Verifying Operational Forecasts of Land-Sea Breeze and Boundary Layer Mixing Processes

Ewan Short*

School of Earth Sciences, and ARC Centre of Excellence for Climate Extremes, The University of Melbourne, Melbourne, Victoria, Australia.

Corresponding author address: School of Earth Sciences, The University of Melbourne, Melbourne, Victoria, Australia.

E-mail: shorte1@student.unimelb.edu.au

ABSTRACT

Forecasts issued by the Australian Bureau of Meteorology (BoM) are based on automated post-processed model data that is edited by human forecasters. Two types of edits are commonly made to the wind fields. These edits aim to improve how the influence of boundary layer mixing and land-sea breeze processes are represented in the forecast. In this study we compare the diurnally varying component of the BoM's official edited wind forecast, with that of station observations and unedited model datasets, to assess changes to error and bias resulting from these edits. We consider coastal locations across Australia over June, July and August 2018, aggregating data over three spatial scales. The edited forecast generally only produces a lower mean absolute error than model guidance at the coarsest spatial scale (500×150 km^2 to 2000×150 km^2), but can achieve lower seasonal biases over all spatial scales. However, the edited forecast only reduces errors or biases at particular times and locations, and rarely produces lower errors or biases than all model guidance products simultaneously. To better understand the biases in the diurnal wind cycles, we fit modified ellipses to the temporal hodographs of seasonally averaged diurnal wind cycles. Biases in the official forecast diurnal cycle vary with location for multiple reasons, including biases in the directions sea-breezes approach coastlines, amplitude and shape biases in the hodographs, and disagreement as to whether sea-breeze or boundary layer mixing processes contribute most to the diurnal cycle.

1. Introduction

Modern weather forecasts are typically produced by models in conjunction with human forecasters. Forecasters working for the Australian Bureau of Meteorology (BoM) construct a seven day forecast by loading model data into a software package called the Graphical Forecast Editor (GFE), then editing this model data using tools within the GFE. Forecasters working for the United States National Weather Service also use GFE and utilise a similar approach. Forecasters can choose which model to base their forecast on, and refer to this as a choice of *model guidance*. Edits are typically made to account for processes that are under-resolved at the resolutions of the model guidance products, or to correct for perceived biases of the model guidance being used. In Australia, the resulting gridded forecast datasets are provided to the public through the BoM's online MetEye data browser (Bureau of Meteorology 2019), and are also translated into text and icon forecasts algorithmically.

Forecasters, and the weather services that employ them, have good reasons for ensuring the diurnally varying component of their wind forecasts are as accurate as possible. In addition to the significant contribution diurnal wind cycles make to overall wind fields (e.g. Dai and Deser 1999), diurnal wind cycles are important for the ventilation of pollution, with sea-breezes transporting clean maritime air inland, where it helps flush polluted air out of the boundary layer (Miller et al. 2003; Physick and Abbs 1992). Furthermore, diurnal wind cycles affect the function of wind turbines (Englberger and Dörnbrack 2018) and the design of wind farms (Abkar et al. 2016), as daily patterns of boundary layer stability affect turbine wake turbulence, and the losses in wind power that result.

Australian forecasters generally make two types of edits to the surface wind fields on a routine daily basis. The first involves modifying the surface winds after sunrise at locations where the

forecaster believes the model guidance is providing a poor representation of boundary layer mixing processes. Boundary layer mixing occurs as the land surface heats up, producing an unstable boundary layer which transports momentum downward to the surface layer. Before this mixing occurs, winds are typically both weaker and ageostrophically oriented due to surface friction (Lee 2018), and so mixing can affect both the speed and direction of the surface winds. Australian forecasters perform these edits using a GFE tool which allows them to specify a region over which to apply the edit, a height z and a percentage p, with the tool then calculating a weighted average of the surface winds and winds at z, weighted by p.

The second type of edit involves changing the afternoon and evening surface winds around those coastlines where the forecaster believes the model guidance is resolving the sea-breeze poorly. Similarly to with boundary layer mixing, these edits are performed using a GFE tool that allows forecasters to trace out the relevant coastline graphically, choose a wind speed and a time, with the tool then smoothly blending in winds of the given speed perpendicular to the traced coastline at the given time.

To our knowledge, no published work has assessed the diurnal component of human edited forecasts, although some previous studies have assessed the performance of different operational models at specific locations. Svensson et al. (2011) examined thirty different operational model simulations, including models from most major forecasting centres utilising most commonly used boundary layer parametrisation schemes, and compared their performance with a large eddy simulation (LES), and observations at Kansas, USA, during October 1999. They found that both the models and LES failed to capture the roughly 6 kn (1 kn ≈ 0.514 m s^{-1}) jump in wind speeds shortly after sunrise, and underestimated morning low level turbulence and wind speeds.

Other studies have assessed near-surface wind forecasts, verifying the total wind speeds, not just the diurnal component. Pinson and Hagedorn (2012) studied the 10 m wind speeds from the

European Centre for Medium Range Weather Forecasting (ECMWF) operational model ensemble across western Europe over December, January, February 2008/09. They found that the worst performing regions were coastal and mountainous areas, and attributed this to the small scale processes, e.g. sea and mountain breezes, that are under-resolved by the ensemble's coarse 50 km spatial resolution.

The present study has two goals. First, to describe a method for comparing the diurnal cycles of human edited wind forecasts to those of unedited model guidance forecasts, in order to assess where and when human edits produce a reduction in error or bias. Second, to apply this methodology across Australian coastal locations to assess both boundary layer mixing and land-sea breeze forecaster edits. The remainder of this paper is organised as follows. Section 2 describes the methodology, and datasets to which it is applied, section 3 provides results, and sections 4 and 5 provide a discussion and a conclusion, respectively.

2. Data and Methods

This study compares both human edited and unedited Australian Bureau of Meteorology (BoM) wind forecasts with automatic weather station (AWS) data across Australia. The comparison is performed by first isolating the diurnal perturbations of each dataset by subtracting 24-hour running means, then comparing these perturbations on an hour-by-hour basis.

a. Data

Four datasets are considered in this study; the human edited official BoM wind forecast data that is issued to the public, observational data from automatic weather stations (AWS) across Australia, unedited data from the ECMWF's high resolution 10-day forecast model (HRES), and unedited model data from the operational Australian Community Climate and Earth System Simulator (AC-

CESS), noting that HRES and ACCESS are two of the model guidance products most commonly used by Australian forecasters for winds. We consider just the lead-day one forecasts of the official forecast, HRES and ACCESS, for reasons discussed below.

This study primarily considers the austral winter months of June, July and August 2018. This short time period was chosen to reduce the effect of changing seasonal and climatic conditions, changing forecasting practice and staff, and of changes to the ACCESS and HRES models. Results for December, January and February 2017/18 are occasionally mentioned to strengthen conclusions or provide a seasonal contrast.

ACCESS is a nested model: in this study we consider the component covering the Australian region from 65.0° south to 16.95° north, and 65.0° east to 184.57° east. This model runs at a 0.11° ($\approx$ 12 km) horizontal grid spacing, with a standard time-step of 5 minutes: occasionally a shorter time step of 2.5 minutes is used to overcome numerical instabilities (Bureau of Meteorology 2016). HRES runs at an $\approx$ 9 km horizontal grid spacing, with a 7.5 minute time-step (Modigliani and Maass 2017).

Both ACCESS and HRES use parametrisation schemes to simulate sub-grid scale boundary layer turbulence, and the resultant mixing. ACCESS uses the schemes of Lock et al. (2000) and Louis (1979) for unstable and stable boundary layers respectively (Bureau of Meteorology 2010). HRES uses similar schemes that the ECMWF develop in-house (European Center for Medium Range Weather Forecasting 2018).

The Bureau's official forecast dataset is produced on a state by state basis at forecasting centres located in most state capitals. To construct the official forecast dataset, forecasters make a choice of model guidance in the GFE, which then interpolates or upscales the model data onto a standard 3 km spatial grid for Victoria and Tasmania, or a 6 km grid for the rest of the country. GFE displays model data at hourly intervals by taking the model guidance output at each hour UTC, with the

exception of the HRES model data which is only provided to the BoM at 3 hourly intervals, and is therefore linearly interpolated to hourly intervals by the GFE. Forecasters then make edits to these 3 or 6 km hourly grids to produce the official forecast datasets.

We therefore compare the official forecast and model guidance datasets as they appear in the GFE, i.e. we compare the upscaled or interpolated datasets on the standardised 3 or 6 km, hourly grids. This both ensures a consistent comparison between model guidance products of different spatial resolutions, and an assessment of how the official forecast compares to the model guidance products as they actually appear to forecasters in the GFE. This is the standard approach the BoM takes when verifying any forecast variable.

These datasets are compared with observations from Australian automatic weather stations (AWS), which typically record wind speed and direction each minute. After basic quality control, 10 minute averages of speed and direction are taken at each station at each hour UTC, usually over the ten minutes leading up to each hour. To calculate verification results, each station is matched with the nearest 3 or 6 km grid-point in the datasets described above.

b. Assessing Diurnal Cycles

Forecasters edit model guidance wind data to account for under-resolved sea-breeze and boundary layer mixing processes. Instead of attempting to assess each type of edit individually, we study the overall diurnal signal by subtracting a twenty four hour centred running mean *background wind* from each zonal and meridional hourly wind data point, to create wind *perturbation* datasets.

To compare errors in the official forecast, ACCESS and HRES diurnal cycles we calculate the Euclidean distances between the official or model guidance perturbation vectors at each hour UTC, and the corresponding AWS perturbation vectors at each hour UTC, viewing the Euclidean distance as a measure of absolute error. For example, to assess whether the official forecast perturba-

tions, $\boldsymbol{u}_{\mathrm{O}}$, or ACCESS perturbations, $\boldsymbol{u}_{\mathrm{A}}$, produce lower absolute errors when compared with the observed AWS perturbations, $\boldsymbol{u}_{\mathrm{AWS}}$, we calculate the *difference of absolute errors* (DAE),

$$\mathrm{DAE}_{\mathrm{OA}} = |\boldsymbol{u}_{\mathrm{AWS}} - \boldsymbol{u}_{\mathrm{A}}| - |\boldsymbol{u}_{\mathrm{AWS}} - \boldsymbol{u}_{\mathrm{O}}|\,. \tag{1}$$

The analogously defined quantities $\mathrm{DAE}_{\mathrm{OH}}$ and $\mathrm{DAE}_{\mathrm{HA}}$ provide a comparison of the official forecast and HRES perturbations, and of the HRES and ACCESS perturbations, respectively. We can then take means of the DAE on an hourly basis; i.e. average all the 00:00 UTC DAE values, all the 01:00 UTC values, and so forth, and denote such an average by $\overline{\mathrm{DAE}}$.

Note that $\overline{\mathrm{DAE}}$ compares just *one aspect* of the official forecast with model guidance: it does not, for instance, assess whether the variability of the official forecast is more realistic than that of model guidance. Thus, any statements about performance made throughout this paper refer solely to $\overline{\mathrm{DAE}}$, or subsequently defined metrics, and no claim is being made that these are sufficient to completely characterise the accuracy, or value to the user, of how the diurnal wind cycle is represented in competing forecasts.

Sea-breeze and boundary layer mixing processes depend on the background atmospheric conditions in which they occur. By comparing wind perturbations rather than the overall wind fields we are not claiming these background conditions are irrelevant. However, when a forecaster makes an edit of a wind forecast to better resolve these processes, they are implicitly assuming that future background conditions will be close enough to the preceding 24 hour mean state, or to model predictions of the mean state, to justify making the edit. Thus, it makes sense to compare forecast perturbations to observed perturbations, as long as differences are interpreted as a consequence not only of how the forecaster or model resolves the diurnal cycle, but of how differences in the background state contribute to differences in the perturbations. To minimise the importance of background state differences, this study focuses exclusively on lead-day one forecasts.

Given the large degree of turbulence or random variability in both the AWS, official, and model datasets, care must be taken to ensure we do not pre-emptively conclude the official forecast has outperformed model guidance when $\overline{\text{DAE}} > 0$ purely by chance. The method for estimating confidence in $\overline{\text{DAE}}$ is based on a method proposed by Griffiths et al. (2017). Time series formed from the DAE values at a particular time, say 00:00 UTC, across the three month time period, are treated as an independent sample of a random variable E. The sampling distribution for each $\overline{\text{DAE}}$ can be modelled by a Student's t-distribution, and from this we calculate the probability that E is positive, denoted $\Pr(E > 0)$.

Although temporal autocorrelations of DAE, i.e. correlations between DAE values at a particular hour from one day to the next, are in practice small or non-existent, they are still accounted for by reducing the "effective" sample size to $n(1-\rho_1)/(1+\rho_1)$, where n is the actual sample size and ρ_1 is the lag-1 autocorrelation (Zwiers and von Storch 1995; Wilks 2011). In the language of statistical hypothesis testing, the null hypothesis that $E = 0$ would be rejected at significance level α if $\Pr(E > 0) > 1 - \frac{\alpha}{2}$ or $\Pr(E < 0) > 1 - \frac{\alpha}{2}$. However, in this study we prefer to simply state the value of $\Pr(E > 0)$, referring to this as a *confidence score*, and noting $\Pr(E < 0) = 1 - \Pr(E > 0)$. We say the official forecast outperforms model guidance with "high confidence" if $\Pr(E > 0) \geq 95\%$, or that model guidance outperforms the official forecast with "high confidence" if $\Pr(E > 0) \leq 5\%$, with high confidence implicit whenever it is not explicitly mentioned.

Following the "fuzzy verification" approach outlined by Ebert (2008), forecast and observational perturbation datasets are compared not only at individual stations, but are also averaged over two coarser spatial scales before being compared. The individual stations we consider are the 8 capital city *airport stations*, marked by stars in Fig. 1, as their high operational significance means that they are typically the most well maintained. An intermediate spatial scale is formed by averaging data over the 10 stations closest to each capital city airport station, with some flexibility allowed to

ensure stations are roughly parallel to the nearest coastline. These station groups are referred to as the *city station groups*. The coarsest spatial scale is formed by averaging over all stations within 150 km of the nearest coastline, and grouping these by state. The Western Australian coastline is subdivided into three pieces, and stations along the Gulf of Carpentaria, north Queensland Peninsula, and Tasmanian coastlines are neglected, in order to ensure each station group corresponds to an approximately linear segment of coastline to better resolve the land-sea breeze after spatial averaging (e.g. Vincent and Lane 2016). These eight station groups are referred to as the *coastal station groups*.

To compare errors in the perturbations over the two coarser spatial scales, we modify the definition of DAE in equation (1) so that each perturbation dataset is first spatially averaged over either the city or coastal station groups. Confidence scores are calculated for the city and coastal station groups in the same way as for the individual airport stations, treating the spatially averaged data as a single time series. This provides a conservative way to deal with spatial correlation between the stations in each group (Griffiths et al. 2017).

To compare biases in the diurnal cycles of each dataset, we calculate the *difference of biases* (DB),

$$\mathrm{DB}_{\mathrm{OA}} = |\overline{\boldsymbol{u}}_{\mathrm{AWS}} - \overline{\boldsymbol{u}}_{\mathrm{O}}| - |\overline{\boldsymbol{u}}_{\mathrm{AWS}} - \overline{\boldsymbol{u}}_{\mathrm{A}}|, \tag{2}$$

with $\mathrm{DB}_{\mathrm{OH}}$ and $\mathrm{DB}_{\mathrm{HA}}$ defined analogously, where the over-bars denote temporal averages of the perturbations at a particular hour, over June, July and August 2018. These temporally averaged perturbations can be viewed as the climatological diurnal wind cycles over the three month study period for each dataset. Biases over the city and coastal station groups are calculated by taking the spatial average before the temporal average. Uncertainty in the DB is estimated through bootstrapping (Efron 1979). This is done by performing resampling with replacement on the underlying perturbation datasets, and calculating the DB multiple times using these resampled datasets. This

provides a distribution of DB values, which analogously to with DAE, we treat as a sample from a random variable B, and use this to estimate $\Pr(B > 0)$.

Another approach to forecast verification is to assess structural features of the phenomena being forecast rather than errors or biases of point predictions; this approach is particularly important at small spatiotemporal scales (e.g. Mass et al. 2002; Rife and Davis 2005). Gille et al. (2005) obtained summary statistics on the observed structure of mean diurnal wind cycles by using linear regression to calculate the coefficients u_i, v_i $i = 0, 1, 2$, for the fits

$$u = u_0 + u_1 \cos(\omega t) + u_2 \sin(\omega t), \tag{3}$$

$$v = v_0 + v_1 \sin(\omega t) + v_2 \sin(\omega t), \tag{4}$$

where ω is the angular frequency of the earth and t is the local solar time in seconds. These fits trace out ellipses in the x, y plane, and descriptive metrics like the eccentricity of the ellipse and the angle the semi-major axis makes with lines of latitude, can be calculated directly from the coefficients u_1, u_2, v_1 and v_2. Gille et al. (2005) applied this fit to scatterometer data, which after temporal averaging resulted in just four zonal and meridional values per location, and as such the fit performed very well.

However, equations (3) and (4) do not provide a good fit for the hourly data considered here, primarily because they assume a twelve hour symmetry in the evolution of the diurnal cycle. In practice, asymmetries between daytime heating and nighttime cooling (e.g. Svensson et al. 2011) result in surface wind perturbations accelerating rapidly just after sunrise, but remaining comparatively stagnant at night (e.g. Fig. 9). Thus, we instead fit the equations

$$u = u_0 + u_1 \cos(\alpha(\psi, t)) + u_2 \sin(\alpha(\psi, t)), \tag{5}$$

$$v = v_0 + v_1 \sin(\alpha(\psi, t)) + v_2 \sin(\alpha(\psi, t)), \tag{6}$$

to the climatological perturbations, with α the function from $[0,24) \times [0,2\pi) \to [0,2\pi)$ given by

$$\alpha(\psi,t) \equiv \pi \left[\sin\left(\pi \frac{(t-\psi) \bmod 24}{24} - \frac{\pi}{2} \right) + 1 \right], \tag{7}$$

with t the time in units of hours UTC, and ψ providing the time when the wind perturbations vary least with time, noting that the same value of ψ is used for both the zonal and meridional perturbations. For each climatological diurnal wind cycle, we solve for the seven parameters u_0, u_1, u_2, v_0, v_1, v_2 and ψ using non-linear regression.

3. Results

In this section, the methods described in section 2 are applied to Australian forecast and station data over the months of June, July and August 2018. First, differences in absolute errors (DAE) and differences in biases (DB) over this time period are assessed. Second, structural indices are compared to elucidate the physical reasons for biases.

a. Absolute Errors

Figure 2 provides the mean difference of absolute error $\overline{\text{DAE}}$ values and confidence scores defined in section 2 for the coastal station groups shown in Fig. 1, for $\overline{\text{DAE}}_{\text{OA}}$, $\overline{\text{DAE}}_{\text{OH}}$ and $\overline{\text{DAE}}_{\text{HA}}$, which represent the official forecast versus ACCESS, official forecast versus HRES, and HRES versus ACCESS comparisons, respectively. The results indicate that for the majority of station groups and hours, both the unedited ACCESS and HRES models outperform the official forecast. The lowest $\overline{\text{DAE}}$ values occur at the NT station group at 23:00 and 00:00 UTC for both $\overline{\text{DAE}}_{\text{OA}}$ and $\overline{\text{DAE}}_{\text{OH}}$. Although the official forecast outperforms at least one of ACCESS or HRES at multiple times and station groups, the only group and time where it outperforms both is 05:00 UTC over the South WA station group. HRES generally outperforms ACCESS from 10:00 - 14:00 UTC, with the South WA station group being the main exception.

Figures 3 and 4 provide case studies of the NT and South WA station groups, respectively. Figure 3 a) provides a time series of DAE for the NT station group at 23:00 UTC. The time series shows significant temporal variability, with DAE frequently dropping below -2 kn. Figures 3 b) and c) show hodographs of the winds and wind perturbations, respectively, at each hour UTC on the 3rd of July, which provides an interesting example.

Figure 3 b) shows that the official wind forecast on this day was likely based on edited ACCESS from 00:00 to 06:00 UTC, then edited HRES from 07:00 to 13:00 UTC, then unedited ACCESS from 15:00 to 21:00 UTC. At 22:00 and 23:00 UTC, the official forecast winds acquire stronger east-northeasterly components than the other datasets. Figure 5 a) shows the first ten values from wind soundings at Darwin Airport at 12:00 UTC on July 3rd and 00:00 UTC on July 4th. In both instances the winds are east-southeasterly, and so the rapidly changing wind perturbations at 22:00 UTC in the official forecast likely reflect a boundary layer mixing edit that has been applied either too early, or has strengthened the southeasterly component of the winds too much. Similar issues create low DAE values on the 8th of June and 9th and 10th of July.

Figure 4 a) provides a time series of DAE for the South WA station group at 05:00 UTC. As with the NT station group there is significant temporal variability, with DAE frequently exceeding 1 kn. Figures 4 b) and c) provide hodographs of the winds and wind perturbations, respectively, on the 9th of June, another interesting example. The perturbation hodograph shows both HRES and ACCESS under-predicting the amplitude of the diurnal wind cycle on this day. Figure 5 shows wind soundings at Perth Airport, the nearest station to provide wind soundings, between 12:00 UTC on the 8th June and 12:00 UTC on the 9th June. The 8th June 12:00 UTC sounding shows surface northerlies of around 6 kn, becoming west to northwesterlies of over 20 kn 2.4 km above the surface. However, the subsequent sounding at 00:00 UTC on the 9th of June shows that the

winds acquire a strong northerly component of 30 kn in the first 500 m of the atmosphere, with the final sounding indicating a strong northwesterly wind at 725 m persisting until 12:00 UTC.

In Fig. 4 c), the official forecast perturbations from 04:00 to 07:00 UTC show stronger westerly perturbations than either ACCESS or HRES, improving the amplitude of the official forecast's diurnal wind cycle. However, the AWS perturbations are more northerly than those of the official forecast, and so the official forecast winds have been strengthened in a slightly incorrect direction. One explanation for this discrepancy is that the official forecast has been edited based on the June 8$^{\text{th}}$ 12:00 UTC sounding, with the winds above the surface changing direction in the subsequent 12 hours. A similar explanation can be given for the high DAE scores on the 3$^{\text{rd}}$ of August, although in this case the official forecast slightly improves both the magnitude and direction of the 05:00 UTC wind perturbations.

Fig. 6 presents the $\overline{\text{DAE}}$ values and confidence scores for the airport stations, and city station groups, for the official forecast versus HRES comparison, i.e. $\overline{\text{DAE}}_{\text{OH}}$. The results for the airport stations are noisier than the results for the coastal station groups in Figs. 2 c) and d), although they share some similarities. For instance, the official forecast outperforms HRES at 01:00 and 02:00 UTC at both the Darwin airport station and the NT coastal station group. There are four other instances where the official forecast outperforms HRES with at least 90% confidence, although this could simply be occurring by chance due repeated testing (Wilks 2011, p. 178).

For the city station groups, HRES outperforms the official forecast almost uniformly. The main exception is the Darwin city station group, where the official forecast outperforms HRES at 02:00 UTC, and there is ambiguity as to whether the official forecast or HRES performs better at 01:00, 03:00 and 04:00 UTC, and from 15:00 to 22:00 UTC. The analogous $\overline{\text{DAE}}_{\text{OA}}$ official forecast versus ACCESS comparisons (not shown) are similar, with the airport station results noisy, but ACCESS outperforming the official forecast over the city station groups for the vast majority

of times and locations. Over the December, January, February 2017/18 season, HRES also outperforms the official forecast almost uniformly over the city station groups, although the official forecast versus ACCESS comparisons are more ambiguous.

Figure 7 provides the $\overline{\text{DAE}}$ values and confidence scores for the airport stations, and city station groups, for the HRES versus ACCESS comparison. As with Fig. 6, the results for the airport stations are noisy, but more often than not show that HRES outperforms ACCESS. The results for the city station groups show HRES usually outperforms ACCESS, the main exceptions being the Darwin and Canberra city station groups. Results for the December, January, February 2017/18 season are again similar, but here HRES outperforms ACCESS over the city station groups almost uniformly.

b. Seasonal Biases

Figure 8 provides the difference of biases (DB) and confidence scores defined in section 2, for the coastal station groups for DB_{OA}, DB_{OH} and DB_{HA}, which represent the the official forecast versus ACCESS, official forecast versus HRES, and HRES versus ACCESS comparisons, respectively. At the NT station at 03:00 UTC, the official forecast outperforms both ACCESS and HRES with confidence $\geq 93\%$. However, both ACCESS and HRES outperform the official forecast at 23:00 and 00:00 UTC, and from 05:00 to 11:00 UTC, consistent with the $\overline{\text{DAE}}$ results of Fig. 2. Figure 9 a) shows that these biases are mostly a consequence of amplitude biases in the official forecast's diurnal cycle.

At the South WA station group from 01:00 to 05:00 UTC, the official forecast outperforms HRES with confidence scores of at least 88%. Figure 9 b) shows that HRES underestimates the westerly perturbations at these times, with these perturbations likely associated with boundary layer mixing processes, as discussed in section 3 a. Each of the official forecast, ACCESS and

HRES underestimate the amplitude of the diurnal cycle between 02:00 and 10:00 UTC, including both the westerly perturbations and the southerly sea-breeze perturbations.

At the NSW station group from 17:00 to 19:00 UTC, the official forecast outperforms both ACCESS and HRES with confidence scores of at least least 95% and 75%, respectively. Figure 9 c) shows that these times correspond to "dimples" in the perturbation temporal hodographs that are present in all four datasets. The official forecast hodograph closely resembles that of ACCESS, except for this dimple, which has been exaggerated relative to ACCESS. Figure 9 c) also shows that although HRES exaggerates the amplitude of the easterly sea-breeze perturbations, it captures the narrower shape of the AWS hodograph better than the official forecast or ACCESS.

At the SA station group from 02:00 to 05:00 UTC and 09:00 to 12:00 UTC, the official forecast outperforms both ACCESS and HRES, although confidence scores do not exceed 88% and 65% respectively. Figure 9 d) shows that although the official forecast captures the amplitude of the perturbations from 01:00 to 05:00 UTC almost perfectly, its diurnal cycle is out of phase with that of the AWS during this period, explaining why the official forecast only slightly outperforms ACCESS in the results of Figures 8 a) and b).

For comparison, Fig. 10 presents the DB values and confidence scores for DB_{OH}, which represents the official forecast versus HRES comparison, for the airport stations and city station groups. Some regions exhibit consistent results across all three spatial scales, for example, the official forecast is less biased than HRES with at least 80% confidence at Sydney airport, the Sydney city station group, and the NSW coastal station group, from 14:00 to 18:00 UTC.

c. Ellipse Fits

The hodographs in Fig. 9 are roughly elliptical in shape, suggesting that descriptive quantities can be estimated by fitting equations (5) and (6) to the zonal and meridional climatological per-

turbations, as described in section 2. Figure 11 gives the R^2 values for the fits of the zonal and meridional perturbations to equations (5) and (6), respectively. The fit performs best at the coastal station group spatial scale, with R^2 generally above 95%.

Figure 12 provides four descriptive quantities based on the fits of equations (5) and (6) to the averaged perturbations: these are maximum perturbation speed, eccentricity of the fitted ellipse, angle the semi-major axis makes with lines of latitude, and the time at which the maximum perturbation speed is achieved. Fig. 12 a) shows that at Brisbane airport the maximum AWS perturbation is at least 1 kn greater than the official forecast, ACCESS and HRES, and Fig. 12 c) shows that the orientation of the AWS fitted ellipse is at least 20 degrees anti-clockwise from the other datasets. Figures 13 a) and b) show hodographs of the Brisbane airport climatological perturbations and ellipse fits, respectively. Although the ellipse fits suppress some of the asymmetric details, they capture the amplitudes and orientations of the real climatological diurnal cycles well. In this case the results show that the average AWS sea-breeze approaches from the northeast, whereas the official forecast, HRES and ACCESS sea-breezes approach more from the east-northeast.

To check whether this just represents a direction bias of the Brisbane Airport weather station, Fig. 13 c) shows the climatological perturbations at the nearby Spitfire Channel station (see Fig. 1). While the amplitude bias is smaller at Spitfire Channel than Brisbane Airport, the directional bias is at least as high. A similar directional bias is evident at the nearby Inner Beacon station (not shown), although the bias is smaller than at Spitfire Channel and Brisbane Airport. Similar biases are also evident at these stations in analogous figures for December, January and February 2017/18 (not shown), with the semi-major axis of the official forecast's ellipse fit oriented 29° clockwise from AWS's at Brisbane airport. Figure 1 shows there are two small islands to the east of Brisbane airport; the more north-northeasterly orientation of the Brisbane Airport sea-breeze suggests these

islands may be redirecting winds between the east coast of Brisbane and the west coasts of these islands, and that this local effect is not being captured in the official forecast, ACCESS or HRES.

Another example is the Hobart Airport station. Figure 12 c) shows that the semi-major axis of the AWS ellipse fit is oriented 31, 35 and 62 degrees anti-clockwise from the semi-major axes of the HRES, official forecast and ACCESS ellipse fits, respectively. Figures 11 a) and b) show that the ellipse fit for the AWS perturbations at Hobart airport only achieve R^2 values of 59% and 68% for the u and v components, respectively, but figures 13 d) and e) show that the fit still captures orientations accurately, although it underestimates the maximum AWS perturbation. Figure 13 f) provides the climatological perturbations at the Hobart (city) station, which also show a large difference in orientation between ACCESS and AWS. Given the timing of the westerly perturbations in ACCESS, and the fact that the prevailing winds around Tasmania are westerly, these results suggest that ACCESS is exaggerating the boundary layer mixing processes involved in the diurnal cycle around Hobart. These biases are not present during December, January and February 2017/18, as strong south to southeasterly sea-breeze perturbations are now dominant in all four datasets, although the semi-major axis of ACCESS's ellipse fit is still oriented 14 degrees clockwise to that of AWS.

At the South WA station group (not shown) the semi-major axes of the ACCESS and official forecast ellipse fits are oriented at least 49 degrees anti-clockwise from those of the AWS and HRES ellipse fits, and the HRES perturbations peak between 1.2 and 2.5 hours after the other datasets. These differences occur because eccentricity values are low for this station group, and Figure 9 b) shows that the westerly perturbations associated with boundary layer mixing are weaker for HRES than the other datasets. A similar issue affects the VIC station group, explaining why the semi-major axes of the AWS ellipse fit is oriented at least 49 degrees anti-clockwise from those of the other datasets.

The Darwin Airport, Darwin Airport station group, and NT station group (not shown) provide further examples. Here the ellipse fits produce favourable R^2 values, although the fits slightly underestimate the AWS max perturbation speed at the Darwin Airport station due to this dataset's highly asymmetric hodograph. At all three spatial scales there are timing differences between the perturbation maximums of up to 8.2 hours. These timing differences occur because for some scales and datasets, the later north to northwesterly sea-breeze perturbations dominate the diurnal wind cycle, but for other scales and datasets the earlier easterly to southeasterly boundary layer mixing effects dominate.

4. Synthesis

For land-sea breeze and boundary layer mixing edits to reduce absolute errors in the subsequent days wind forecast, these edits should reduce the absolute errors in the diurnal component of the wind fields. However, Figs. 2 and 6 indicate that this is only possible when absolute error is considered at coarse spatial scales, as at individual airport stations results are noisy and ambiguous, and over the intermediate city station group scale HRES outperforms the official forecast almost uniformly.

Taking the effective resolutions of the models considered in this study to be approximately $7\Delta x$ (e.g. Skamarock 2004; Abdalla et al. 2013), where Δx is the horizontal grid spacing, we have effective resolutions of ≈ 84 km and ≈ 63 km for ACCESS and HRES respectively. From resolution considerations alone, one might expect that forecaster edits would be able to reduce errors at the individual airport station scale, and the intermediate city station group scale (see Fig. 1), as motion at these scales is unresolved or only partially resolved by ACCESS and HRES.

To further investigate the effect of spatial scale on error, consider first just the zonal components of the AWS and official forecast wind perturbations, denoted by u_{AWS} and u_{O} respectively. Con-

sidering just the values at a particular hour UTC, over the entire June, July, August time period, the mean square error $\mathrm{mse}(u_{\mathrm{AWS}}, u_{\mathrm{O}}) = \overline{(u_{\mathrm{AWS}} - u_{\mathrm{O}})^2}$ can be decomposed $\mathrm{mse}(u_{\mathrm{AWS}}, u_{\mathrm{O}}) =$

$$\underbrace{\mathrm{var}(u_{\mathrm{AWS}}) + \mathrm{var}(u_{\mathrm{O}}) - 2\,\mathrm{cov}(u_{\mathrm{AWS}}, u_{\mathrm{O}})}_{\text{error variance}} + \underbrace{(\overline{u}_{\mathrm{AWS}} - \overline{u}_{\mathrm{O}})^2}_{\text{squared bias}} \tag{8}$$

where var, cov and the over-bar denote the sample variance, covariance and mean respectively. The first three terms are the variance of $u_{\mathrm{AWS}} - u_{\mathrm{O}}$, i.e. the error variance, and the last term is the square of the bias between u_{AWS} and u_{O}. Equation (8) can also be applied to the MSEs of HRES. Note that the mean square errors (MSEs) of the official forecast and HRES are closely related to $\overline{\mathrm{DAE}}_{\mathrm{OH}}$, which is the difference between the mean absolute errors of the official forecast and HRES; similarly, the squared bias components of the MSEs are closely related to $\mathrm{DB}_{\mathrm{OH}}$.

Figure 14 shows the terms of equation (8) for both the official forecast and HRES for Adelaide Airport, the Adelaide city station group, and the SA coastal station group. At all three scales the official forecast varies more than HRES, which is also the case at the other locations considered in this study. At Adelaide airport the variance of AWS is significantly larger than either the official forecast or HRES, but this additional variability is mostly uncorrelated to either dataset. This is unsurprising from representation considerations alone (e.g. Zaron and Egbert 2006), as the official forecast and HRES data represent averages over 6 km spatial grid-cells, whereas the AWS data represent point values. As a result, error variance terms are much larger than the squared bias terms, and of comparable magnitudes for both datasets. This is consistent with the comparatively noisy DAE results of Figs. 6 a) and b).

At the intermediate Adelaide city station group scale, the AWS variances are of similar magnitudes to those of HRES, but smaller than those of the official forecast, with the official forecast's additional variability mostly uncorrelated to AWS. This results in larger error variance terms for the official forecast, consistent with HRESs almost complete outperformance of the official fore-

cast in Figs. 6 c) and d). Over the coarse SA coastal station group scale, variances in all three datasets are now small enough that the error variance terms no longer dwarf the bias terms. Although the error variance of the official forecast is still larger than that of HRES, HRES's zonal biases at 05:00 UTC are now sufficient to result in a larger MSE at this time, consistent with the DAE results of Fig. 2 c) and d).

Analogous points can be made for the other locations considered in this study, the main exception being Darwin airport, Darwin city station group, and the NT coastal station group, where zonal biases in HREF around 01:00 - 03:00 UTC are large enough to overcome the official forecast's larger error variance, producing the results of Fig. 6 and Figs. 2 c) and d). The results of Fig. 6 c) and d) are therefore generally a consequence of the official forecast being more variable than HRES, with this additional variability mostly random, in the sense of being uncorrelated with AWS. Similarly, the official forecast is generally more variable than ACCESS, explaining why the official forecast also struggles to outperform ACCESS at these scales, and ACCESS is generally more variable than HRES, explaining why HRES generally outperforms ACCESS in the DAE results of Fig. 7. In the coastal station group DAE results of Fig. 2, the random variability in each dataset is reduced, and biases are now large enough to actually affect errors in the diurnal component of the forecast.

These results show that switching model guidance products or performing edits can add more random noise to the diurnal component of the official forecast than what can be offset by reductions in bias, or improved correlations with AWS. Because the official forecast is built from multiple model datasets, most commonly HRES and ACCESS, blending datasets with different means will tend to produce greater variance than any of the component datasets. If the choice of model guidance is made primarily on which model best captures more slowly evolving synoptic scale features, then switching model guidance may add random variability to the diurnal component of

the official forecast. Furthermore, unless all forecasters follow identical thought processes when making edits, the edits will also add random variability. It is less clear why ACCESS shows greater random variability than HRES: one cause may be ACCESS's shorter time-step.

These results have implications for forecasting practice. Model guidance products are indeed biased in how they resolve diurnal wind cycles (e.g. Fig. 13), and there is therefore scope for forecaster edits to reduce these biases. However, editing model guidance generally fails to reduce error in the forecast diurnal cycle, even at scales finer than the effective resolutions of the models, as the cycle itself is mostly hidden by random variability. Averaging over large areas reduces this random variability, and so biases have a greater impact on forecast error, but even at large scales Fig. 2 shows model guidance still outperforms the official forecast more often than not.

Reducing the random variability of the official forecast, or the model guidance datasets that comprise it, will therefore improve the capacity of these types of edits to reduce error. One way to do this would be to move to an ensemble forecasting system, another would be to post process model guidance products, such as by averaging multiple time steps around the hour, before including them in GFE.

5. Conclusion

In this study we have presented methods for verifying the diurnal component of wind forecasts, with the intended application being the assessment of the edits Australian forecasters make to model guidance datasets in order to better resolve land-sea breeze and boundary layer mixing processes. We considered both errors and seasonal biases at each hour UTC, over three spatial scales, but the methods are immediately generalisable to other spatiotemporal scales.

When the methods are applied to Australian forecast data, the results indicate that the official edited forecast only produces lower absolute errors in the diurnal wind cycle when averaged over

coarse spatial scales of $500 \times 150\ \text{km}^2$ to $2000 \times 150\ \text{km}^2$: this scale corresponds to the aggregation of data within 150 km of the Australian coastline, subdivided into linear segments of coastline and by state (see Fig. 1). Even at these scales, reductions in error are isolated to particular locations and times of day, and the official forecast rarely has lower mean absolute error than both commonly used model guidance products simultaneously. This suggests that forecaster skill in improving diurnal wind processes lies more in making the choice of model guidance than in making edits.

By contrast, the official forecast can produce lower seasonal biases than model guidance at all three spatial scales, but again, it rarely produces lower biases than both standard model guidance products simultaneously. Reduced seasonal biases do not translate into reduced errors at the two smaller spatial scales because the diurnal cycle is mostly masked by the random variability in each dataset. Furthermore, because the official forecast exhibits much greater random variability than HRES, HRES almost uniformly outperforms the official forecast over the intermediate 50×50 km^2 to 200×200 km^2 spatial scale. The same is true for ACCESS, although to a slightly lesser extent, and also explains why HRES mostly outperforms ACCESS at this scale.

We also compare structural features of the diurnal wind cycles of each dataset by fitting modified ellipses to hodographs of seasonally averaged diurnal wind cycles, then deriving metrics from these ellipses. This approach reveals structural biases in the official forecast, including directional biases in the approach of the sea-breeze at Brisbane airport, eccentricity biases along the coast of NSW, and amplitude biases along the southwest coast of WA. It also reveals biases in model guidance datasets, such as ACCESS's overemphasis of boundary layer mixing processes around Hobart.

Future research could extend this study in multiple directions. One important question is whether the random variability in the official forecast, or the model guidance products that comprise it, could be reduced through ensemble forecasting or post-processing, as reducing random variability

would both decrease errors, and increase the value of land-sea breeze and boundary layer mixing edits. Another goal could be to identify precisely the spatiotemporal scales at which diurnal wind cycles can be identified against background noise, so as to better understand the scales at which land-sea breeze and boundary layer mixing edits can add value to a forecast.

Acknowledgments. Funding for this study was provided for Ewan Short by the Australian Research Council's Centre of Excellence for Climate Extremes (CE170100023). Datasets and software were generously provided by the Australian Bureau of Meteorology's Evidence Targeted Automation team. Thanks are due to Michael Foley, Deryn Griffiths, Nicholas Loveday, Ben Price and Alexei Hider for providing support at the Bureau of Meteorology's Melbourne and Darwin offices, and to Professors Craig Bishop and Todd Lane from the University of Melbourne, and Carly Kovacik from the United States' National Weather Service, for some helpful conversations.

References

Abdalla, S., L. Isaksen, P. A. E. M. Janssen, and N. Wedi, 2013: Effective spectral resolution of ECMWF atmospheric forecast models. 19–22, doi:10.21957/rue4o7ac, [Available online at https://www.ecmwf.int/node/17358 - Accessed 07/12/2019].

Abkar, M., A. Sharifi, and F. Port-Agel, 2016: Wake flow in a wind farm during a diurnal cycle. *Journal of Turbulence*, **17 (4)**, 420–441, doi:10.1080/14685248.2015.1127379.

Bureau of Meteorology, 2010: Operational implementation of the ACCESS numerical weather prediction systems. Tech. Rep. NMOC Operations Bulletin No. 83, Bureau of Meteorology, Melbourne, Victoria. [Available online at http://www.bom.gov.au/australia/charts/bulletins/apob83.pdf - Accessed 25/04/2019].

Bureau of Meteorology, 2016: APS2 upgrade to the ACCESS-R numerical weather prediction system. Tech. Rep. BNOC Operations Bulletin No. 104, Bureau of Meteorology, Melbourne, Victoria. [Available online at http://www.bom.gov.au/australia/charts/bulletins/apob107-external.pdf - Accessed 25/04/2019].

Bureau of Meteorology, 2019: Meteye. Bureau of Meteorology, [Available online at http://www.bom.gov.au/australia/meteye/].

Dai, A., and C. Deser, 1999: Diurnal and semidiurnal variations in global surface wind and divergence fields. *Journal of Geophysical Research*, **104**, 31 109–31 125, doi:10.1029/1999JD900927.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteor. Appl.*, **15 (1)**, 51–64, doi:10.1002/met.25.

Efron, B., 1979: Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7 (1)**, 1–26, doi:10.1214/aos/1176344552.

Englberger, A., and A. Dörnbrack, 2018: Impact of the diurnal cycle of the atmospheric boundary layer on wind-turbine wakes: a numerical modelling study. *Boundary-Layer Meteorology*, **166 (3)**, 423–448, doi:10.1007/s10546-017-0309-3.

European Center for Medium Range Weather Forecasting, 2018: *Part IV: Physical processes*, 223. No. 4, IFS Documentation, European Center for Medium Range Weather Forecasting, [Available online at https://www.ecmwf.int/node/18714 - Accessed 25 April 2019].

Gille, S. T., S. G. Llewellyn Smith, and N. M. Statom, 2005: Global observations of the land breeze. *Geophysical Research Letters*, **32 (5)**, doi:10.1029/2004GL022139.

552 Griffiths, D., H. Jack, M. Foley, I. Ioannou, and M. Liu, 2017: Advice for automation of forecasts:
553 a framework. Tech. rep., Bureau of Meteorology, Melbourne, Victoria. [Available online at
554 http://www.bom.gov.au/research/publications/researchreports/BRR-021.pdf].

555 Lee, X., 2018: *Fundamentals of boundary-layer meteorology.* Springer atmospheric sciences,
556 Springer.

557 Lock, A. P., A. R. Brown, M. R. Bush, G. M. Martin, and R. N. B. Smith, 2000: A new bound-
558 ary layer mixing scheme. Part I: scheme description and single-column model tests. *Monthly*
559 *Weather Review*, **128 (9)**, 3187–3199, doi:10.1175/1520-0493(2000)128⟨3187:ANBLMS⟩2.0.
560 CO;2.

561 Louis, J.-F., 1979: A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer*
562 *Meteorology*, **17 (2)**, 187–202, doi:10.1007/BF00117978.

563 Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution
564 produce more skillful forecasts? *Bulletin of the American Meteorological Society*, **83 (3)**, 407–
565 430, doi:10.1175/1520-0477(2002)083⟨0407:DIHRPM⟩2.3.CO;2.

566 Miller, S. T. K., B. D. Keim, R. W. Talbot, and H. Mao, 2003: Sea breeze: Structure, forecasting,
567 and impacts. *Reviews of Geophysics*, **41 (3)**, doi:10.1029/2003RG000124.

568 Modigliani, U., and C. Maass, 2017: Detailed information of implementation of IFS cy-
569 cle 41r2. ECMWF, [Available online at https://confluence.ecmwf.int/display/FCST/Detailed+
570 information+of+implementation+of+IFS+cycle+41r2 - Accessed 27/04/2019].

571 Physick, W. L., and D. J. Abbs, 1992: Flow and plume dispersion in a coastal valley. *Journal*
572 *of Applied Meteorology*, **31 (1)**, 64–73, doi:10.1175/1520-0450(1992)031⟨0064:FAPDIA⟩2.0.
573 CO;2.

Pinson, P., and R. Hagedorn, 2012: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. *Meteor. Appl.*, **19 (4)**, 484–500, doi:10.1002/met.283.

Rife, D. L., and C. A. Davis, 2005: Verification of temporal variations in mesoscale numerical wind forecasts. *Monthly Weather Review*, **133 (11)**, 3368–3381, doi:10.1175/MWR3052.1.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review*, **132 (12)**, 30193032, doi:10.1175/MWR2830.1, URL https://doi.org/10.1175/MWR2830.1, https://doi.org/10.1175/MWR2830.1.

Svensson, G., and Coauthors, 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 Meteorology*, **140 (2)**, 177–206, doi:10.1007/s10546-011-9611-7.

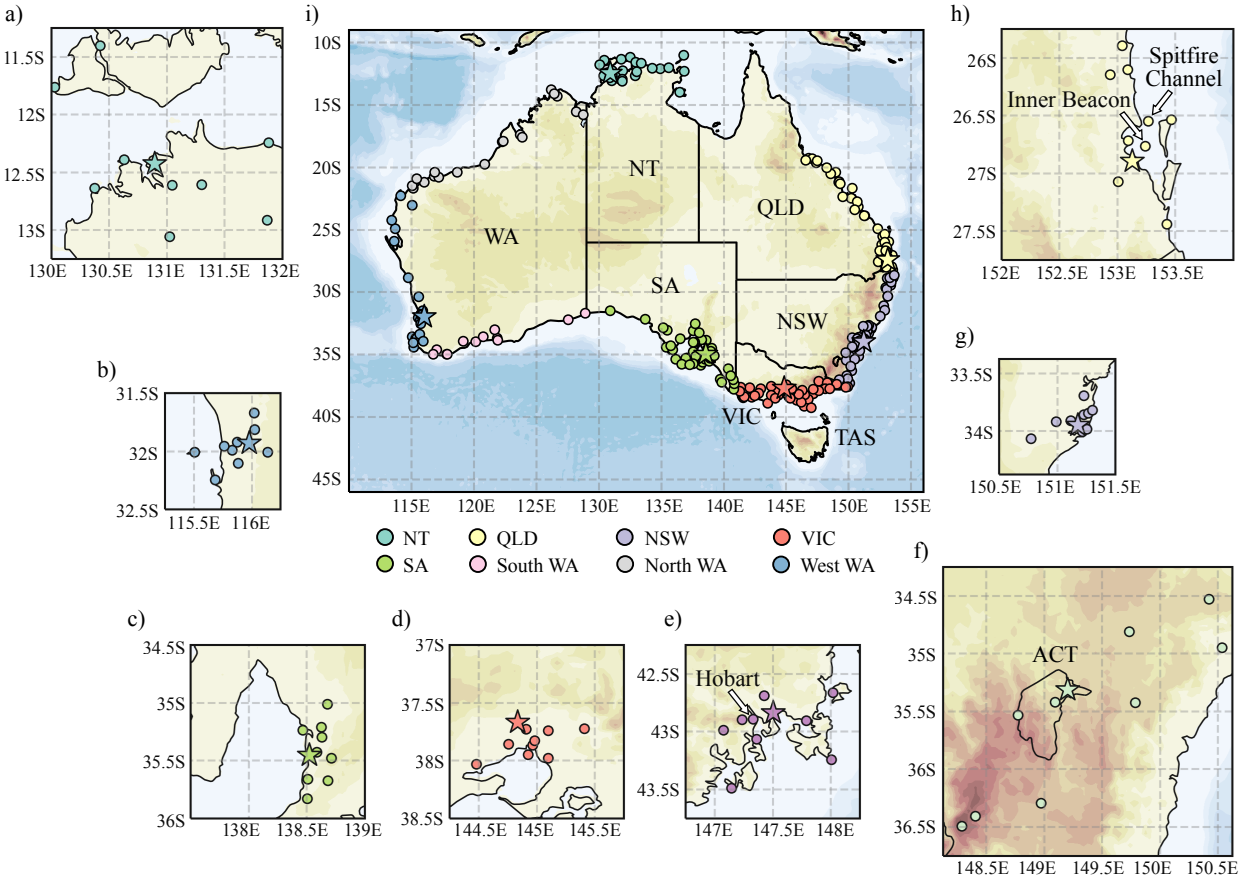
Vincent, C. L., and T. P. Lane, 2016: Evolution of the diurnal precipitation cycle with the passage of a Madden-Julian Oscillation event through the Maritime Continent. *Monthly Weather Review*, **144 (5)**, 1983–2005, doi:10.1175/MWR-D-15-0326.1.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*. International geophysics series: v. 100, Elsevier.

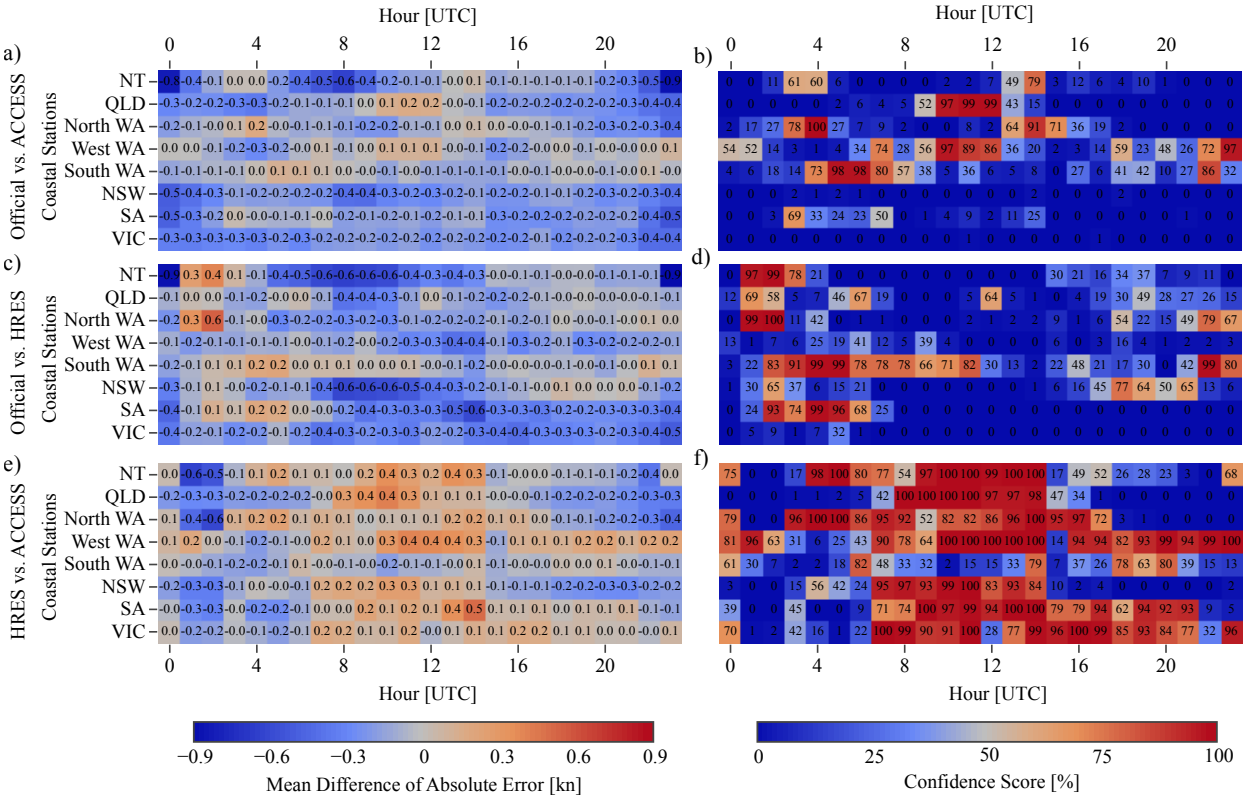
Zaron, E. D., and G. D. Egbert, 2006: Estimating open-ocean barotropic tidal dissipation: The hawaiian ridge. *Journal of Physical Oceanography*, **36 (6)**, 1019–1035, doi:10.1175/JPO2878.1.

Zwiers, F. W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *Journal of Climate*, **8 (2)**, 336–351, doi:10.1175/1520-0442(1995)008⟨0336:TSCIAI⟩2.0.CO;2.

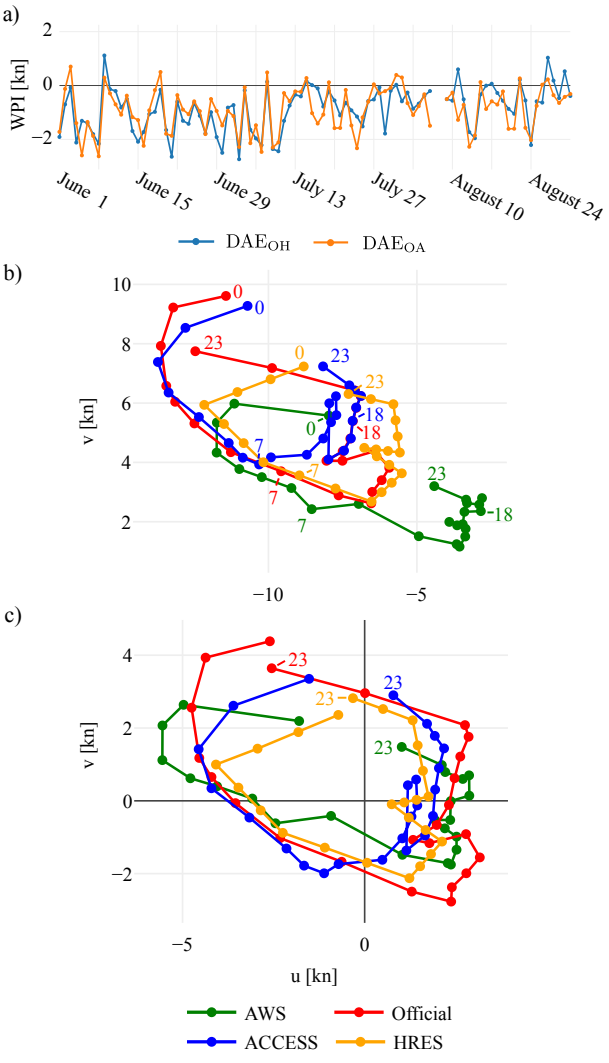
LIST OF FIGURES



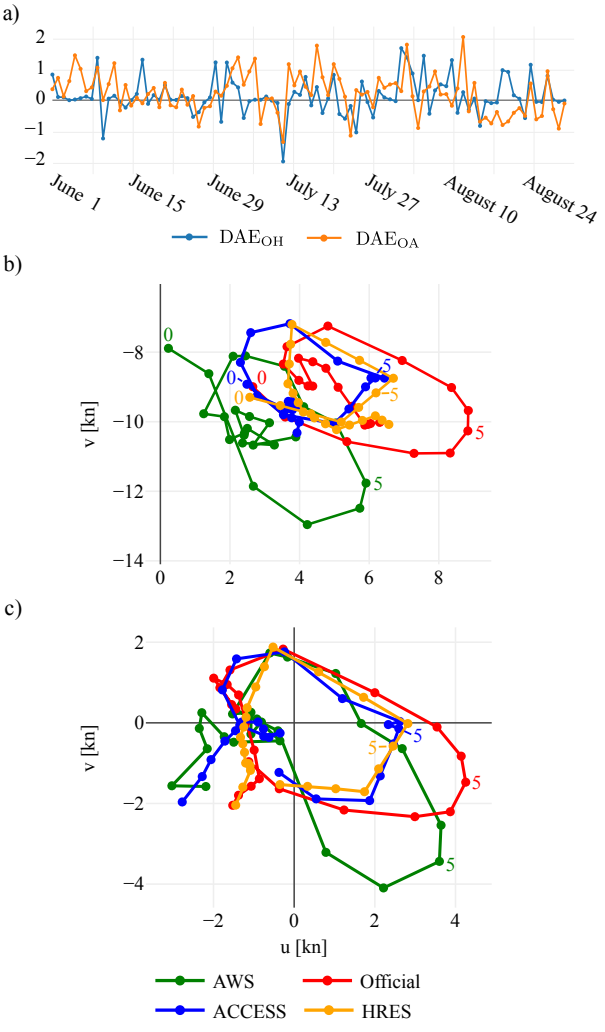
649 FIG. 1. Locations of the automatic weather stations considered in this study, where stars give the locations
650 of the capital city *airport stations*. Stations are divided into the Darwin, Perth, Adelaide, Melbourne, Hobart,
651 Canberra, Sydney and Brisbane *city station groups*, a) to h), respectively, and the *coastal station groups*, i).
652 Height and depth shading intervals every 200 and 1000 m, respectively.



653 FIG. 2. Heatmaps of mean difference of absolute error $\overline{\text{DAE}}$ values, a), c), e), and confidence scores, b), d),
654 f), for each coastal station group (see Fig. 1) and hour of the day, for the official forecast versus ACCESS, a) and
655 b), official forecast versus HRES, c) and d), HRES versus ACCESS, e) and f). Positive $\overline{\text{DAE}}$ values indicate that
656 the former dataset in each pair is on average $\overline{\text{DAE}}$ kn closer to observations than the latter dataset (see equation
657 1), where 1 kn ≈ 0.514 m s^{-1}. Confidence scores provide the probability the population or "true" value of $\overline{\text{DAE}}$
658 is greater than zero (see section 2).



659 FIG. 3. Time series, a), of the difference in absolute error DAE defined in equation (1) for the official forecast
660 versus ACCESS, DAE_{OA}, and official forecast versus HRES, DAE_{OH}, for the NT coastal station group shown in
661 Fig. 1 at 23:00 UTC. Also, temporal hodographs in hours UTC showing hourly changes in winds, b), and wind
662 perturbations from a 24 hour running mean, c), at the NT coastal station group on the 3rd of July 2018.



663 FIG. 4. As in Fig. 3, but for, a), the South WA coastal station group at 05:00 UTC, and b) and c), the winds
664 and wind perturbations, respectively, over the South WA coastal station group on the 9th June 2018.

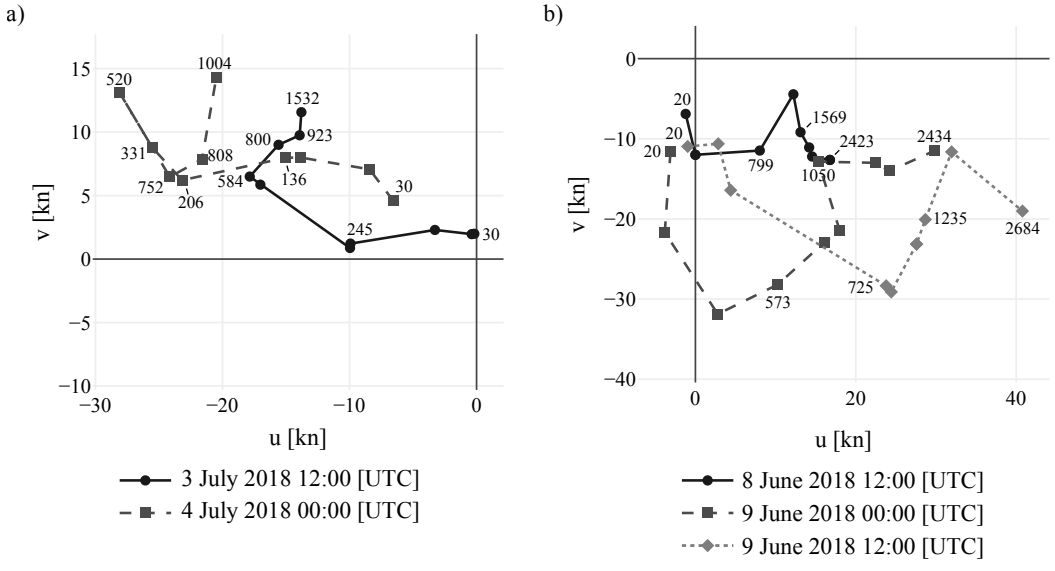
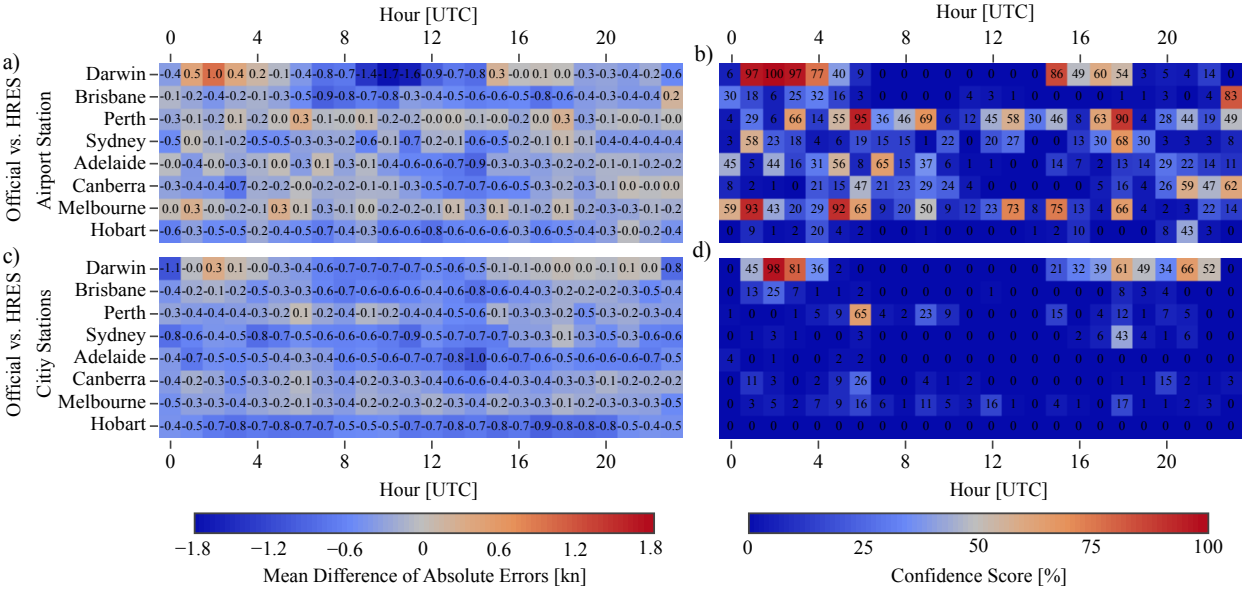
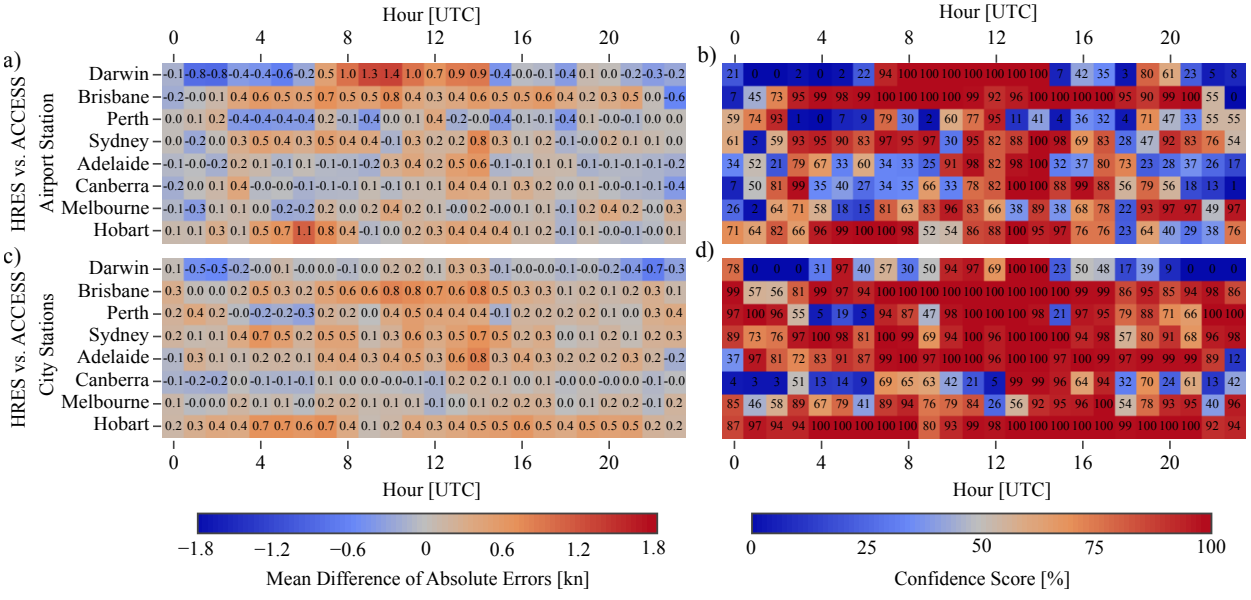


FIG. 5. Vertical wind soundings at, a), Darwin Airport, and b), Perth Airport, with heights given in metres.



665 FIG. 6. As in Fig. 2, but for the official versus HRES mean difference of absolute error $\overline{\mathrm{DAE}}_{\mathrm{OH}}$ values, a) and
666 c), and confidence scores, b) and d), for the airport stations, a) and b), and city station groups, c) and d).



667 FIG. 7. As in Fig. 6, but for the HRES versus ACCESS mean difference in absolute error $\overline{\text{DAE}}_{\text{HA}}$ values and
668 confidence scores.

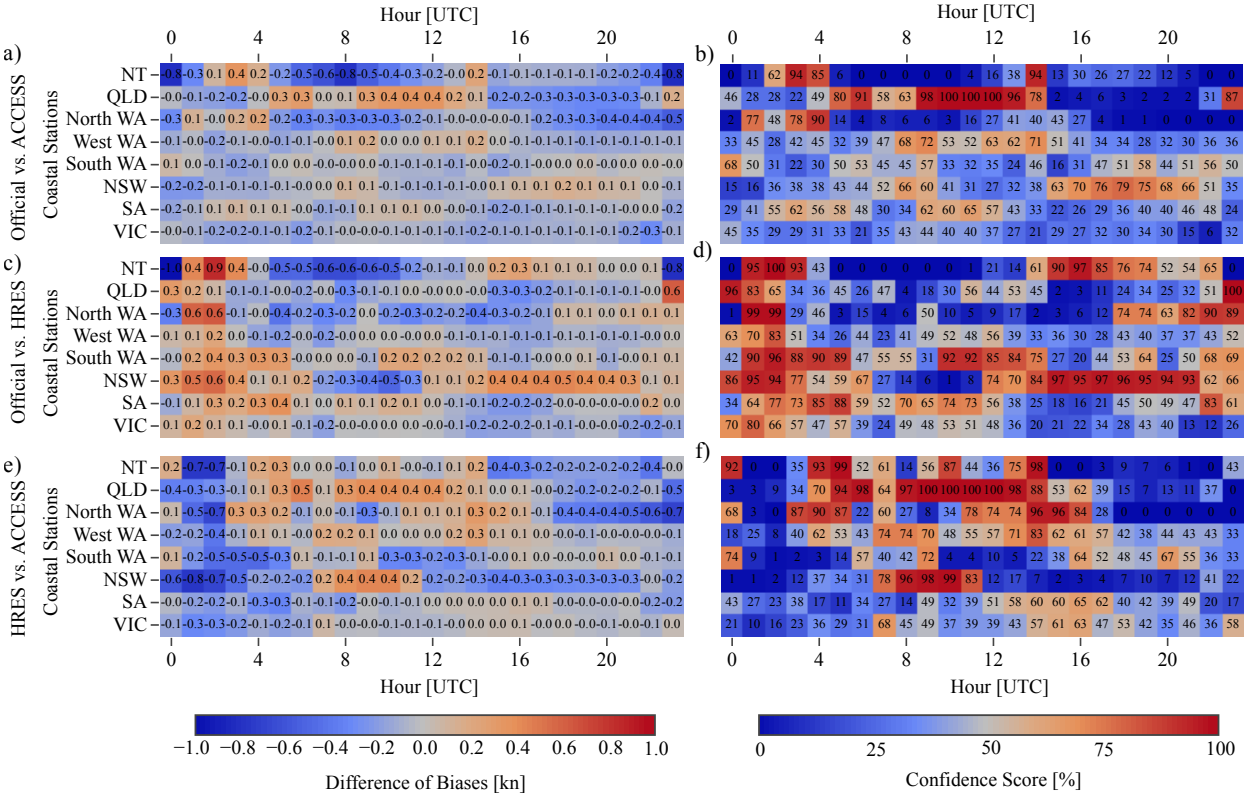
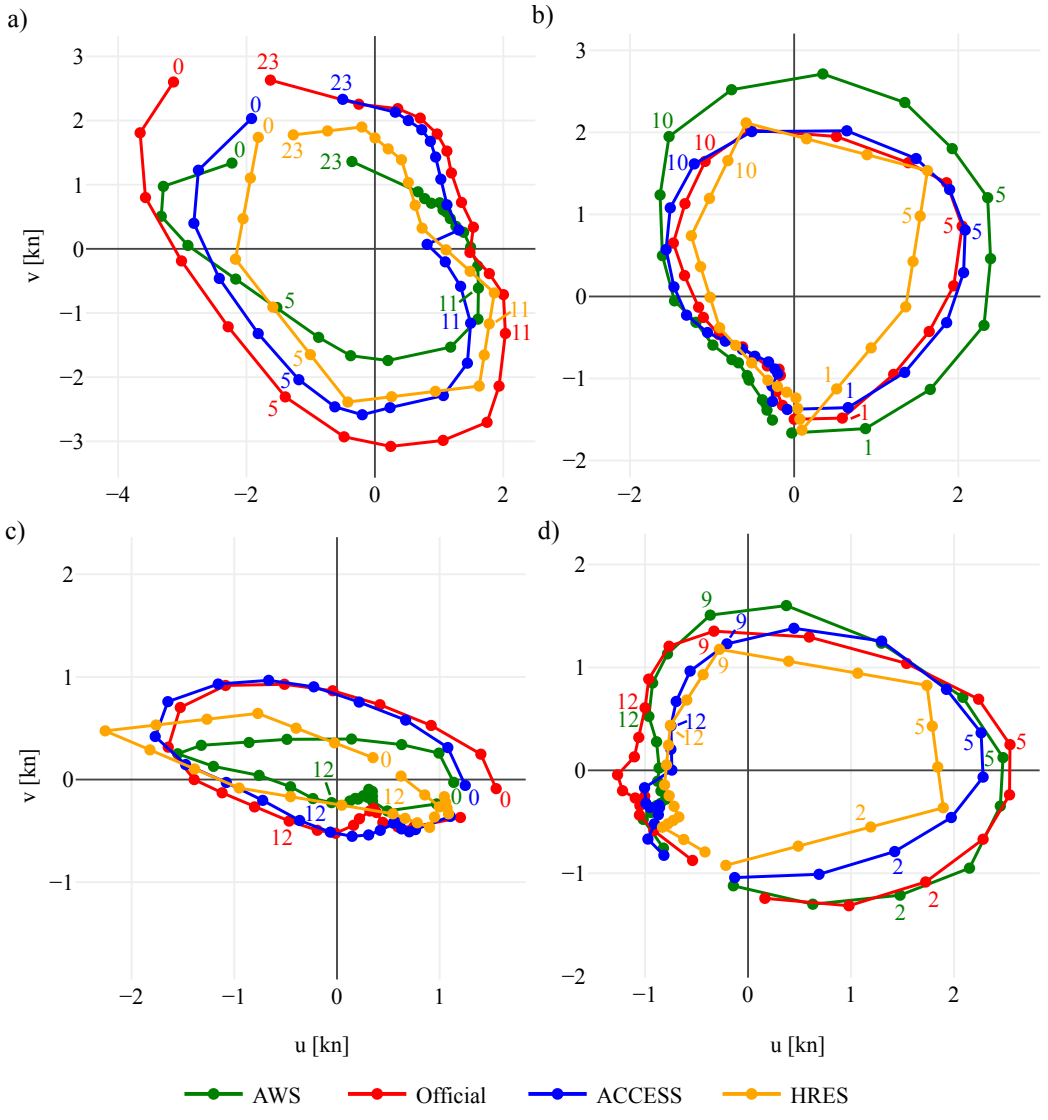


FIG. 8. As in Fig. 2, but for the difference of biases (DB) values and confidence scores.



669 FIG. 9. Temporal hodographs in hours UTC of wind perturbations spatially averaged over the, a), NT, b)
670 South WA, c) NSW and d), SA coastal station groups (see Fig. 1), and temporally averaged over June, July and
671 August 2018.

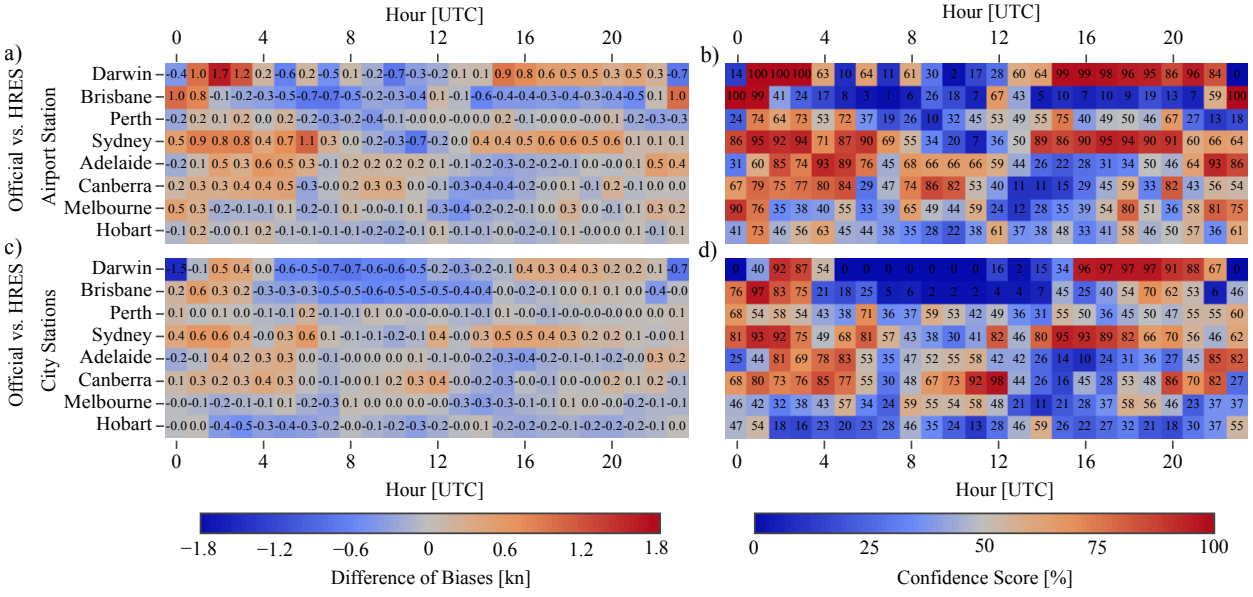
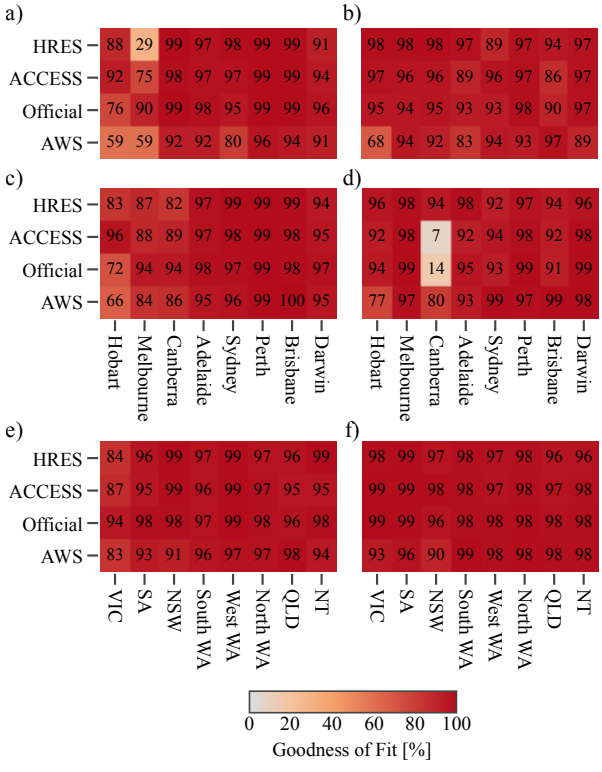


FIG. 10. As in Fig. 6, but for the difference of biases (DB) values and confidence scores.



672 FIG. 11. R^2 values as percentages for the fit of equation (5) to the zonal perturbations, a), c) and e), and
673 equation (6) to the meridional perturbations, b), d) and f), for the airport stations, a) and b), city station groups,
674 c) and d), and coastal station groups, e) and f), shown in Fig. 1.

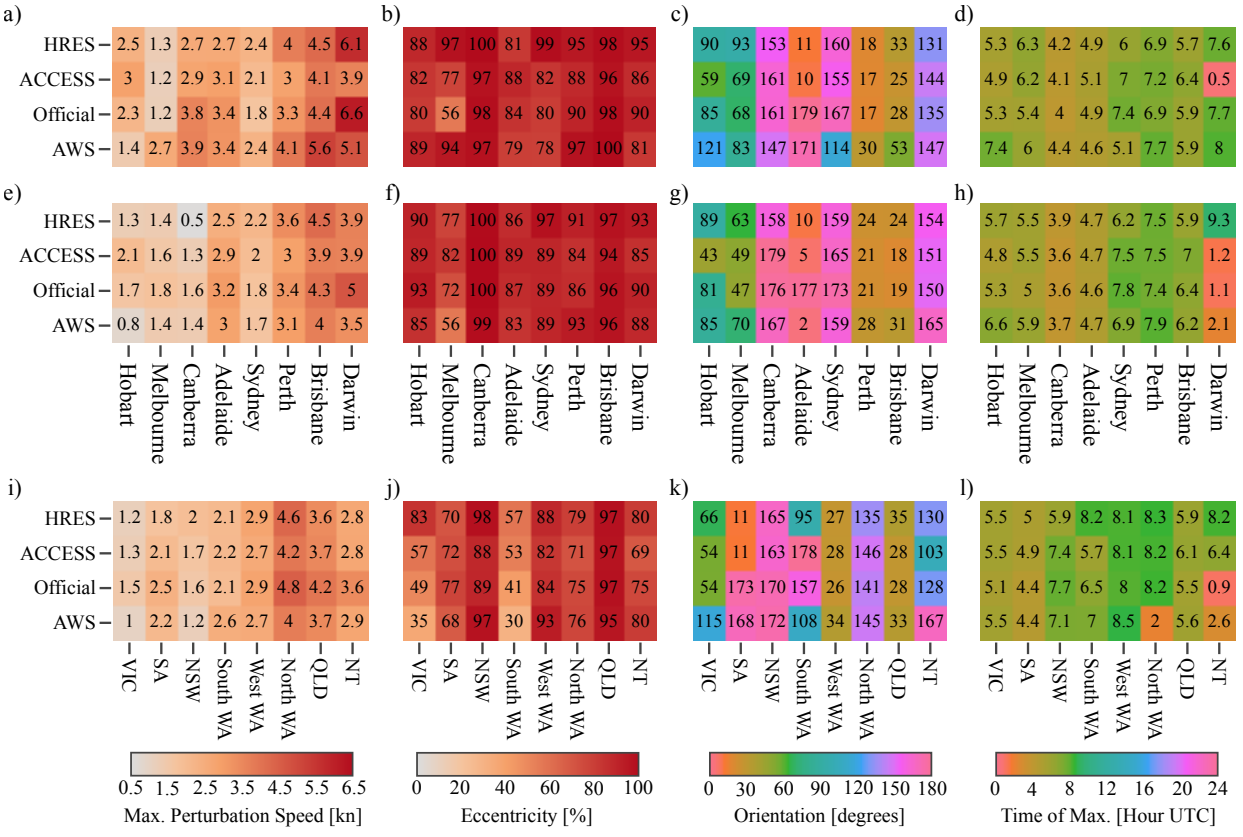


FIG. 12. Metrics derived from fitting ellipse equations (5) and (6) to wind perturbations at the Australian capital city airport stations, a) to d), and to wind perturbations spatially averaged over the city station groups and coastal station groups shown in Fig. 1, e) to h) and i) to l) respectively, with perturbations also temporally averaged over June, July and August 2018 in each case. Metrics given are the maximum perturbation speed, a), e) and i), eccentricity of fitted ellipse, b), f) and j), orientation semi-major axis makes with lines of latitude, c), g) and k), and time of maximum perturbation, d), h) and l).

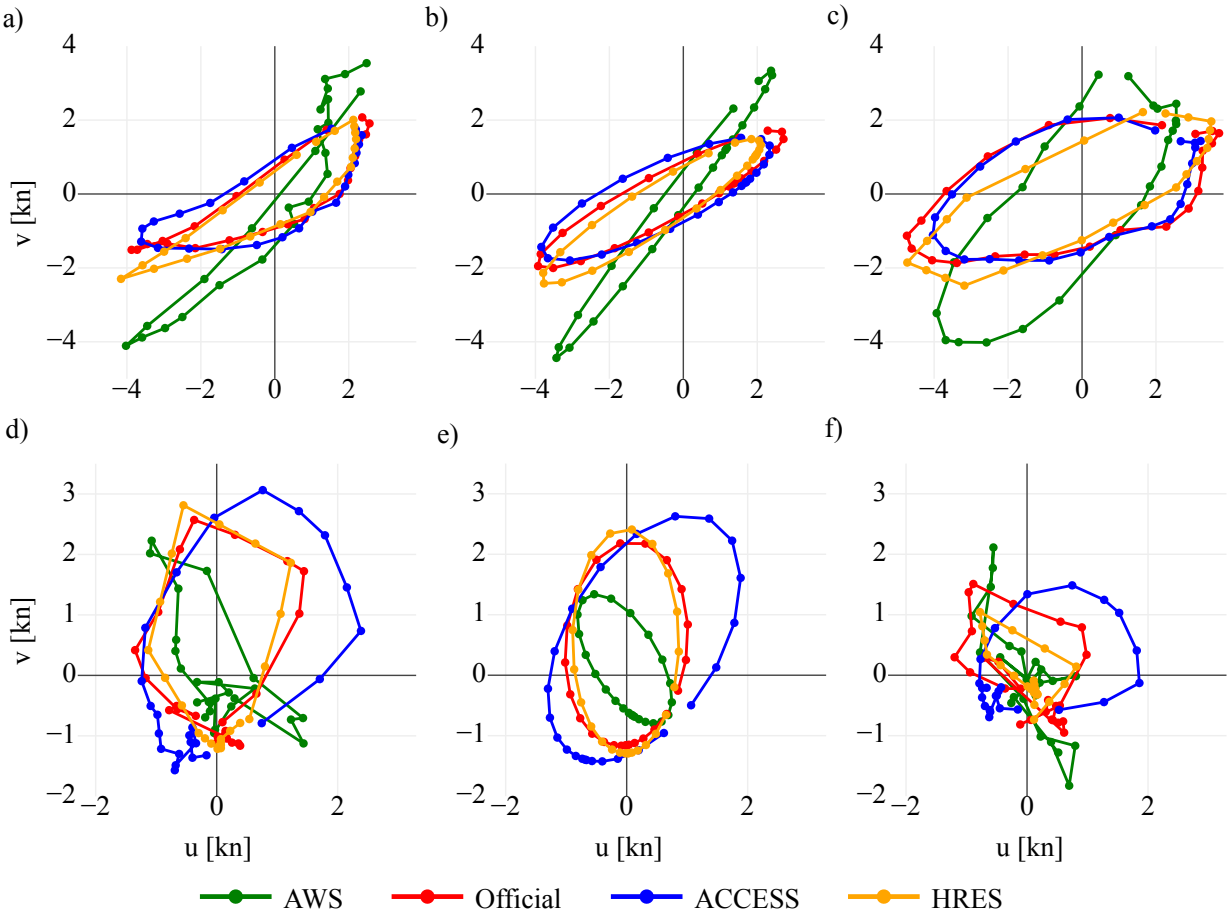


FIG. 13. Temporal hodographs of wind perturbations at each hour UTC averaged over June, July and August 2018, at Brisbane and Hobart airports, a) and d), and the associated ellipse fits, b) and e). For comparison, c) and f) provide the hodographs of the averaged perturbations at the Spitfire Channel and Hobart city stations, respectively (see Fig. 1).

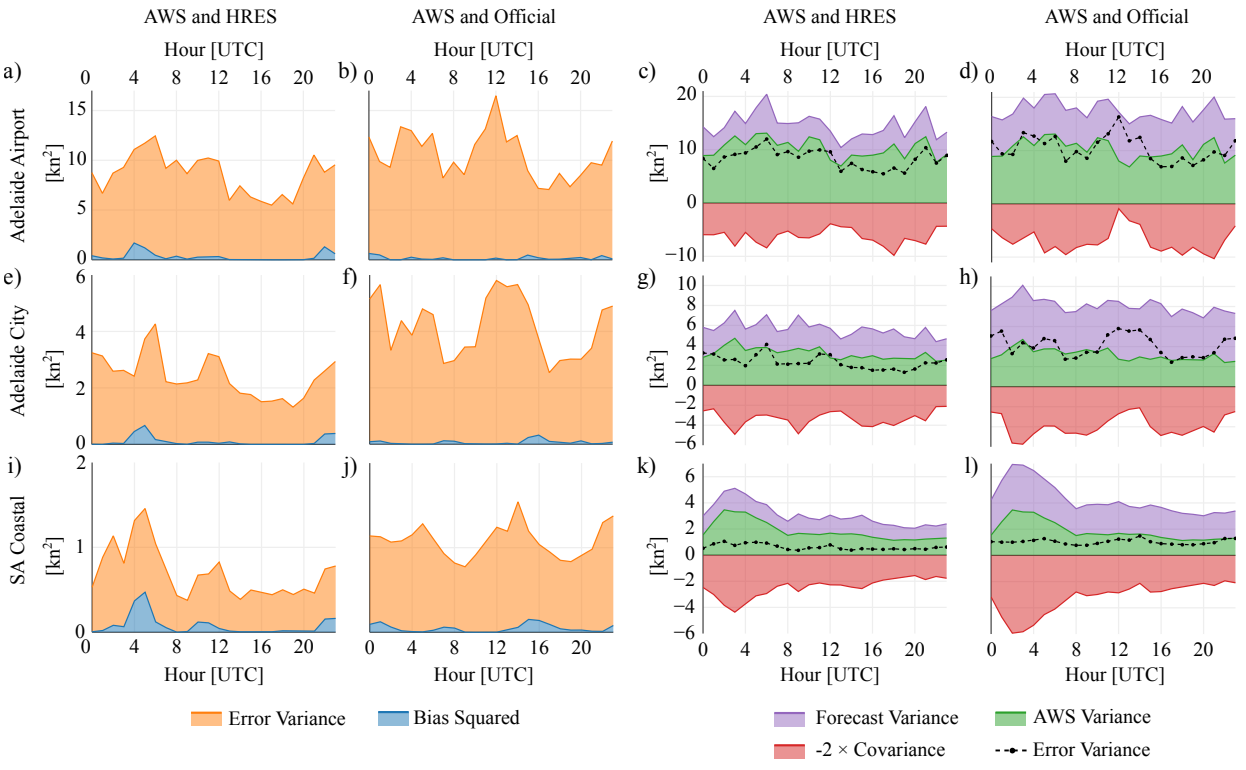


FIG. 14. Mean square error between the AWS and HRES zonal perturbations $\overline{(u_{\mathrm{AWS}} - u_{\mathrm{H}})^2}$, a), e), and i), decomposed into the error variance $\mathrm{var}\,(u_{\mathrm{AWS}} - u_{\mathrm{H}})$ and squared bias $(\overline{u}_{\mathrm{AWS}} - \overline{u}_{\mathrm{H}})^2$ terms of equation (8). Also, the decomposed mean square error between the AWS and official forecast zonal perturbations, b), f) and j). Additionally, the HRES and AWS error variance term $\mathrm{var}\,(u_{\mathrm{AWS}} - u_{\mathrm{H}})$ decomposed into the $\mathrm{var}\,(u_{\mathrm{AWS}})$, $\mathrm{var}\,(u_{\mathrm{H}})$ and $-2 \cdot \mathrm{cov}\,(u_{\mathrm{AWS}}, u_{\mathrm{H}})$ terms, c), g) and k), and analogously for the official forecast and AWS error variance term $\mathrm{var}\,(u_{\mathrm{AWS}} - u_{\mathrm{O}})$, d), h) and l). Decompositions given for Adelaide Airport, a) to d), the Adelaide city station group, e) to h), and the SA coastal station group, i) to l) (see Fig. 1.)