Performance of Hourly Operational Consensus Forecasts (OCFs) in the Australian Region

CHERMLEE ENGEL AND ELIZABETH EBERT

Bureau of Meteorology Research Centre, Melbourne, Victoria, Australia

(Manuscript received 13 November 2006, in final form 18 April 2007)

ABSTRACT

This paper presents an extension of the operational consensus forecast (OCF) method, which performs a statistical correction of model output at sites followed by weighted average consensus on a daily basis. Numerical weather prediction (NWP) model forecasts are received from international centers at various temporal resolutions. As such, in order to extend the OCF methodology to hourly temporal resolution, a method is described that blends multiple models regardless of their temporal resolution. The hourly OCF approach is used to generate forecasts of 2-m air temperature, dewpoint temperature, RH, mean sea level pressure derived from the barometric pressure at the station location (QNH), along with 10-m wind speed and direction for 283 Australian sites. In comparison to a finescale hourly regional model, the hourly OCF process results in reductions in average mean square error of 47% (air temperature), 40% (dewpoint temperature), 43% (RH), 29% (QNH), 42% (wind speed), and 11% (wind direction) during February–March with slightly higher reductions in May. As part of the development of the approach, the systematic and random natures of hourly NWP forecast errors are evaluated and found to vary with forecast hour, with a diurnal modulation overlaying the normal error growth with time. The site-based statistical correction of the model forecasts is found to include simple statistical downscaling. As such, the method is found to be most appropriate for meteorological variables that vary systematically with spatial resolution.

1. Introduction

Forecasters in Australian regional weather forecasting centers have a wealth of numerical model guidance available to them, including regional models produced in the Australian Bureau of Meteorology (the Bureau) and a subset of the guidance from global models produced in other international centers. When guidance is available from a number of different models, consensus-forecasting techniques, which combine the models, have been found to be more accurate on the average than techniques that try to predict the best “model of the day” (Hibon and Evgeniou 2005; Fritsch et al. 2000). Combining multiple individual forecasts to increase accuracy is an approach used in various fields from business to psychology (Clemen 1989). The operational consensus forecast (OCF) scheme developed at the Bureau combines multimodel guidance. Daily OCF has been shown to produce objective guidance for forecast fields such as maximum and minimum daily air temperatures that is competitive with subjective official forecasts (Woodcock and Engel 2005). A 6-month real-time trial showed reductions of mean absolute error (MAE) in minimum and maximum air temperature forecasts of about 40%, relative to component forecasts at days 0–2. MAEs were also lower at day 1 than were the matching official forecasts of maxima and minima by 8% and 10%, and the MAEs outperformed official subjective forecasts at over 71% and 75% of sites, respectively (Woodcock and Engel 2005). The official test comparison was conducted over 1 yr.

While daily OCF forecasts provide useful guidance for public weather forecasts, fire-weather and aviation forecasts require higher temporal resolution numerical guidance. Aviation forecasters use numerical guidance to assist in the production of Terminal Aerodrome Forecasts (TAFs) to the airline industry worldwide. TAFs provide localized forecast information with 3- or 6-h temporal resolutions out to 12 or 24 h, respectively.

DOI: 10.1175/2007WAF2006104.1

© 2007 American Meteorological Society
TAFs are produced at regular times every day for each aerodrome as well as on special request, with the base time of TAF issuance varying with station (Bureau of Meteorology 2005, chapter 7). Fire-weather forecasters use numerical guidance to assist in the production of public weather and spot fire forecasts. Routine fire-weather forecasts include predictions of maximum air temperature, along with dewpoint temperature (or RH) and mean wind speed at the time of maximum. In addition, fire-weather warnings are dependent also upon the timing of events such as significant wind direction and speed changes and indications of other events such as rainfall or lightning (Bureau of Meteorology 1997, chapter 5). Consensus techniques offer an opportunity to greatly enhance the quality of the numerical guidance made available to forecasters.

Due to communications and other costs, global numerical guidance from international centers is received routinely via the Global Telecommunications System (GTS) at temporal resolutions of 6 to 24 h with reduced spatial resolution. Hourly temporal resolution numerical guidance is currently only available to Bureau forecasters from three locally produced, nested regional models. To reduce the amount of disk space used for archiving purposes, some meteorological variables of the Australian models are made available every 3 h at reduced spatial resolution. In this paper, we describe a self-consistent hourly temporal resolution consensus product produced from a multimodel ensemble with a mix of 6-, 3-, and 1-h temporal resolutions. This “blended” hourly OCF scheme is then evaluated out to 42 h for 283 stations around Australia.

Throughout this paper, models are referenced by abbreviations defined in Table 1. Section 2 describes the observational and model data used throughout this report. Section 3 provides a description of the daily OCF methodology and highlights features relevant to the extension of the scheme to higher temporal resolution and to more complex meteorological variables. The performance of the bias correction and weighting parameters on an hourly basis is evaluated in section 4. Section 5 provides the justification for the inclusion of coarse international model forecasts within the site-based consensus. Section 6 provides a methodology for coping with the low temporal resolution of the international models. Finally, section 7 provides a summary discussion and section 8 details the conclusions.

### 2. Data

#### a. Numerical models

The numerical weather prediction (NWP) model guidance used in this study is available routinely within the Bureau. The forecasts include locally produced Local Area Prediction System (LAPS; Puri et al. 1998) model guidance at 37.5-, 12.5-, and 5.0-km resolution.
and lower spatiotemporal resolution fields from other operational centers (Table 1). Gridlines corresponding to three low-resolution international models are displayed in Fig. 1 for reference. For a description of the models, refer to the relevant official documentation. The suite of models uses different numerical formulations, physics packages, and initial conditions. Therefore, their forecasts and associated random errors can be considered somewhat independent.

In this study forecasts were derived from the first 42 h of 1200 UTC model issuances. The NWP fields utilized included 2-m air temperature, dewpoint temperature, and RH; 10-m meridional and zonal wind guidance; and barometric pressure. The mean sea level pressure derived from the barometric pressure at the station location (QNH) forecasts were derived using barometric pressure and topographic fields.

b. Observational data

This study utilizes site-specific Automatic Weather Station (AWS) measurements of 2-m air temperature and dewpoint temperature, 10-m wind direction and speed, and QNH from 283 stations. Relative humidity values were derived from air temperature and dewpoint temperature observations. Although spurious observations may affect the verification and bias corrections of particular weather stations, the sites used within this study are used routinely for aviation and fire-weather forecasting. As such, systematic observation errors are likely to have been reported and the number of one-off errors is likely to be small in comparison to the overall national sample size. Therefore, extra quality control was not performed on the observational data.

This study is based on a dataset covering two 60-day periods: from 15 January 2004 to 15 March 2004 and from 1 April 2004 to 31 May 2004, including only the days that all models were present. The first 30 days in each period were used only to determine bias correction and weighting parameters; the last 30 days were used for forecast evaluation as well. With some data missing due to archival failure, we had 24 forecast days over the February–March period and 25 forecast days during May. The datasets were chosen, restricted by availability, to cover two different weather periods.
During the April–May period there were generally 283 stations available for each variable except QNH, which had 262. The January–March period was similar apart from eight stations being unavailable during that period. A map of the station locations is shown in Fig. 1.

3. OCF methodology

a. Description

The OCF methodology of Woodcock and Engel (2005) is a simple statistical scheme, which takes a weighted average of bias-corrected component model forecasts on a site-by-site and day-by-day basis. The scheme is based upon the premise that each model-derived forecast \( f_i \) has three components: the true value \( o \), a systematic error component or bias \( b_i \) that can be approximated and removed, and a random error component \( e_i \) that can be minimized through compositing (\( i \) indicating each separate model). The success of the OCF scheme is based upon the estimation of bias and weighting parameters.

Bias and weighting parameters are based on a 30-day moving-bias window. Model biases \( b_i \) are approximated using the best easy systematic estimator (BES; Wonnacott and Wonnacott 1972, section 7.3) over the errors in the sample:

\[
\hat{b}_i = \text{BES} = (Q_1 + 2Q_2 + Q_3)/4, \quad (1)
\]

where \( Q_1 \), \( Q_2 \), and \( Q_3 \) are the error samples first, second, and third quartiles, respectively. This is more robust than a simple arithmetic mean. Normalized weighting parameters \( \hat{w}_i \) are calculated by using the inverse MAE from the bias-corrected error samples of the \( n \) contributing model forecasts over the past 30 days, with

\[
\hat{w}_i = (\text{MAE}_i)^{-1} \left[ \sum_{i=1}^{n} (\text{MAE}_i)^{-1} \right]^{-1}. \quad (2)
\]

Using these parameters, OCF based on \( n \) model forecasts \( f_i \) is given by

\[
\text{OCF} = \sum_{i=1}^{n} [\hat{w}_i(f_i - \hat{b}_i)]. \quad (3a)
\]

Breaking the forecasts \( f_i \) into the aforementioned components,

\[
\text{OCF} = \sum_{i=1}^{n} [\hat{w}_i(o + b_i + e_i) - \hat{b}_i]]. \quad (3b)
\]

Gathering terms, this becomes

\[
\text{OCF} = o + \sum_{i=1}^{n} [\hat{w}_i(b_i - \hat{b}_i)] + \sum_{i=1}^{n} (\hat{w}_i e_i). \quad (3c)
\]

The final two terms in (3c) highlight the importance of the bias removal and weighting schemes. It is important to remove as much systematic error as possible as systematic errors not removed may propagate through to the resultant forecast. Characterization of the random nature of the error distributions, as part of the weighting scheme, aids minimization of the random errors via compositing with highly variable models penalized for their reduced reliability.

b. Simple statistical downscaling

Grid points from the NWP models used in this study do not represent weather at specific latitudes and longitudes. Rather, they reflect values representative of an area whose size is dependent upon the spatial resolution of the grid. In this case, site-based direct model output (DMO) guidance is derived from model nearest-layer grid values through bilinear interpolation without any attempt to accommodate differences between grid-point and site elevations, or underlying surface variations such as land or sea. Therefore, DMO forecasts contain systematic errors due solely to the mismatch in representativeness between characteristics of the grid area and the specific location in question, including elevation, roughness length, and many other aspects (Tustison et al. 2001). The bias removal process in OCF, while accounting for true systematic model errors, is to a large extent a simple statistical downscaling scheme.

Meteorological processes can be considered in terms of differing spatial scales (Seed 2003). Let us visualize the atmospheric conditions as having large-scale (ls) and small-scale (ss) components, where the small-scale component is regarded as a deviation from the large-scale value. A site-specific observation is then composed of \( o_{ls} \), the average observational value over the area covered by a grid box, and \( o_{ss} \), which is the further deviation required to obtain the observation at a specific location, that is,

\[
o = o_{ls} + o_{ss}. \quad (4a)
\]

As mentioned, forecasts from NWP models do not resolve the spatial variability contained in site-based observations. Numerical forecasts from the international centers used in this study are received at particularly coarse spatial resolutions. As such, international model
forecasts \( (f_{\text{int}}) \) can be visualized as only containing large-scale components, with

\[
f_{\text{int}} = f_{\text{ls}}.
\]

Since the international model forecasts, received at reduced resolution, contain no explicit small-scale processes, their error \( (e_{\text{int}}) \) is as follows:

\[
e_{\text{int}} = f_{\text{int}} - o = f_{\text{ls}} - (o_{\text{ls}} + o_{\text{ss}}).
\]

Using Eq. (1), their site-based bias, based on the past 30 days of errors, \( (\hat{b}_{\text{int}}) \), is approximately

\[
\hat{b}_{\text{int}} = \frac{f_{\text{int}} - o}{\text{ls}} = \frac{f_{\text{ls}} - o_{\text{ls}} - o_{\text{ss}}}{\text{ls}}.
\]

As such, the bias parameter includes the large-scale bias plus the mean deviation for smaller-scale components. Thus, the bias correction process can be seen to add into the forecast statistical information about smaller-scale weather processes with

\[
f_{\text{int}} - \hat{b}_{\text{int}} = (f_{\text{ls}} - \hat{f}_{\text{ls}} - o_{\text{ls}}) + \overline{\sigma}_{\text{ss}} = \hat{f}_{\text{ls}} + \overline{\sigma}_{\text{ss}},
\]

where \( \hat{f}_{\text{ls}} \) is the bias-corrected large-scale forecast.

Since the finescale models run at spatial resolutions high enough to resolve some mesoscale phenomena, the bias correction procedure could be considered as correcting both large-scale and mesoscale biases, while adding in small-scale statistical information. The site-based estimation of representativeness errors by statistical values is inherently more appropriate for some variables (such as air temperature), than for others (such as wind speed), for which hourly guidance is required. This finding has implications for the appropriateness of extending the bias correction strategy to other meteorological variables.

4. Evaluation of bias correction and weighting parameters on an hourly basis

In Woodcock and Engel (2005), the bias correction methodology and moving-window sizes were evaluated using 2-m air temperature maxima and minima forecasts valid over 12-h periods. In this section we evaluate the estimation of bias correction and weighting parameters on an hourly basis, including all but one of the meteorological variables evaluated in this report. Hourly DMO forecast errors reveal modeling deficiencies in both phase and amplitude, and representations of shorter-lived, smaller-scale atmospheric features. a. Diurnal influence on site-based hourly DMO errors

To correctly interpret aggregate (over all stations) hourly DMO error statistics, we investigated the nature of the errors on an individual-site basis. Exploratory analysis of DMO errors was performed on an hour-by-hour, site-by-site basis by viewing site-specific box and whisker plots of error versus forecast hour (refer to Fig. 2 for two specific sites). Mean errors were found to vary with hour, with a diurnal modulation overlaying the normal error growth with time. The random error (variance) component of DMO errors, represented by the lengths of the boxes and whiskers, exhibited specific diurnal influences with greater variance at certain times of day. Given that a forecast hour of 12 from a 1200 UTC L1 issuance corresponds to approximately 1000 local time, this is equivalent to late afternoon in Hobart (Fig. 2a) and early morning at Mount Wellington (Fig. 2b).

Station-based error samples exhibited highly localized characteristics. This is highlighted in Fig. 2, with large differences in error characteristics exhibited between the two stations that are 22 km apart. This is most likely due to highly localized processes influencing site-based observations (refer to Fig. 1, inset).

b. Length of moving bias window

On an individual-station basis, a perfect bias correction scheme would result in zero median error and unchanged error variance. The impact of bias removal on the aggregate median error and error variance statistics was assessed using moving-window sizes of from 5 to 30 days in intervals of 5 days. The OCF BES bias correction [Eq. (1)] and weighting [Eq. (2)] strategies were applied on a station-by-station, hour-by-hour basis. Due to problems with observations and/or model data, corresponding errors were sometimes missing. To emulate an operational system, leeway was given to each bias window, allowing for three missing days. To investigate the relationship of the impact of bias correction and the spatial resolution of the NWP model, the analysis was repeated for a mesoscale (L1) and global (UK) model. QNH statistics were not generated. Station-based analysis was performed, but omitted from this report for brevity. The national statistics included represent an aggregation of the statistics from each of the individual stations that exhibit unique systematic and random errors.

Figure 3a demonstrates the effect of bias window size on 2-m air temperature error statistics \( \text{DMO} - \text{observation} \) using median error and variance plots. The median error statistics after bias correction do not exhibit
large variations with window size with similar values for windows sizes of 5–30 days. In comparison, the error variance statistics decrease as the window size increases. The initial reduction of the overall variance (using a 5-day window) reflects the overall error variance due mainly to individual stations biases. The further reduction in error variance experienced using larger window sizes was due to improved representativeness of the sample BES.

While the reductions in error variance due to bias correction did not explicitly converge, the improvement monotonically decreased with increasing window size and was beginning to asymptote after a window size of 10 days. The reduction in error variance was more apparent in the global (UK) model (not shown) than in the mesoscale (L1) model. This reflects the larger systematic error component in the global model, related to its coarser spatial resolution and the simplicity of the downscaling method used. The findings were very similar for both dewpoint temperature and relative humidity (not shown).

The response to the bias removal procedure (for air temperature, dewpoint temperature, and relative humidity) was very similar in May (not shown) to the

![Fig. 2. The (L1 – observed) 2-m air temperature box and whisker plots, over the February–March period. Fifty percent of values fall within the box, while 97% fall within the whiskers. The light gray line represents MAE, while the dark shows RMSE. Hobart and Mount Wellington are separated by about 22 km and have an elevation difference of approximately 1257 m (Fig. 1).](image-url)
February–March period. While there were some variations in the median error and error variance statistics, the same bias window sizes were favored. Balancing requirements of skill, robustness to missing data, and processing time, the daily OCF window size of 30 days, allowing for 15 missing days, was found to be appropriate for the estimation of bias and weighting parameters. Bias correction of wind speed forecasts is discussed further in the next section.

c. Wind speed and direction

Meridional and zonal wind components are modeled directly by NWP models. These components can be combined to produce wind speed and direction forecasts. Glahn and Lowry (1972) highlighted the systematic underestimation of wind speeds when using separate regressions for wind components. As such, separate wind direction and speed bias equations were adopted.

There are two possible approaches for the bias correction of wind direction forecasts: first, bias correction of individual meridional and zonal wind components and second, bias correction of wind direction forecasts where wind direction errors are defined as the angle between the forecast and observed wind vectors. Bias correction was found to not be beneficial to wind direction forecasts using either approach; the latter was selected for further evaluation.

Wind speed on the other hand seemed to benefit from bias correction (Fig. 3b), with the improvement

Fig. 3. Aggregate (L1—observed) median error and error variance statistics over all stations, for the February–March period; the meteorological variables are as indicated. Values are plotted after bias correction, for the specified window sizes (refer to key).
(due to bias correction) still increasing with a window size of 30 days. There was some variation of error statistics with forecast hour for all wind speed statistics (Fig. 3b). When viewing the subset of wind errors when the observed wind is above 5 m s\(^{-1}\), while bearing in mind the smaller sample size, there are biases remaining (Fig. 3c).

Further investigation into the relationship between the DMO forecast and the observed values was performed for all variables, using contour–scatterplots such as the one shown in Fig. 4. Plots for variables such as air and dewpoint temperature revealed that errors were only weakly dependent on the forecast value; these plots are not shown here. Some nonlinear characteristics were observed near the observational limit of the AWS instruments. In contrast, wind speed DMO errors were found to have a dependence on observed wind strength (Fig. 4).

Weak wind speeds and directions are hard to both measure and predict. Figure 4a highlights the existence of two separate regimes, with a concentration of values along the forecast axis where the observed wind speed is around 0 m s\(^{-1}\). Given that these forecasts were from a 1200 UTC base time, a 6-h forecast corresponds to approximately 0400 LT (in the Australian region). The concentration of forecast–observation pairs indicates an inability to forecast still environments overnight, with corresponding forecasts varying from weak to strong winds. This may be due to the greater importance of localized small-scale processes, such as inversions, under these conditions. Although the DMO forecasts had problems with predicting extremely weak winds throughout the whole forecast period, weak winds occurred more often over the nighttime period.

AWS anemometers have wind speed sensors that have a starting threshold of approximately 1 m s\(^{-1}\) (Potts et al. 1997). While the starting threshold of the anemometer wind speeds would contribute to the separate regimes, in particular the concentration on the 0 m s\(^{-1}\) observed axis, the observations would still correspond to weak winds. The corresponding forecasts, which are generally stronger than the possible 1 m s\(^{-1}\) observations, indicate deficiencies in the modeling and/or downscaling procedures that contribute to nonlinear characteristics of the wind speed errors.

Another important characteristic of wind speed forecast errors is the linear dependence on the forecast (or observed value). A cutoff wind speed of 5 m s\(^{-1}\), below which forecasters consider winds to be weak, is displayed in the plots in Fig. 4. The mean forecast for each observed value (the dotted line in Fig. 4) revealed that DMO forecasts tend to overpredict weak winds and underpredict strong winds. While instances of strong observed wind speeds are of importance to applications such as aviation forecasting, they are not well represented in the error statistics due to less frequent occurrence.

Given the complex nature of wind speed DMO errors, a bias correction scheme based upon constant (scalar) values representing all forecast values (section 3b) does not work well for these forecasts. Use of this type of approach will result in a residual bias for the least represented, but potentially more important, strong wind speeds. Future work will address this issue.

5. Inclusion of lower spatiotemporal resolution forecasts

International NWP forecasts are received at lower spatiotemporal resolutions compared to mesoscale model output produced in Australia (Table 1). The lower temporal resolution means that not all models
are available at all forecast hours. Thus, in order to include international NWP forecasts, a blending procedure is required to achieve self-consistent objective forecasts. This procedure is described in section 6. The lower spatial resolution increases the representativeness issues discussed in section 3b. This section assesses the impact of including these coarse-scale models on the accuracy of site-based consensus forecasts.

Model guidance was used to produce both a full composite (based on all available models) and a finescale composite (based on finescale Australian models alone). The consensus algorithm described in section 3a was applied on a site-by-site, hour-by-hour basis. The models used were those introduced in Table 1, with a 24-h lag introduced to the European Centre for Medium-Range Weather Forecasts (ECMWF) model to mimic the late arrival time experienced in operations. Forecasts were produced for the hours on which all model guidance was available, for example, every 6 h out to 42 h. Differences in error statistics were then assessed.

The mean percentage reduction in mean square error (MSE) due to the addition of the coarse-scale international model forecasts was between 9% and 28.5% during February–March (Table 2, column 3). The reductions in MSE for the full composite in comparison with the uncorrected component models were 52% (air temperature), 45% (dewpoint temperature), 47% (wind speed), 9% (wind direction), and 45% (RH) during February–March with slightly higher reductions in May (Table 2, column 2). The reductions for the finescale composite were considerably lower (Table 2, column 1).

The finescale models produced in Australia share both initial states and physical packages and are nested within the same global model. As a result, the cross correlations of the model errors are ~0.9 (before and after bias correction). The lowest cross correlations of errors after bias correction (~0.65) were found between coarse-scale (international) and finescale (Australian) models. Correlations between coarse-scale model errors (after bias correction) were slightly higher (~0.75), perhaps due to common representativeness errors. The redundancy affected the impact of compositing. The finescale composite had statistics closer to that of the individual models (Fig. 5).

Wind direction forecasts were found to have the lowest reductions in MSE for the full composite in comparison with the uncorrected component models at 9%. When including only statistics where the observed wind speed was above 5 m s⁻¹, the reduction was increased to 18% (Table 2, column 2). With the simple statistical

<table>
<thead>
<tr>
<th>Meteorological variable</th>
<th>Finescale composite vs uncorrected finescale models</th>
<th>Full composite vs all uncorrected models</th>
<th>Full composite vs finescale composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Feb–15 Mar 2004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air temperature</td>
<td>25.9</td>
<td>52.0</td>
<td>28.5</td>
</tr>
<tr>
<td>Dewpoint temperature</td>
<td>15.8</td>
<td>45.4</td>
<td>23.6</td>
</tr>
<tr>
<td>RH</td>
<td>22.8</td>
<td>44.6</td>
<td>24.4</td>
</tr>
<tr>
<td>QNH</td>
<td>29.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wind speed</td>
<td>28.2</td>
<td>46.9</td>
<td>19.8</td>
</tr>
<tr>
<td>Wind speed (observed wind speed above 5 m s⁻¹ threshold)</td>
<td>39.6</td>
<td>54.7</td>
<td>19.6</td>
</tr>
<tr>
<td>Wind direction</td>
<td>0.8</td>
<td>9.3</td>
<td>8.2</td>
</tr>
<tr>
<td>Wind direction (observed wind speed above 5 m s⁻¹ threshold)</td>
<td>0.5</td>
<td>17.7</td>
<td>19.9</td>
</tr>
<tr>
<td>1 May–31 May 2004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air temperature</td>
<td>37.9</td>
<td>60.2</td>
<td>28.5</td>
</tr>
<tr>
<td>Dewpoint temperature</td>
<td>26.9</td>
<td>46.8</td>
<td>18.0</td>
</tr>
<tr>
<td>RH</td>
<td>29.3</td>
<td>46.7</td>
<td>20.4</td>
</tr>
<tr>
<td>QNH</td>
<td>35.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wind speed</td>
<td>32.1</td>
<td>47.9</td>
<td>16.2</td>
</tr>
<tr>
<td>Wind speed (observed wind speed above 5 m s⁻¹ threshold)</td>
<td>36.0</td>
<td>50.8</td>
<td>20.9</td>
</tr>
<tr>
<td>Wind direction</td>
<td>0.3</td>
<td>5.2</td>
<td>5.2</td>
</tr>
<tr>
<td>Wind direction (observed wind speed above 5 m s⁻¹ threshold)</td>
<td>0.9</td>
<td>14.0</td>
<td>19.1</td>
</tr>
</tbody>
</table>
FIG. 5. Aggregate (forecast – observed) MSE statistics, over all stations, for the February–March period; the meteorological variables are as indicated. Plotted are the bias-corrected component international (white; in alphabetical order) and local (striped; in alphabetical order) models along with the finescale (gray) and full (black) composite statistics. (Here, L0 is omitted from the plot because of restricted station coverage, and QNH is omitted because of a lack of international input.)
bias correction not working for wind direction, and therefore having no downscaling, the increased reduction in wind direction MSE reflects the enhanced value of coarse-scale wind direction forecasts in scenarios where the wind direction varies less with scale. This has implications for the future usefulness of OCF wind direction forecasts.

For most meteorological variables and months, the MSE statistics of the full composite were lower than those of all the input models. The exceptions were when one model clearly outperformed the other ensemble members. This was the case for dewpoint temperature and wind direction (where the observed wind speed was greater than 5 m s\(^{-1}\)) in May (not shown). In those cases, the simple compositing algorithm led to a slight degradation of the MSE statistic in comparison with the best model. Overall, the inclusion of the international model forecasts increased the accuracy of the consensus.

6. Blending scheme

While the inclusion of the international models increases the accuracy on the hours that they are available, hourly consensus forecasts based on a simple mix of model temporal resolutions can result in “jumpy” forecasts. This is due to cases where the higher temporal resolution models deviate significantly from the lower temporal resolution models, resulting in obvious anomalies on the hours when they are not available. As such, a blending procedure was developed to produce self-consistent hourly OCF forecasts.

a. Blending procedure

Due to the availability of 3-h NWP forecasts (not used within this study), the blending procedure consisted of two stages: blending 6- and 3-h forecasts, then 3- and 1-h forecasts. The blending procedure is based on the following assumptions:

1) The accuracy of a consensus forecast increases with the number of models on which it is based.
2) High temporal resolution models contain extra information about the temporal structure of the forecast time series.
3) The difference between two forecasts at a given forecast hour has some temporal persistence.

The blending procedure is as follows:

1) generate consensus based on largest number of models available and use as “best guesses” or “fence posts,” available every 6 h;
2) generate consensus based on subset of models available every 3 h;
3) determine the difference between consensus forecasts at the fencepost hours;
4) interpolate the differences to intervening hours;\(^1\) and
5) adjust intervening forecasts by the interpolated differences.

Repeat steps 1–5 blending from 3- to 1-h forecasts.

When fencepost forecasts were missing (usually due to missing observations and therefore bias corrections), the consensus based solely on the 1-h models was returned for the full time series. This was to ensure consistency. A graphical demonstration of the blending process is provided in Fig. 6.

The performance of the blending scheme is affected by the validity of the assumptions upon which it is based. While the first assumption is true on average, there are instances where increasing the number of models in the consensus may degrade the resultant forecast. In regard to the second assumption, while the hourly resolution models attempt to simulate the temporal variability of atmospheric variables, the accuracy of those variations is not assured.

b. Performance of blending scheme

“Blended” hourly forecasts were produced, along with a naïvely interpolated forecast (where consensus

\(^1\) While cubic spline interpolation was used for interpolation across fencepost hours, the two-stage blending process is approximately equivalent to a single-stage 6- to 1-h blending process when no 3-h models are used.
forecasts are interpolated between fencepost hours by cubic spline functions). These were both compared against each other and the original “unblended” forecasts (where consensus forecasts are based on all forecasts available at a particular hour with no regard to consistency). The performance of the blending algorithm was investigated on three levels: exploratory analysis of the effect on single forecast runs (for specific stations), the impact on error statistics of specific stations, and the impact on the national error statistics over both the February–March and May periods.

Comparisons of national aggregate blended and unblended forecast statistics are displayed in Fig. 7. Sharp decreases in unblended MSE statistics (dash–dot line) at fencepost hours reflect the increase in consensus size. Blended forecast error variances (heavy black line) were more consistent across the forecasting period, similar to the uncorrected L1 model (dashed line). They were generally lower than, or of similar magnitude to, unblended forecasts. The interpolated forecasts (heavy gray line) had behavior specific to the meteorological variables.
Blended forecasts outperformed both the unblended and interpolated forecasts for both air temperature and relative humidity. This reflected the inability of interpolation to reconstruct the diurnal cycle. Six-hourly forecasts sample the diurnal cycle with only four points. This is at the Nyquist limit for signal reconstruction. As a result, the interpolated forecasts dampened the diurnal cycle. Hence, the placements of fencepost hours had varied impacts on specific times during the day, as well as specific locations. There were forecast hours where the interpolated forecast performed very poorly, while at other times of day it outperformed the blended forecast for some variables. In comparison with the interpolated forecasts, the blending processes resulted in a reduction in average MSEs of 24% (air temperature) and 13% (RH) during February–March with slightly higher reductions in May (Table 3, column 2).

The blended wind speed forecasts outperformed the unblended forecasts, but the interpolated forecasts sometimes performed better (Fig. 7d). Even though finescale (mesoscale) models attempt to simulate hourly variability, the more constant interpolated forecasts verify better in some cases, especially in regard to the (implicit) timing of wind speed changes, which are hard to predict. In these cases, the mesoscale forecasts are penalized for timing (phase) errors. While smoother forecasts may verify more favorably, their usefulness to forecasters is questionable. This is especially true when events such as fronts are passing through. A mistimed event may be of more use than a missed one to an operational forecaster, as a kind of “heads up.” Interpolated wind speed forecast errors exhibited much more variability across the forecasting period, indicating that blended forecasts may be more reliable.

The statistics based solely on the occasions that the observed wind strength was above 5 m s$^{-1}$ indicated that the interpolated wind speed forecasts outperformed the blended forecasts in general, with an average reduction in MSE of −10% over May (Table 3, column 2).

Little was gained by performing the blending procedure in comparison to a naïve interpolation for dewpoint temperature in May (0% average reduction in MSE). The interpolated forecast was often quite close to the blended forecast. This reflects the fact that dewpoint temperature generally has a smaller diurnal cycle than air temperature or relative humidity (Table 3, column 2).

The QNH consensus was composed solely of nested locally produced finescale models. Blending showed minor increases in skill accordingly. Improvement in QNH as a result of the blending process was due to the L1 and L3 QNH forecasts having a temporal resolution of 3 h.

Overall, wind direction forecasts were not greatly enhanced by the addition of the international models (Fig. 5e; Table 2, column 3). This was reflected in the minor differences in skill (both positive and negative) attributable to the blending process (Fig. 7e; Table 3, column 1). When analyzing the subset of wind direction fore-
casts corresponding to strong observed wind strength (Table 2, column 3), the slightly more favorable response to the larger ensemble was seen (Table 3, column 1). In the later forecast hours, where the additional models appeared to have the best effect, the blending showed more skill. In the earlier periods, the blending both enhanced and degraded the forecast errors depending on the actual forecast hour (and month). Larger mean reductions in the MSE due to blending (in comparison with not blending) were experienced for wind direction forecasts when the observed wind strength was greater than 5 m s$^{-1}$ (Table 3).

While the blending algorithm helped with the consistency of error variance across forecast hours, the blended forecast error variances were still slightly more variable than those of a single model (such as L1), across forecast hours. Despite this, the comparative mean percentage reduction in MSE between blended forecasts and uncorrected L1 forecasts was above 40% for all meteorological variables except QNH and wind direction.

In summary, although the blending scheme achieved the aim of producing self-consistent forecasts, the accuracy of any particular feature was not guaranteed. The blending process resulted in more consistent error variances across forecast hours. Given the simplicity of the blending approach, the increase in the consistency of the hourly OCF forecasts was quite substantial in comparison with the unblended forecasts. The blended forecasts showed more skill than did the interpolated forecasts for air temperature and relative humidity, and QNH. There was little gained over a naïve interpolation for dewpoint temperature. Hourly OCF outperformed the hourly (uncorrected) L1 forecasts currently available in regional offices.

7. Future improvements

The bias correction procedure of the OCF methodology addresses the differences in spatial resolution between the NWP models and site-based observations. While OCF can readily be extended from daily to hourly temporal resolution, the simple statistical downscaling performed is more appropriate for meteorological variables such as 2-m air temperature (which has a relatively systematic relationship with height and, therefore, topography) than other variables such as wind speed and wind direction, which interact in a more complex manner with topography. More complex statistical or dynamical downscaling (accompanied by large-scale bias correction) may benefit the OCF technique as a whole; this will be the subject of future study.

While the simple blending technique introduced in section 6a performs better than naïve interpolation, more sophisticated techniques have yet to be explored. For example, acknowledgment of the simple statistical downscaling included in the site-based OCF scheme highlights the possibility of producing gridded OCF forecasts. By correcting and combining model forecasts on the spatial resolutions at which they are made available, it would be possible to generate large-scale gridded consensus forecasts that can then be further down-scaled. The gridded forecasts (containing spatial information not available in site-based forecasts) could be utilized in generating consistent hourly forecasts by predicting how weather patterns, rather than site-based time series, develop in the intermediate hours. Using the available gridded forecast information, we may also be able to produce grid-based probability forecasts (Stensrud and Yussouf 2005).

8. Conclusions

It has been shown that the OCF scheme can be successfully extended down to hourly temporal resolution. OCF guidance generally outperforms the component models, even after bias correction, when considering aggregate site-based statistics. The systematic and random natures of direct model output forecast errors vary on an hour-by-hour and station-by-station basis. The bias-corrected and -weighted consensus in the OCF methodology reduces the error by 5%–60%, depending on the meteorological variable (Table 2, column 2). A window size of 15–30 days has been shown to be appropriate for determining bias correction and weighting parameters for certain meteorological variables. The combination of model forecasts exhibits greater skill when coarse-scale international forecasts are included. A simple blending scheme extended the skill to hourly resolution to produce consistent hourly forecasts. The blending scheme exhibits greater skill than interpolation, for air temperature, relative humidity, and QNH. The blending scheme also exhibits greater accuracy than do the currently available hourly forecasts.

Although the OCF scheme has been shown to perform well, there are limitations to the bias removal procedures. This was highlighted by the performance of wind strength predictions, suggesting that a more complicated downscaling procedure could further increase the skill of the OCF scheme. This is the subject of continuing research.

Acknowledgments. The author would like to thank many people including Terry Hart, Tom Keenan, Tony Bannister, Graham Mills, Jeff Kepert, Alan Seed, and
Frank Woodcock, Andrew Amad-Corson and Jim Fraser were a great source of help with the management of model data. Comments on earlier versions of the manuscript made by Frank Woodcock and Todd Lane were extremely useful.

REFERENCES


