 Q. J. R. Meteorol. Soc. 141: 1345–1353, April 2015 B DOI:10.1002/qj.2444

RMetS
Royal Meteorological Society

Evaluation of boundary-layer type in a weather forecast model utilizing long-term Doppler lidar observations

N. J. Harvey,* R. J. Hogan and H. F. Dacre
Department of Meteorology, University of Reading, UK

*Correspondence to: N. J. Harvey, Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK. E-mail: n.j.harvey@reading.ac.uk

Many studies evaluating model boundary-layer schemes focus on either near-surface parameters or short-term observational campaigns. This reflects the observational datasets that are widely available for use in model evaluation. In this article, we show how surface and long-term Doppler lidar observations, combined in such a way as to match model representation of the boundary layer as closely as possible, can be used to evaluate the skill of boundary-layer forecasts. We use a two-year observational dataset from a rural site in the UK to evaluate a climatology of boundary-layer type forecast by the UK Met Office Unified Model. In addition, we demonstrate the use of a binary skill score (Symmetric Extremal Dependence Index, SEDI) to investigate the dependence of forecast skill on season, horizontal resolution and forecast lead time. A clear diurnal and seasonal cycle can be seen in the climatology of both model and observations, with the main discrepancies being the model overpredicting cumulus-capped and decoupled stratocumulus-capped boundary layers and underpredicting well-mixed boundary layers. Using the SEDI skill score, the model is most skilful at predicting the surface stability. The skill of the model in predicting cumulus-capped and stratocumulus-capped stable boundary-layer forecasts is low, but greater than a 24 h persistence forecast. In contrast, the prediction of decoupled boundary layers and boundary layers with multiple cloud layers is lower than persistence. This process-based evaluation approach has the potential to be applied to other boundary-layer parametrization schemes with similar decision structures.

Key Words: stable boundary layer; decoupled boundary layer; stratocumulus; SEDI; skill scores; process evaluation

Received 12 February 2014; Revised 5 August 2014; Accepted 1 September 2014; Published online in Wiley Online Library 11 November 2014

1. Introduction

Climate models vary substantially in their predictions of boundary-layer clouds in a warmer climate. This leads to an uncertainty in radiative feedback and is one of the largest sources of uncertainty in climate prediction (Bony *et al.*, 2006; Webb *et al.*, 2006). For example, Bony and Dufresne (2005) have shown that the climate models with the largest climate sensitivity are those that have the largest changes in boundary-layer cloud in their future climate.

On a more local scale, the boundary-layer parametrization scheme used in a given numerical weather prediction model can affect the forecasts of weather phenomenon such as tornadoes (Stensrud and Weiss, 2002), hurricanes (Powell, 1980) and convective clouds (Zampieri *et al.*, 2005). Even within a single scheme, small differences in parameter values or initial conditions can cause forecasts to change dramatically, for instance changing from clear sky to overcast conditions (Martin *et al.*, 2000). Such changes have large impacts on surface temperatures and also feedback on the timing and location of deep convection (Baldauf *et al.*, 2011). Accurate near-surface temperature forecasts are important for a range of users, including electricity companies, as demand for electricity varies with temperature, and local road authorities, who are concerned with values of near-surface temperature relative to a threshold below which roads should be treated to prevent ice formation. Therefore, there is a strong need for accurate and comprehensive methods for the evaluation of boundary-layer schemes, from both climate and weather prediction perspectives.

There are many different boundary-layer parametrization schemes used in numerical weather prediction and climate models (e.g. schemes based on a first-order closure with local or non-local diffusivities and schemes based on the prognostic turbulent kinetic energy method). There have been a number of attempts to evaluate these schemes by comparing their output with observations in case studies (Beesley *et al.*, 2000; Betts and Jakob, 2002; Zhang and Zheng, 2004; Cuxart *et al.*, 2006; Hu *et al.*, 2010; Shin and Hong, 2011; Svensson *et al.*, 2011; Xie *et al.*, 2012). However, these are all based on short-term observational campaigns. In addition, they typically focus on only a few variables such as 2 m temperature, 10 m winds and boundary-layer height. The studies of Sengupta *et al.* (2004) and Barrett *et al.* (2009) go further and consider the occurrence and distribution of boundary-layer clouds, but to date there has been no systematic evaluation of boundary-layer

schemes based on surface and above-surface turbulent mixing and cloud type made throughout the depth of the boundary layer.

The new dataset of observed boundary-layer type derived by Harvey *et al.* (2013) provides an opportunity to perform such an evaluation. Surface and above-surface observations are analyzed in such a way as to diagnose boundary-layer types that match the categories used in models as closely as possible, making it possible to evaluate boundary-layer parametrizations. In addition the method is based solely on ground-based Doppler lidar and sonic anemometer data, which are routinely collected at various locations worldwide, and therefore provides a viable method for performing long-term boundary-layer scheme evaluations over different sites. This dataset could also be used in many different ways to characterize other aspects of the boundary layer, such as cloud cover and the structure of turbulence; however, here we restrict our attention to evaluating one particular boundary-layer parametrization scheme.

In this study, two years of data from the Chilbolton Facility for Atmospheric and Radio Research (CFARR), UK, are used to provide such an evaluation of the boundary-layer scheme in the UK Met Office Unified Model (UM). This model has a boundary-layer parametrization scheme that makes explicit use of the concept of boundary-layer type: it uses model variables to diagnose discrete boundary-layer types, which are then used to determine the location and intensity of of the turbulent mixing to apply. In principle, this analysis could be extended to other atmospheric models that use binary decisions inside their boundary-layer parametrizations, since each combination of binary decisions can be interpreted as a boundary-layer type. The evaluation of parametrization schemes is an indispensable part of the development of prediction systems. In this article, we aim to design an evaluation scheme that can be used to quantify both skill and bias in model forecasts, with the intention that this scheme can aid model development and lead eventually to improved forecasts.

This article is organized as follows. In section 2, the methodology of Harvey *et al.* (2013) is outlined briefly, followed by a description of the data used in this study. In section 3, a two-year climatology of boundary-layer type is presented for both the model and observations and in section 4 the Symmetric Extremal Dependence Index (SEDI) is then used to evaluate the skill of both the 4 and 12 km resolution versions of the UM. This measure of skill is also used to assess the predictions of boundary-layer type as a function of forecast lead time and season.

2. Method

2.1. Observational data

Harvey *et al.* (2013) diagnose discrete boundary-layer types from observations according to an extension of the classifications used in the UM (Lock *et al.*, 2000). Thus the verification data are matched to the forecast data as closely as possible, making it easier to verify the model forecast and identify bias. Table 1 lists the seven UM boundary-layer types and their relation to the nine observational types of Harvey *et al.* (2013). The observational boundary-layer types are diagnosed using data from a vertically pointing ground-based Doppler lidar and a sonic anemometer, both located at the CFARR. The sonic anemometer is used to derive the surface sensible heat flux ($\overline{H}$). The Doppler lidar is used to infer the presence of one or more layers of boundary-layer cloud and the skewness (s) and variance (σ_w^2) of the vertical velocity throughout the depth of the boundary layer. Together, s and σ_w^2 provide information on the presence of turbulent mixing in the boundary-layer as well as its source (cloud-top or surface-driven convection).

Each decision in the algorithm incorporates observational uncertainties and, as such, results in a probability of occurrence for each of the nine boundary-layer types for each hour of available data. For a fair comparison with the deterministic hourly UM data, only the most probable boundary-layer type is considered and observational types Ia, Ib and Ic are combined into one type and IIIa and IIIb are combined into one type, as shown in Table 1. In addition, the observational diagnosis is unreliable during periods of precipitation. As a result, all hours where there is precipitation in either the observational data (from a collocated rain gauge) or the model forecast (defined as precipitation rate above 0.02 mm h^{-1}) are removed from the comparison. This removes approximately 20% of the data from the comparison.

Table 1. The UM boundary-layer types of Lock *et al.* (2000) (left column) and their relation to the nine observational boundary-layer types of Harvey *et al.* (2013) (right column).

UM type		Observational type	
I	Stable, possibly with non-turbulent cloud	Ia	Stable boundary layer, no cloud
		Ib	Stratus-topped boundary layer, no cumulus
		Ic	Forced cumulus under stratocumulus
II	Stratocumulus over a stable surface layer	II	Stratocumulus over a stable surface layer
III	Single mixed layer, possibly cloud-topped	IIIa	Single mixed layer, no cloud
		IIIb	Single stratocumulus-topped mixed layer
IV	Decoupled stratocumulus not over cumulus	IV	Decoupled stratocumulus not over cumulus
V	Decoupled stratocumulus over cumulus	V	Decoupled stratocumulus over cumulus
VI	Cumulus-capped layer	VI	Cumulus-capped layer
VII	Shear-dominated unstable layer	III	Type a or b, depending on the presence of cloud

2.2. Model data

The UM (version 5.2 onwards) solves non-hydrostatic, deep-atmosphere dynamics using a semi-implicit, semi-Lagrangian numerical scheme (Cullen *et al.*, 1997; Davies *et al.*, 2005). The model includes a comprehensive set of parametrizations, including schemes for the surface (Best *et al.*, 2011; Clark *et al.*, 2011; Essery *et al.*, 2001), boundary layer (Lock *et al.*, 2000), mixed-phase cloud microphysics (Wilson and Ballard, 1999) and radiation (Edwards and Slingo, 1996). The model also includes an option for convection parametrization (Gregory and Rowntree, 1990), which is used at all resolutions greater than 4 km, with additional downdraught and momentum-transport parametrizations. The model runs on a rotated latitude/longitude horizontal grid with Arakawa C staggering and a terrain-following hybrid height vertical coordinate with Charney–Philips staggering.

Operational forecasts from two versions of the UM are used in this study. The 4 km resolution version of the UM (UK4) is used for the main observational comparison presented in sections 3 and 4 and the North Atlantic European version of the UM (NAE) is used to investigate the effect of horizontal resolution on the boundary-layer type forecasts. The UK4 covers a domain slightly larger than the UK and has 70 levels in the vertical, 16 of which are in the lowest 1 km. The NAE covers a larger domain over the North Atlantic and Europe and has a 12 km horizontal resolution and a coarser vertical resolution of 38 levels, with 7 levels in the lowest 1 km. There are several other differences between the two models, most notably in the convection and data assimilation schemes.

Each forecast is 36 h long and these are initialized four times per day, at 0300, 0900, 1500 and 2100 UTC for the UK4 and 0000, 0600, 1200 and 1800 UTC for the NAE. UK4 data are available for the two-year period 1 September 2009–31 August 2011, whereas NAE data are available for the nine-month period 1 September

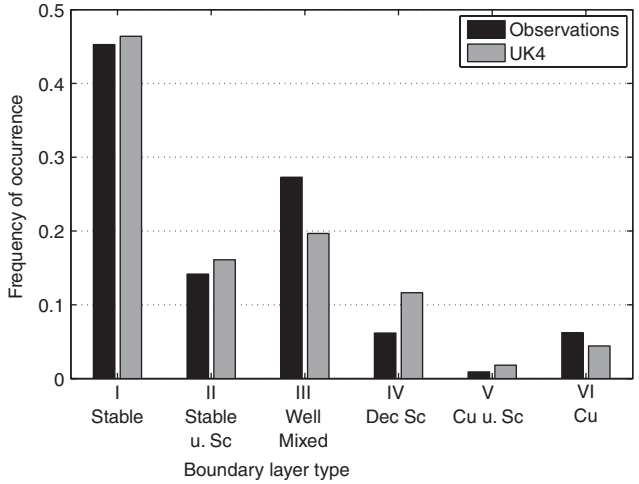

Figure 1. The frequency of occurrence of hourly boundary-layer types for UK4 and observations during the period 1 September 2009–31 August 2011.

2009–31 May 2010. These data are available for the closest nine grid points to CFARR.

The same boundary-layer scheme (Lock *et al.*, 2000; Lock and Edwards, 2011) is used in both the UK4 and NAE models. It categorizes the boundary layer at each grid point and time step into one of the seven different types summarized in Table 1, based on the surface stability, the vertical profile of potential temperature and the presence and type of cloud. The selected boundary-layer type then influences the form of the eddy diffusivity profile used to parametrize the turbulent fluxes within the boundary layer. A first-order K-closure scheme is used and the diffusivity can have contributions from both local and non-local terms, depending on the static stability. Additional diffusivity terms are included if boundary-layer cloud is present. For example, if cumulus cloud is diagnosed then it is assumed that there is turbulent mixing present from the surface up to the cumulus cloud base. In that case, the associated convection is treated entirely by the convection scheme. In stratocumulus-capped boundary layers, there is an additional source of mixing associated with turbulence driven from the cloud-top due to radiative cooling.

3. Evaluation of the model climatology

In this section the climatology of hourly boundary-layer types from the UK4 forecasts is compared with observations. Figure 1 shows the frequency of occurrence of hourly UK4 and observational boundary-layer types for the two years of available model data. For this comparison, only data from the grid point nearest to the CFARR is used. In addition, only data from the first 6 h of each forecast are used (the dependence on lead time is discussed in section 4.2.1).

There is good agreement between the frequency of occurrence of the stable boundary-layer types (I and II) in model and observations, with the model forecasting a slightly higher frequency of each. There is less agreement for the unstable types, with the model forecasting the decoupled stratocumulus types (IV and V) more frequently than occurs in the observations and the well-mixed (III) and cumulus (VI) types less frequently than in the observations. The ranking of the types in terms of frequency of occurrence is similar in the model and the observations, with only the order of the decoupled stratocumulus (IV) and cumulus (VI) types reversed.

Regarding the diurnal evolution of boundary-layer type, Figure 2 shows the frequency of occurrence of each type as a function of time of day for each season. A clear diurnal cycle is present in both model and observations, with the stable types dominating at night and the unstable types during daylight hours. Consistent with this, there is a seasonal cycle in the frequency of occurrence of each type, with higher occurrences of the unstable types during the summer months and higher occurrences of the stable types during the winter months. The transition between these two states occurs fairly rapidly around the time of sunrise, although this is blurred out in the seasonal averages of Figure 2.

The tendency for the model to favour the decoupled stratocumulus type (IV) over the well-mixed (III) type (Figure 1) is apparent in all seasons by the relative sizes of the bars. This discrepancy is largest during the morning daylight hours, particularly in spring and summer. Another feature to note is the difference between the occurrence of unstable types during night-time hours. The observations show the presence of unstable types during night-time during spring, summer and autumn with very little in winter, whereas the opposite is true of the model forecasts.

4. Evaluation of forecast skill

4.1. *Verification measures*

In this section, the skill of the model in predicting the correct boundary-layer type at the correct time is assessed using binary verification measures. These are calculated from joint histograms between the boundary-layer types of UK4 and the observations. Figure 3 shows the joint histogram for the hourly boundary-layer data from the entire two-year period. It shows the total number of occurrences of each combination of observed and modelled boundary-layer type.

If the model provided perfect forecasts, then all occurrences would lie on the diagonal in Figure 3. However, this is not the case here and there is a large spread. Multi-category verification measures do exist for assessing the skill of multi-category variables quantitatively; however, as our contingency table arises from a sequence of binary decisions for both model and observations, we will instead assess the skill using the more intuitive approach of applying binary verification measures to each decision in turn.

For binary events, the problem of forecast verification has a long history dating back to Finley (1884), who studied forecasts of tornadoes. More recently, similar techniques have been used, for instance by Barrett *et al.* (2009), Hogan *et al.* (2009b) and Mittermaier (2012), to evaluate forecasts of cloud properties.

In the present case, the joint histogram is split into 2×2 contingency tables by dividing it into four quadrants based on the decisions made in the diagnosis of boundary-layer type. It is common to refer to quadrants in a contingency table using the letters a, b, c and d, as shown in Table 2, and this convention is followed here. There are five decisions used in determining the observational boundary-layer type, as listed in Table 3. The third column of this table shows how each of the five decisions discriminates uniquely between the boundary-layer types and Figure 4 shows schematically how the histogram is split for each decision. For the 'surface layer stable' and 'cumulus present' decisions, the totals from each combination of events are summed to give the quadrant values. As an example, for the joint histogram shown in Figure 3, the 2×2 contingency table for the stability decision is as follows:

$$\begin{pmatrix} c & d \\ a & b \end{pmatrix} = \begin{pmatrix} 233 & 3596 \\ 5853 & 624 \end{pmatrix}, \tag{1}$$

meaning that of the 6086 observed stable types, 5853 occurred in the model and of the 4220 unstable types, 3596 occurred in the model.

There are many verification measures that can be used to assess the skill of a 2×2 contingency table (e.g. Wilks, 1995; Von Storch and Zwiers, 1999; Casati *et al.*, 2008; Hogan *et al.*, 2009b; Hogan and Mason, 2012). Here, the SEDI is used (Ferro and Stephenson, 2011). This measure was chosen as it has many desirable properties: it is equitable, meaning that all random forecasting systems will receive the same expected score, and it

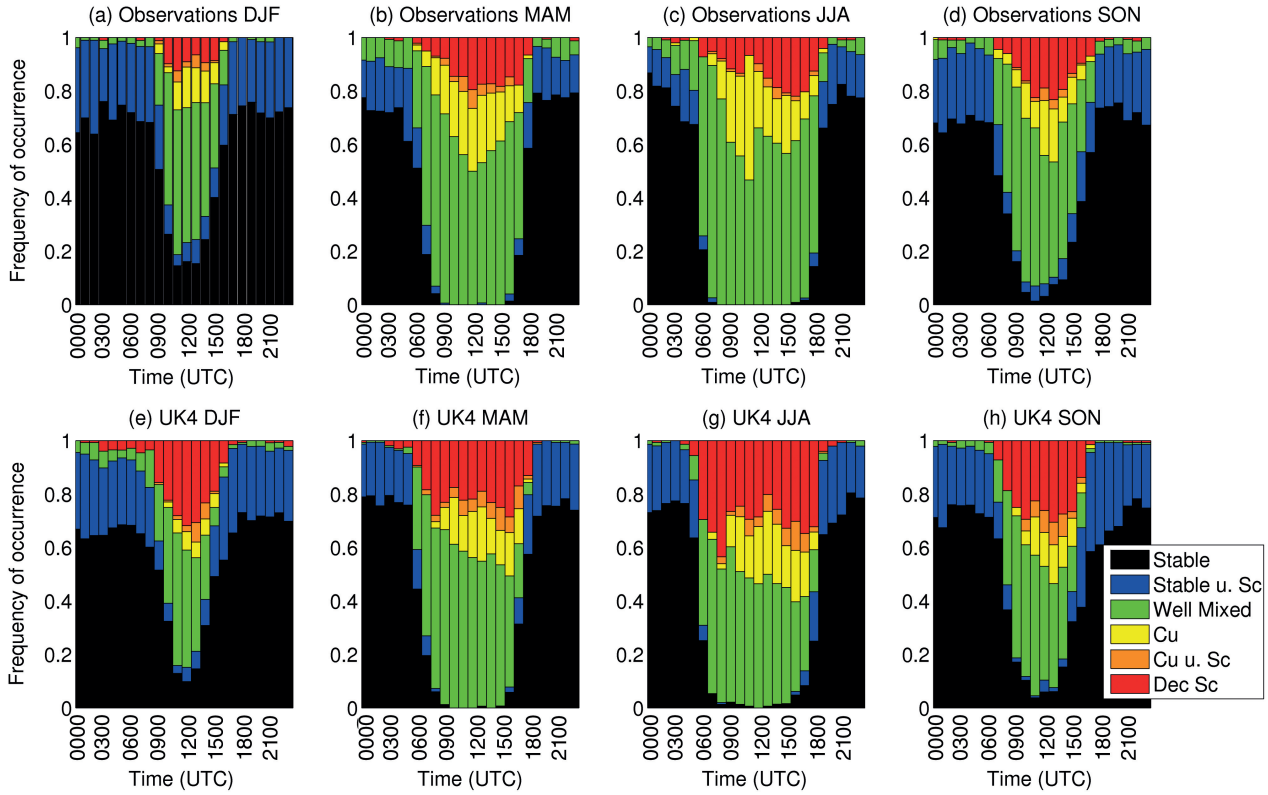

Figure 2. The frequency of occurrence of each type as a function of time of day for (a) and (e) winter, (b) and (f) spring, (c) and (g) summer and (d) and (h) autumn. Panels (a)–(d) show observational boundary-layer types and (e)–(h) show UK4 boundary-layer types.

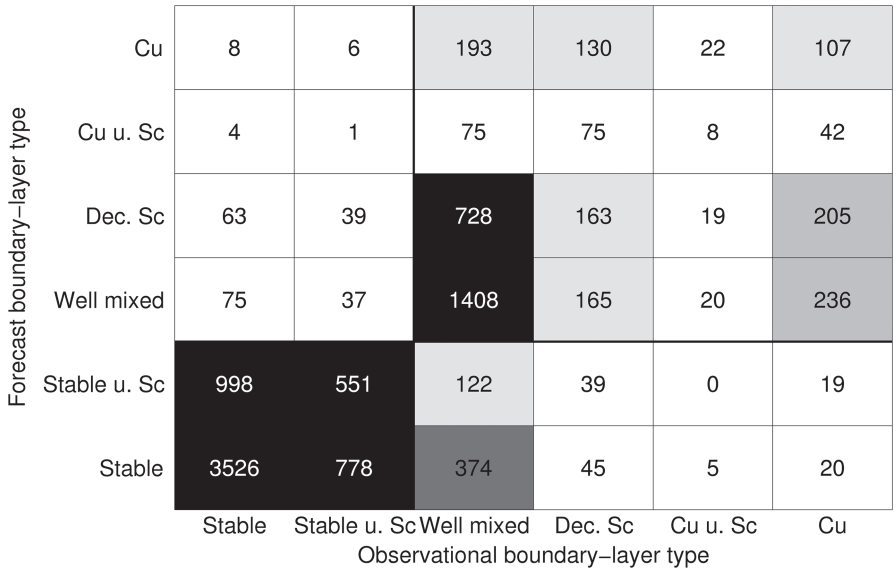

Figure 3. The joint histogram of hourly boundary-layer types for UK4 and observations during the period 1 September 2009–31 August 2011. The darker shading indicates a larger number of events.

Table 2. The construction of a 2 × 2 contingency table.

	Event observed	
Event forecast	Yes	No
No	*c* (misses)	*d* (correct rejections)
Yes	*a* (hits)	*b* (false alarms)

is also difficult to hedge, meaning that it cannot be improved by issuing a forecast that is not the true judgement of the forecaster. In addition, many verification measures tend to give meaningless values for rare events, but SEDI is independent of the frequency of occurrence of an event and therefore can be used for both rare and overwhelmingly common events (which is required here for types V and I, respectively).

The SEDI skill score is defined as

$$\mathrm{SEDI} = \frac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)}, \qquad (2)$$

where H is the hit rate ($H = a/(a+c)$) and F is the false-alarm rate ($F = b/(b+d)$). A SEDI value of 1 indicates perfect forecast skill, whereas a value of 0 indicates no more skill than a random forecast.

4.2. *The SEDI skill score for the UK4 forecasts*

In this section, the SEDI skill score is used in both relative and absolute terms to judge the skill of forecasts relative to each other and relative to two baseline reference forecasts. The first is a persistence forecast for which the boundary-layer type at a

Table 3. Summary of decisions that are assessed using binary verification measures.

Decision	Description	Types	Forecast SEDI	Persistence SEDI
1	Surface layer stable?	I and II vs. III, IV, V and VI	0.938	0.903
2	Cumulus present given unstable surface layer?	V and VI vs. III and IV	0.184	0.108
3	Decoupled given cumulus is not present?	IV vs. III	0.152	0.299
4	More than one cloud layer given cumulus cloud present?	VI vs. V	−0.019	0.083
5	Stratocumulus present given surface layer is stable?	II vs. I	0.271	0.098

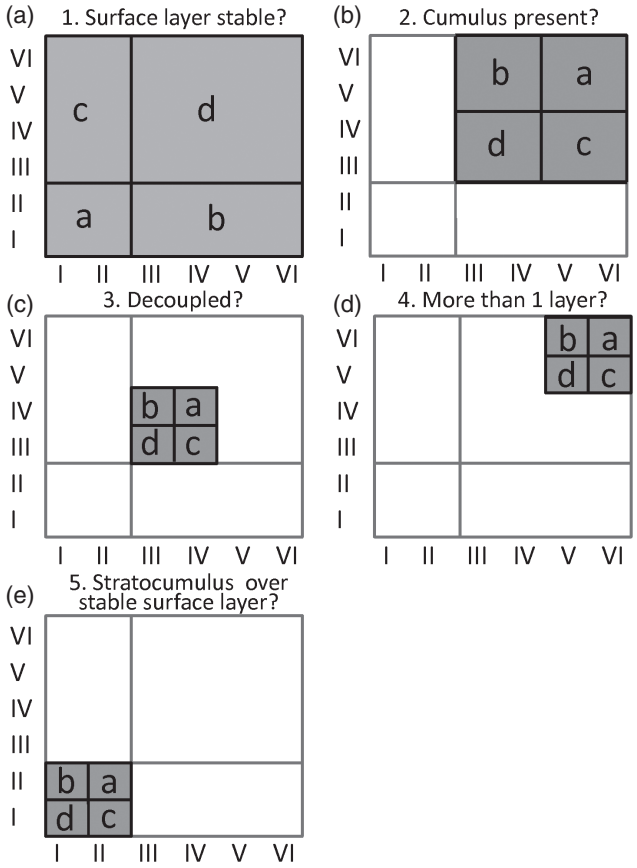

Figure 4. Schematic of how the joint histogram in Figure 3 is split into multiple 2 × 2 contingency tables corresponding to each decision. The number refers to the decision being considered (as in Table 3). The abscissa refers to the observed boundary-layer type and the ordinate to the modelled type.

given hour of a given day is forecast to be the boundary-layer type of the same hour on the previous day and the second is a hypothetical random forecast for which the SEDI skill score is zero. The SEDI skill scores for the full two-year period, using data from only the first 6 h of each forecast (i.e. the same data as discussed in section 3), are shown in Table 3. These are briefly discussed before considering the impact of lead time, season and model resolution.

The highest value of skill by far is for the stability decision (0.938), which may be due to the presence of a strong diurnal cycle. For this decision, the UK4 forecast skill is greater than the skill from persistence. The cumulus and stable stratocumulus decisions have lower forecast skill than the stability decision (0.184 and 0.271, respectively) and again the UK4 forecast skill is greater than the skill from persistence. In contrast, the decoupled and layers decisions have SEDI values lower in the UK4 forecasts than the persistence forecast (0.152 and −0.019, respectively) and, further, the layers decision has a slightly negative SEDI value, which is worse than that expected from a random forecast. The size of the error bars on these values due to sampling is discussed in section 4.2.1. The sharp decrease in SEDI between the stability and cloud-related decisions is probably due to the fact that it is fundamentally more difficult to predict cloud-related variables, as they are sensitive to subtle changes in the vertical temperature structure. This hypothesis is supported by Hogan *et al.* (2009b), who found that the NAE model systematically underpredicts cloud fractions greater than 5% in the lowest 5 km of the atmosphere.

The sensitivity of the SEDI skill scores to the choice of model grid point used has been found to be small. In particular, the values in Table 3 are very similar if, instead of using the nearest model grid point to the CFARR for the observational comparison, the most common boundary-layer type of the nearest nine grid points is used.

4.2.1. *Dependence of skill on forecast lead time*

To test whether the skill of the UK4 forecasts varies with lead time, the SEDI has been calculated for all forecast lead times grouped into 6 h periods. These are: 0–5, 6–11, 12–17, 18–23, 24–29 and 30–36 h for the two-year period (1 September 2009–31 August 2011).

Figure 5 shows the evolution of the SEDI values with lead time. The plots also show error bars for each SEDI value, which are based on the following formula, as presented in Hogan and Mason (2012):

$$S_{\mathrm{err}}^2 = \frac{S_H^2\left[\frac{\mathrm{SEDI}(1-2H)+1}{H(1-H)}\right]^2 + S_F^2\left[\frac{\mathrm{SEDI}(1-2F)+1}{F(1-F)}\right]^2}{\left[\ln F + \ln H + \ln(1-H) + \ln(1-F)\right]^2}, \qquad (3)$$

where $S_H^2 = H(1-H)/(a+c)$ and $S_F^2 = F(1-F)/(b+d)$ are the error variances of H and F. However, this formula assumes that each event in the contingency table is independent. In the case of boundary-layer type diagnosis, this is unlikely to be true since, particularly at night time, there are prolonged periods (i.e. several consecutive hours) with the same type present. To take account of this, the number of independent events for each type is estimated by counting the number of times that there is a transition to that boundary-layer type. For example, the sequence I I I I I would be one event for type I, whereas I V V II II would be three events, one each for types I, II and V. The contingency table coefficients are scaled by the fraction of independent events over total events and these scaled coefficients are then used to calculate the SEDI error variance of Eq. (3).

Figure 5 shows that none of the decisions has a significant increase or decrease in skill with lead time. This contrasts with the behaviour found by Hogan *et al.* (2009a) for cloud occurrence (at all levels) in a similar model, where skill dropped significantly during the 36 h forecast period. This may be because their short lead-time skill scores were higher than the skill scores here, with the exception of the stability decision, for which the skill is aided by the strong dependence on the diurnal cycle. In addition, all hourly periods that contain either observed or modelled precipitation have been removed (see section 2.1). Therefore it is possible that some large-scale weather events, the forecasts of which tend to have a strong dependence on lead time (for instance the passing of a front), have been neglected, thus skewing the results.

The error bars for each lead time in Figure 5 are generally small. One exception is for the layers decision; this is due to the relatively small number of samples. There is therefore no evidence that the SEDI is negative, i.e. that the UK4 forecasts are worse than a random forecast.

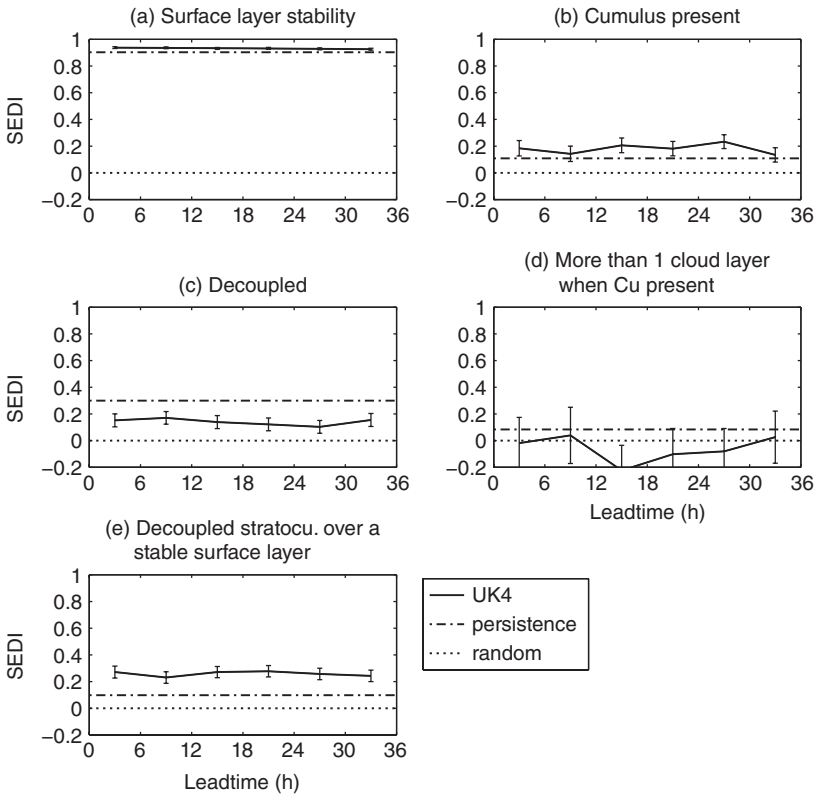

Figure 5. The dependence of skill on forecast lead time. The panels show the skill for (a) the stability decision, (b) the cumulus decision, (c) the decoupled decision, (d) the layers decision and (e) the stable stratocumulus decision. The lines indicate the SEDI values for (solid) UK4 and (dot–dashed) persistence forecasts, while the dotted line indicates the expected SEDI values for a random forecast. The error bars are calculated as described in the text.

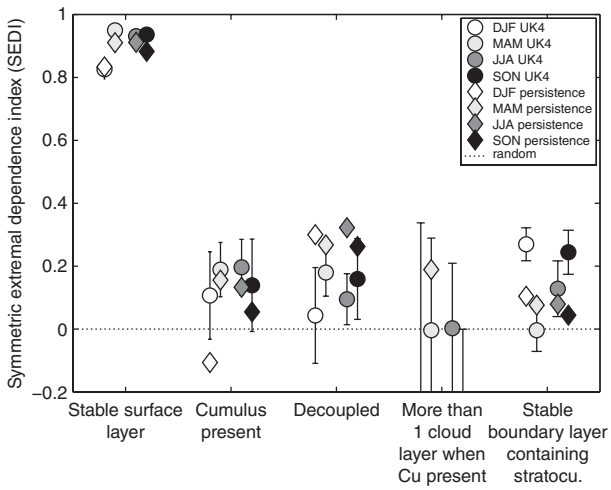

Figure 6. The dependence of skill on season. The circles indicate UK4 forecasts and the diamonds indicate the persistence forecast. The shading indicates the seasons as shown in the legend. The dotted line indicates the expected SEDI values for a random forecast and the error bars on the UK4 SEDI values are calculated as described in the text.

4.2.2. *Dependence of skill on season*

To assess the dependence of the forecast skill on the time of year, Figure 6 summarizes the SEDI score for each decision for each season. In this plot, data from all six forecast lead-time periods have been combined. This is to improve the statistics by increasing the number of samples. This is justified in this case, since, as shown in section 4.2.1, there is very little variation of the forecast skill with lead time, meaning that each forecast can be treated as an alternative realization of the same period. The error bars in Figure 6 are estimated from the variations between the forecasts of different lead times, σ, in the following way:

$$\mathrm{CI} = \pm 1.96 \frac{\sigma}{\sqrt{N-2}}, \qquad (4)$$

where $N = 6$ is the number of forecast lead times used. The scaling of 1.96 corresponds to a confidence interval of 95% assuming a normal distribution.

Figure 6 shows that, in winter, the decisions that discriminate between the unstable types are predicted with less skill than in all other seasons. The stability decision also has the lowest skill during winter. The reason for this drop in skill in winter may be related to the fact that it is seen observationally that during winter the sensible heat flux can remain close to zero throughout the day. This can make it difficult for the model to predict when the transition from stable to unstable occurs, thus reducing the skill. Spring has the highest SEDI scores for stability and decoupled. Summer has the highest score for the cumulus and layers decisions. The prediction of more than one cloud layer when cumulus is present has little or no skill in all seasons.

4.3. *Dependence of skill on model resolution*

The effect of model resolution is investigated by using another model in the operational suite of the UK Met Office. Here the NAE is used, as described in section 2.2. Due to the availability of the NAE data, the shorter period of 1 September 2009–31 May 2010 is used for this analysis.

Figure 7 shows the frequency of occurrence of boundary-layer types for the observations, the UK4 model and the NAE model for the period 1 September 2009–31 May 2010. As before, the observations and the UK4 model agree reasonably well for stable boundary-layer types (I and II), but there is a discrepancy with the NAE model. The NAE model has a much greater frequency of occurrence of stable boundary-layer type I than the other data sets. This is compensated by a lower frequency of occurrence of the stable under stratocumulus type (II).

The NAE model also diagnoses both of the cumulus types (V and VI) much less frequently than the UK4 model. The decrease in occurrence of cumulus types in the NAE model is compensated by an increase in the number of well-mixed boundary-layer types diagnosed. The occurrence of decoupled stratocumulus cloud is very similar in all data sets.

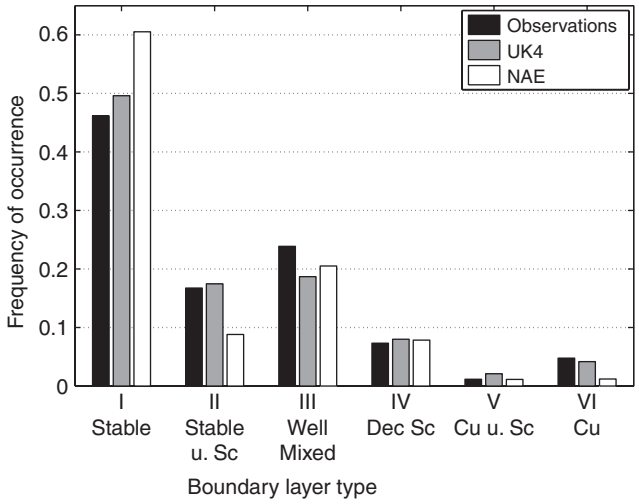

Figure 7. The frequency of occurrence of hourly boundary-layer types for UK4 and NAE models and observations during the period 1 September 2009–31 May 2010.

The SEDI score has been calculated for each of the decisions described in section 4.2.1. As in section 4.2.2, data from all of the forecast lead-time periods has been used. Figure 8 shows the SEDI skill score for each decision in turn. To aid comparison with the 12 km grid of the NAE model, rather than using the nearest grid point to the CFARR in the UK4 grid the most commonly occurring boundary-layer type in the nearest nine grid points is used instead.

Within the 95% confidence intervals (calculated as in Eq. (4)), there is no significant difference in skill between the UK4 and NAE models for any of the decisions. This is supported by Mittermaier (2012), who could not draw any conclusion about the impact of horizontal resolution on the Symmetric Extreme Dependency Score of cloud-base height and total cloud amount in the NAE, UK4 and UKV (a 1.5 km resolution version of the Met Office Unified Model). Conversely, Lean *et al.* (2008) found that increasing horizontal resolution increased the Fractions Skill Score of precipitation events over the UK for a forecast lead time of 6 h. Small differences in skill as model resolution increases were also seen in the NCEP Eta model by Mass *et al.* (2002). They found that more realistic mesoscale structures and evolution were seen as the resolution increased from 36 to 12 km. This gave improvements in precipitation amount, 10 m winds, 2 m temperature and surface pressure. However, there was not much impact on skill as the resolution was further increased from 12 to 4 km.

Also shown in Figure 8 is the SEDI score for the presence of low cloud (below 3 km) in the UK4 and NAE models. This score is significantly higher than all of the cloud decisions related to boundary-layer type. This shows that the model does a reasonable job of predicting low cloud despite incorrectly predicting the cloud type. This is because in the model cloud presence is controlled by the large-scale humidity field, rather than more subtle features in the thermodynamic profile.

5. Conclusions

In this article, we have demonstrated how numerical weather prediction boundary-layer parametrization schemes may be verified utilizing continuous Doppler lidar and sonic anemometer observations. Designing the observational verification data to match the model forecast dataset closely allows the boundary-layer parametrization scheme to be verified in a more direct way, making it easier to identify model bias and areas for model improvement.

Firstly, the climatology of boundary-layer type has been compared. In general, the seasonal and diurnal cycles seen in the model and observations are not dissimilar to the most common boundary-layer type in both model and observations, being stable followed by well mixed. However, there is a tendency for the model to diagnose decoupled stratocumulus-capped boundary-layers over well-mixed boundary layers, particularly during the morning hours in spring and summer. In addition, the model underpredicts the presence of unstable boundary layers during the night-time in spring, summer and autumn but overpredicts during the winter.

The ability of model forecasts to predict boundary-layer type at the correct time has been evaluated in an absolute sense relative to persistence and random forecasts. Overall, there is good skill when predicting stable and unstable boundary-layer types, due to the strong diurnal cycle. Consistent with previous studies, it was shown that the skill of predicting cloud presence is much greater than persistence. However, when considering different cloud types the skill reduces. The skill of the model in predicting cumulus-capped and stratocumulus-capped stable boundary layers is low but greater than persistence. In contrast,

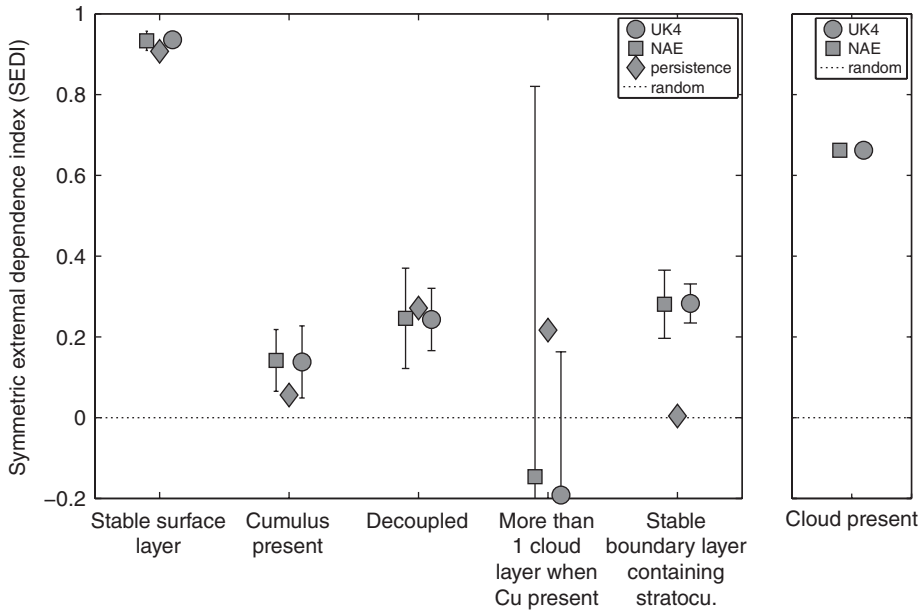

Figure 8. Summary of the SEDI score for each decision for the the modal UK4 boundary-layer type and NAE boundary-layer type closest to the CFARR, with 95% confidence intervals calculated using the forecast lead time data for the period 1 September 2009–31 May 2010. The right panel shows the SEDI score for the presence of cloud (>10%) below 3 km for the same period for the UK for the closest grid point to CFARR.

the prediction of decoupled boundary layers and boundary layers with multiple cloud layers is lower than persistence. This is likely due to the fact that the presence of cloud in the model depends on smoothly varying fields (e.g. temperature and humidity); however, cloud type in the model depends on the gradients in these fields, which are much more difficult to forecast.

The verification method described can also be used to judge the model skill in relative terms. Thus it is possible to determine how changes in the model resolution, lead time and seasonality affect the skill of the forecast. It was found that there is no significant impact of changing model resolution from 12 to 4 km. This is likely to be due to the fact that the boundary-layer scheme used at both 4 and 12 km resolution is the same and at 4 km the model is still not able to resolve turbulent processes within the boundary layer. It would be interesting to evaluate a model running at several hundred metres or better, where the largest eddies in the the boundary layer are resolved and there would be less dependence on the boundary-layer parametrization. No decrease in model skill was found with increasing lead time. However, it was found that decisions that discriminate between boundary-layer types in winter are predicted with less skill than in all other seasons.

An obvious further extension to this study would be to evaluate the model skill at a different site to see whether the model bias identified at the rural site is also present at other locations. It would also be interesting to compare the skill and climatology of boundary-layer type over an urban surface, which may exhibit different seasonal and diurnal evolution.

The UK Met Office is the only modelling centre to use the *Lock* boundary-layer scheme, but many other models have a similar tree of decisions, which is used to determine which parametrization schemes are applied, i.e. whether to apply a local or non-local mixing scheme. An example of this is the European Centre for Medium-range Weather Forecasts (ECMWF) model, which uses an eddy-diffusivity mass-flux framework (Kohler *et al.*, 2011). With this in mind, it would be possible to extend this type of comparison to models from other forecast centres around the world.

Acknowledgements

We thank Adrian Lock, Alan Grant, Stephen Belcher and Peter Jan van Leeuwen for useful discussions. The 1.5 μm Doppler lidar was acquired with NERC grant NE/C513569/1. The sonic anemometer was acquired with NERC grant NE/D005205/1. The UK4 and NAE data were supplied by the Met Office. The instruments at CFARR are operated and maintained by the Rutherton Appleton Laboratory.

References

Baldauf M, Seifert A, Förstner J, Majewski D, Raschendorfer M, Reinhardt T. 2011. Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Weather Rev.* **139**: 3887–3905.

Barrett AI, Hogan RJ, O'Connor EJ. 2009. Evaluating forecasts of the evolution of the cloudy boundary layer using diurnal composites of radar and lidar observations. *Geophys. Res. Lett.* **36**: L17811, doi: 10.1029/2009GL038919.

Beesley J, Bretherton C, Jakob C, Andreas E, Intrieri J, Uttal T. 2000. A comparison of cloud and boundary layer variables in the ECMWF forecast model with observations at surface heat budget of the Arctic ocean (SHEBA) ice camp. *J. Geophys. Res.* **105**: 12337–12349, doi: 10.1029/2000JD900079.

Best MJ, Pryor M, Clark DB, Rooney GG, Essery RLH, Ménard CB, Edwards JM, Hendry MA, Porson A, Gedney N, Mercado LM, Sitch S, Blyth E, Boucher O, Cox PM, Grimmond CSB, Harding RJ. 2011. The joint UK land environment simulator (JULES), model description Part 1: Energy and water fluxes. *Geosci. Model Dev.* **4**: 677–699.

Betts A, Jakob C. 2002. Evaluation of the diurnal cycle of precipitation, surface thermodynamics, and surface fluxes in the ECMWF model using LBA data. *J. Geophys. Res.* **107**: LBA 12-1–LBA 12-8, doi: 10.1029/2001JD000427.

Bony S, Dufresne JL. 2005. Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.* **32**: L20806, doi: 10.1029/2005GL023851.

Bony S, Colman R, Kattsov V, Allan R, Bretherton C, Dufresne J, Hall A, Hallegatte S, Holland M, Ingram W, Randall DA, Soden BJ, Tselioudis G, Webb MJ. 2006. How well do we understand and evaluate climate change feedback processes? *J. Clim.* **19**: 3445–3482.

Casati B, Wilson LJ, Stephenson DB, Nurmi P, Ghelli A, Pocernich M, Damrath U, Ebert EE, Brown BG, Mason S. 2008. Forecast verification: Current status and future directions. *Meteorol. Appl.* **15**: 3–18.

Clark DB, Mercado LM, Sitch S, Jones CD, Gedney N, Best MJ, Pryor M, Rooney GG, Essery RLH, Blyth E, Boucher O, Harding RJ, Huntingford C, Cox PM. 2011. The joint UK land environment simulator (JULES), model description. Part 2: Carbon fluxes and vegetation dynamics. *Geosci. Model Dev.* **4**: 701–722.

Cullen MJP, Davies T, Mawson MH, James JA, Coulter SC, Malcolm A. 1997. An overview of numerical methods for the next generation U.K. NWP and climate model. *Atmos. Ocean* **35**: 425–444.

Cuxart J, Holtslag A, Beare R, Bazile E, Beljaars A, Cheng A, Conangla L, Ek M, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld G-J, Svensson G, Taylor P, Weng W, Wunsch S, Xu K-M. 2006. Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorol.* **118**: 273–303.

Davies T, Cullen MJP, Malcolm AJ, Mawson MH, Staniforth A, White AA, Wood N. 2005. A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Q. J. R. Meteorol. Soc.* **131**: 1759–1782.

Edwards J, Slingo A. 1996. Studies with a flexible new radiation code. I: Choosing a configuration for a large-scale model. *Q. J. R. Meteorol. Soc.* **122**: 689–719.

Essery R, Best M, Cox P. 2001. 'MOSES 2.2 technical documentation'. Hadley Centre Technical report 30, Met Office, Exeter, UK.

Ferro CAT, Stephenson DB. 2011. Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting* **26**: 699–713.

Finley J. 1884. Tornado prediction. *Am. Meteorol. J.* **1**: 85–88.

Gregory D, Rowntree P. 1990. A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Weather Rev.* **118**: 1483–1506.

Harvey NJ, Hogan RJ, Dacre HF. 2013. A method to diagnose boundary-layer type using Doppler lidar. *Q. J. R. Meteorol. Soc.* **139**: 1681–1693, doi: 10.1002/qj.2068.

Hogan RJ, Mason IB. 2012. Deterministic forecasts of binary events. In *Forecast Verification a Practitioner's Guide in Atmosphere Science*, Jolliffe IT, Stephenson DB. (eds.): 31–59. Wiley-Blackwell: Chichester, UK.

Hogan RJ, Grant ALM, Illingworth AJ, Pearson GN, O'Connor EJ. 2009a. Vertical velocity variance and skewness in clear and cloud-topped boundary layers as revealed by Doppler lidar. *Q. J. R. Meteorol. Soc.* **135**: 635–643.

Hogan RJ, O'Connor EJ, Illingworth AJ. 2009b. Verification of cloud-fraction forecasts. *Q. J. R. Meteorol. Soc.* **135**: 1494–1511.

Hu X, Nielsen-Gammon J, Zhang F. 2010. Evaluation of three planetary boundary layer schemes in the WRF model. *J. Appl. Meteorol. Climatol.* **49**: 1831–1844.

Kohler M, Ahlgrimm M, Beljaars A. 2011. Unified treatment of dry convective and stratocumulus topped boundary layers in the ECMWF model. *Q. J. R. Meteorol. Soc.* **137**: 43–57.

Lean HW, Clark PA, Dixon M, Roberts NM, Fitch A, Forbes R, Halliwell C. 2008. Characteristics of high-resolution versions of the Met Office unified model for forecasting convection over the United Kingdom. *Mon. Weather Rev.* **136**: 3408–3424.

Lock A, Edwards J. 2011. 'The parameterization of boundary layer processes', UM Documentation Paper 24, Met Office, Exeter, UK.

Lock AP, Brown AR, Bush MR, Martin GM, Smith RNB. 2000. A new boundary-layer mixing scheme –1. Scheme description and single-column model tests. *Mon. Weather. Rev.* **128**: 3187–3199.

Martin GM, Bush MR, Brown AR, Smith RNB. 2000. A new boundary-layer mixing scheme –2. Tests in climate and mesoscale models. *Mon. Weather Rev.* **128**: 3200–3217.

Mass C, Ovens D, Westrick K, Colle B. 2002. Does increasing horizontal resolution produce more skilful forecasts. *Bull. Am. Meteorol. Soc.* **83**: 407–430.

Mittermaier M. 2012. A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Q. J. R. Meteorol. Soc.* **138**: 1794–1807, doi: 10.1002/qj.1918.

Powell M. 1980. Evaluations of diagnostic marine boundary-layer models applied to hurricanes. *Mon. Weather Rev.* **108**: 757.

Sengupta M, Clothiaux E, Ackerman T. 2004. Climatology of warm boundary layer clouds at the ARM SGP site and their comparison to models. *J. Clim.* **17**: 4760–4782.

Shin H, Hong S. 2011. Intercomparison of planetary boundary-layer parametrizations in the WRF model for a single day from CASES-99. *Boundary-Layer Meteorol.* **139**: 261–281.

Stensrud D, Weiss S. 2002. Mesoscale model ensemble forecasts of the 3 May 1999 Tornado outbreak. *Weather and Forecasting* **17**: 526–543.

Svensson G, Holtslag AAM, Kumar V, Mauritsen T, Steeneveld G, Angevine WM, Bazile E, Beljaars A, de Bruijn E, Cheng A, Conangla L, Cuxart J, Ek M, Falk MJ, Freedman F, Kitagawa H, Larson VE, Lock A, Mailhot J, Masson V, Park S, Pleim J, Soderberg S, Weng W, Zampieri M. 2011. Evaluation of the diurnal cycle in the atmospheric boundary layer over land as represented by a variety of single-column models: The second GABLS experiment. *Boundary-Layer Meteorol.* **140**: 177–206.

Von Storch H, Zwiers FW. 1999. *Statistical Analysis in Climate Research*. Cambridge, UK.

Webb MJ, Senior CA, Sexton DMH, Ingram WJ, Williams KD, Ringer MA, McAvaney BJ, Colman R, Soden BJ, Gudgel R, Knutson T, Emori S, Ogura T, Tsushima Y, Andronova N, Li B, Musat I, Bony S, Taylor K. 2006. On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Clim. Dyn.* **27**: 7–38.

Wilks DA. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press: London.

Wilson DR, Ballard SP. 1999. A microphysically based precipitation scheme for the UK Meteorological Office unified model. *Q. J. R. Meteorol. Soc.* **125**: 1607–1636.

Xie B, Fung JCH, Chan A, Lau A. 2012. Evaluation of nonlocal and local planetary boundary layer schemes in the WRF model. *J. Geophys. Res.* **117**: D12103, doi: 10.1029/2011JD017080.

Zampieri M, Malguzzi P, Buzzi A. 2005. Sensitivity of quantitative precipitation forecasts to boundary layer parameterization: A flash flood case study in the western Mediterranean. *Nat. Hazards Earth Syst. Sci.* **5**: 603–612.

Zhang D, Zheng W. 2004. Diurnal cycles of surface winds and temperatures as simulated by five boundary layer parameterizations. *J. Appl. Meteorol.* **43**: 157–169.

