boyden (1963) as

\begin{equation}
BI = Z_{700-1000} - T_{700} - 200, \quad (A4)
\end{equation}

where $Z_{700-1000}$ is the thickness of the layer between the 700- and 1000-hPa levels in the atmosphere, in decimeters and $T_{700}$ is the temperature at 700 hPa in degrees Celsius. Boyden (1963) found that sferics were almost entirely contained within the BI = 94 contour and that there was a sharp increase in the number of stations reporting thunderstorms when the value of BI rose from 93 to 94. Hence, the suggestion of the 94 threshold to forecast thunderstorm activity. The analysis for all forecasts has been restricted to the central domain of the UKV model, where the resolution is a fixed 1.5-km grid.

APPENDIX B

Operational Verification Procedure

a. Verification method

The following nine-step procedure can be used to generate operational verification products from the methodology described in this paper.

1) Grid the observed lightning strike locations onto the model grid.
2) From the model and observation data, generate a binary mask of lightning forecasts and locations, similar to Figs. 1d–f.
3) Add up the total points in each mask to get $n_a$ and $n_d$. Obtain $a_{dom}$, $b_{dom}$, $c_{dom}$, and $d_{dom}$ from Eq. (1). Calculate $SEDS_{dom}$ as described in the next section of this appendix, below.
4) Calculate $D_{a-f}$ and $D_{f-a}$ as shown in Fig. 7.
5) Determine $D_{dis}$ from Eq. (3).
6) Repeat steps 4 and 5 with a lightning forecast mask that covers 100% of the domain (every model grid box) to obtain $D_h$ (where $D_h = D_{dis}$ for this mask).
7) Produce four further masks that have a single lightning forecast point in them: one mask for each of the four corners of the domain. Repeat steps 4 and 5 for each mask to obtain four distances. The maximum of these four distances is $D_m$.
8) Calculate QSDS using Eq. (4) and the results obtained from steps 4–7.
9) Calculate the flash bias and intensity score using Eqs. (5) and (6).

The results of this analysis can be plotted in a fashion similar to Fig. 11.

b. Calculation of the $SEDS_{dom}$ skill score from Eq. (1)

Hogan et al. (2009) give the SEDS equation as

$$SEDS_{dom} = \frac{\ln((a+b)/n) + \ln((a+c)/n) - 1}{\ln(a/n)} \quad (B1)$$

This can be calculated from any contingency table derived from Eq. (1) by substituting $a_{dom}$ for $a$, $b_{dom}$ for $b$, $c_{dom}$ for $c$, and $d_{dom}$ for $d$ in Eq. (B1). The only remaining input to Eq. (B1) is $n$, which can be determined as $a_{dom} + b_{dom} + c_{dom} + d_{dom}$.

REFERENCES


Mäkelä, A., E. Saltikoff, J. Julkunen, I. Juga, E. Gregow, and S. Niemelä, 2013: Cold-season thunderstorms in Finland and


