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.

Ben ?. Price

Bureau of Meteorology, Casuarina, Northern Territory, Australia

Derryn ?. Griffiths and Alexei ?. Hider

Bureau of Meteorology, 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 are based on model data that is edited by human forecasters. Two types of edits are commonly made to the wind fields: these aim to improve how boundary layer mixing processes and the land-sea breeze are resolved in the forecast. In this study we compare the diurnally varying component of the edited wind forecast, with those of station observations and unedited model guidance datasets, to assess the additional accuracy provided by the edits. We consider coastal locations across Australia over June, July and August 2018, assessing performance at three different spatial scales, and on both a daily and seasonal basis. The results show that the edited forecast generally only produces lower daily errors than model guidance at the coarsest spatial scale (1000 - 2000 km), but can achieve lower seasonal biases over all three spatial scales. However, the edited forecast only reduces errors at particular times and locations, and rarely produces lower errors than all model guidance products simultaneously. This suggests that forecaster skill lies mostly in making the choice of model guidance, rather than in making edits. To better diagnose the causes of errors in the diurnal wind cycles, we fit a modified ellipse to the climatological diurnal cycle hodographs. Performance varies 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. Is this also how things work at the U.S National Weather Service and U.K. Met Office? 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 synoptic scale model resolutions, or to correct known biases of the models being used. The resulting gridded forecast datasets are then provided to the public through the BoM's online MetEye data browser (Bureau of Meteorology 2019); the gridded forecast datasets are also translated into text and icon forecasts algorithmically.

Australian forecasters generally make two types of edits to the surface wind fields on a routine daily basis. The first is to edit the surface winds after sunrise at locations where the forecaster believes the model guidance is providing a poor representation of boundary layer mixing processes. Bounday 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. How do the boundary layer mixing tools in GFE currently work? While I was in Darwin you picked a height z and a percentage p, and the tool essentially formed an average of the surface winds and winds at x 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.

How do the sea-breeze tools in GFE currently work? While I was in Darwin you traced out the relevant coastline graphically, chose a wind speed and a time, and GFE would add in winds perpendicular to the traced coastline at this speed, and smoothly blend them in spatially and temporally.

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. Dai and Deser (1999) fitted the first two harmonics to seasonal averages of wind speed at different times of day, and showed that over land surfaces the average amplitude of the wind speed diurnal cycle varied from 1.2 to 2.1 kn, (knots are used throughout this paper because this is the unit forecasters work with, and the unit that is used in Jive) and that the fitted harmonics accounted for 50 to 70% of the daily variability. Table 1 provides the mean wind speeds for the eight Australian capital city airport stations shown in Fig. 1, over December, January, February 2017/18, and June, July and August 2018, suggesting that the amplitude of the mean diurnal cycles are approximately 10 to 34% of the mean wind speeds across Australia.

Beyond their contribution to the overall wind field, 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.

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 sim-

ulation (LES), and observations at Kansas, USA, during October 1999. They found that both the models and LES failed to capture the sudden ≈ 6 kn 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 resolved by 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 at ECMWF's coarse 50km spatial resolution.

Any attempt to validate model data against observations must confront the *representation problem* (e.g. Zaron and Egbert 2006). Because models cannot resolve physical processes occurring at sub-grid scales, a value predicted by an operational model for a given grid-cell must be interpreted as a prediction of the filtered, or Reynolds averaged value over that grid-cell. Therefore, comparing model data with observational data can be an unfair test of model performance, and for this reason model forecasts are often verified against reanalysis hind-casts that use the same model (e.g. Lynch et al. 2014).

However, the way the representation problem applies to the verification of forecasts issued to the public is more nuanced. In this case, a forecast issued by a national weather service is attempting to represent either reality itself, or the filtered version of reality *that is of interest to the end users.* Thus, Pinson and Hagedorn (2012) disregarded the representation problem entirely, arguing that the end user is not interested in spatiotemporal scales of models, only the "best forecast" at the time and place of their choice. However, different users will have different ideas about what the "best" forecast entrails. Some users may desire a forecast that minimises error between the forecast and observations, others a forecast that most accurately reproduces the observed wind

speed distribution, regardless of temporal details. Ideally, operational forecasts should therefore be assessed according to the specific representation needs of particular end users.

Note that BoM verifies its wind forecasts by comparing hourly forecast data with station observations that are first averaged 10 minutes either side of the hour: implicit in this practice is the assumption that end users do not care about wind turbulence at temporal scales less than 20 minutes. Furthermore, the fact that the BoMs forecast is formed from model datasets with different resolutions, and the choice of model guidance can change even over the course of a single day, (e.g. Fig. 3 b), means it is difficult to determine precisely what BoM forecasts intend to represent, and hence addressing the representation problem is difficult. Note that clicking locations on the MetEye map seems to bring up the forecast for the nearest station or population centre. Does this imply that the Official forecast is intending to represent spatial scales at least as fine as the distance between stations? Or is it intending to represent averaged parameters at the ACCESS-R or ACCESS-C resolutions? Note also that grids are only provided every 3 hours in MetEye; do these grids represent the hourly values from the Official forecast, provided just at these three hour intervals, or a three hourly average?

Related to the representation problem is the question of how mesoscale models, which run at spatial resolutions of between 1 and 10 km, should be verified. BoM now regularly runs mesoscale models over some Australian capital cities as part of its daily forecasting routine, and the edits performed by human forecasters are also often mesoscale in nature. Note that mesoscale models resolve topography and its effects on the atmosphere in more detail, and explicitly simulate most convective and boundary layer processes. In this sense they are more realistic than coarser scale models, although they can actually perform worse than coarse models on standard verification scores whenever there are timing or location differences between features in the models and in observations. Mass et al. (2002) found that in the northwestern United States, mean square errors

in forecasts produced using mesoscale models decreased with resolution down to 10 to 15 km, whereas in the eastern United States where the topography is much flatter, this threshold was considerably larger, at 20 to 40 km.

Mass et al. (2002) therefore argued that existing verification approaches needed reform, suggesting that verification could instead be performed on spatially or temporally averaged parameters, an approach now known as *upscaling* (Ebert 2008). Alternatively, Mass et al. (2002) argued that "feature based" identification metrics be developed, which reward models for realistically simulating atmospheric features, even if the timing or location of these features is incorrect. Rife and Davis (2005) developed such a method for the verification of surface winds, defining a wind "object" as a wind change of at least one standard deviation occurring within a 12 hour interval, then assessing whether a mesoscale model could replicate the "objects" present in observations.

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 an increase in accuracy, and to do so in a way that respects the representation and mesoscale verification challenges discussed above. Second, to apply this methodology across Australian coastal locations to better understand the performance of 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, 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 model data from ECMWF, (is this the mean of the ECMWF operational ensemble?) and unedited model data from the Australian Community Climate and Earth System Simulator (ACCESS). ECMWF and ACCESS are two of the model guidance products most commonly used by Australian forecasters. (I haven't considered the BoM's Operational Consensus Forecast in this study, because it was not used in the Official forecast for winds while I was in the NT. If it is used frequently in other states, the study can be reinterpreted as a verification of both forecaster edits and OCF against unedited model guidance. Also, with the Bureau's permission, I would like to make the datasets used in this paper open access, and host them on, for instance, the NCI catalogue page.) The Official and ECMWF data are at ? and ? degree spatial resolutions respectively. What are the resolutions of these datasets as they're used in Jive? I know ACCESS uses nested grids, so what is the resolution of the ACCESS dataset when used in GFE and Jive? Are the outer grids interpolated to the resolution of the inner grids, or are the inner grids upscaled? Official, ACCESS and AWS data exists at each UTC hour, but ECMWF data only exists at a three hour resolution. Why is this? What are the actual time-steps of the models? To be consistent with the other data sets, ECMWF is therefore linearly interpolated to an hourly resolution: note that this is also what happens when forecasters load ECMWF wind data into the GFE, and the linearly-interpolated ECMWF data is therefore the appropriate model guidance dataset to compare with the Official forecast. To facilitate comparison with observations, Official, ACCESS and ECMWF

data is (tri-linearly?) interpolated in all three spatial dimensions to the locations of the weather stations. AWS wind data is recorded every minute at each station, and the hourly AWS datasets used in this study are formed by taking 10 minute averages either side of each UTC hour. (My memory is that Jive uses a ten minute average either side, but need to confirm this as it could be five minutes either side.) Stations are quality controlled by... I've excluded the list of known problematic wind stations in the Jive documentation, but I can't remember what the other quality control methods were.

Both ACCESS and ECMWF 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). ECMWF use similar schemes that they develop in-house (European Center for Medium Range Weather Forecasting 2018). This study 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 developments to the ACCESS and ECMWF models.

b. Assessing Diurnal Cycles

Forecasters typically edit model guidance wind data to account for under-resolved sea-breezes 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 hour centred running mean *background wind* from each zonal and meridional hourly wind data point. This provides a collection of zonal and meridional wind *perturbation* datasets.

One measure of the accuracy of the Official, ACCESS and ECMWF diurnal cycles is to compare the Euclidean distances of the perturbations at each hour with the corresponding AWS per-

turbations. For example, to assess whether the Official forecast perturbations, $\boldsymbol{u}_{\mathrm{O}}$, or ACCESS perturbations, $\boldsymbol{u}_{\mathrm{A}}$, best match the AWS observations, $\boldsymbol{u}_{\mathrm{AWS}}$, we calculate the *Wind Perturbation Index* (WPI), defined by

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

The analogously defined quantities $\mathrm{WPI}_{\mathrm{OE}}$ and $\mathrm{WPI}_{\mathrm{EA}}$ can then be used to provide a comparison of the Official and ECMWF perturbations, and of the ACCESS and ECMWF perturbations, respectively. We can then take means of the WPI on an hourly basis; i.e. all the 00:00 UTC WPI values are averaged, all the 01:00 UTC values are averaged, and so forth, and denote such an average by $\overline{\mathrm{WPI}}$.

$\overline{\mathrm{WPI}}$ is the difference of two mean absolute errors. A $\overline{\mathrm{WPI}}_{\mathrm{OA}}$ value of 0.5 kn at 00:00 UTC means that the Official 00:00 UTC perturbations are, on average, 0.5 kn closer to the observed perturbations than are those of ACCESS. The WPI compares just *one aspect* of the Official forecast with model guidance; it says nothing, for instance, about whether the variability of the Official forecast is closer to that of the AWS than the model guidance. As such, any statements about performance made throughout this paper refer solely to WPI, and no claim is being made that WPI is sufficient to completely characterise the accuracy, or value to the user, of how the surface diurnal wind cycle is represented in competing forecasts.

Note that sea-breeze and boundary layer mixing processes depend crucially 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 either climatology, or model predictions of background conditions, to justify making the edit. Thus, it makes sense to

compare forecast perturbations to observed perturbations, as long as errors are interpreted as the consequence not only of how the forecaster or model resolves the diurnal cycle, but of how errors in the background state contribute to errors in the perturbations. To minimise the significance of background state errors, this study focuses exclusively on lead-day one forecasts.

Given the large degree of turbulence and unpredictable variability in both the AWS, Official, and model datasets, care must be taken to ensure we do not pre-emptively conclude Official has outperformed the model guidance when $\overline{\text{WPI}} > 0$ purely by chance. The method for estimating confidence in $\overline{\text{WPI}}$ is based on a method proposed by Griffiths et al. (2017) as a general framework for BoM verification metrics. Note first that WPI is defined so as to minimise the temporal autocorrelations within each dataset, and to avoid having to consider correlations between the zonal and meridional components within and between datasets. Time series formed from the WPI values at a particular time, say 00:00 UTC, across the three month time period, can therefore be idealised as an independent random sample of a random variable W. The sampling distribution for each $\overline{\text{WPI}}$ can be modelled by a Student's t-distribution, and from this we calculate the probability that W is positive, denoted $\Pr(W > 0)$. Although temporal autocorrelations of WPI, i.e. correlations between WPI values at a particular hour from one day to the next, are in practice small or non-existent thanks to how WPI is defined, 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). Note that in the standard language of statistical hypothesis testing, we would reject the null hypothesis that $W = 0$ at significance level α if $\Pr(W > 0) > 1 - \frac{\alpha}{2}$ or $\Pr(W < 0) > 1 - \frac{\alpha}{2}$. However, in this study we are interested in both whether $W > 0$ or whether $W < 0$, so prefer to simply state the value of $\Pr(W > 0)$, referring to this as a *confidence score*, and noting $\Pr(W < 0) = 1 - \Pr(W > 0)$. We say Official outperforms model guidance with "high confidence" if $\Pr(W > 0) \geq 95\%$, or that model guidance outperforms

Official with "high confidence" if $\Pr(W > 0) \leq 5\%$, with high confidence implicit whenever it is not explicitly mentioned. Much of this explanation is probably unnecessary, but would like to get feedback before I trim it.

To investigate the consequences of the representation and mesoscale verification challenges discussed in section 1, we apply *upscaling* (Ebert 2008), where forecast and observational data are first averaged to coarser spatiotemporal scales before being compared. In this study we consider three spatial and two temporal scales. The finest spatial scale is that of the individual station. This study focuses on the 8 capital city airport stations, marked by stars in Fig. 1, as their high operational significance means that they are typically the most accurate and well maintained. The next spatial scale is formed by taking 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 *airport station groups*. The coarsest spatial scale is formed by taking all stations within 150 km of the nearest coastline, and grouping these by state. This is done because Australian forecasts are currently produced on a state by state basis at forecasting centres based in each state capital, with each forecasting centre utilising different forecasting practices. Indeed, the Official gridded forecast typically shows slight discontinuities across state boundaries (Bureau of Meteorology 2019). Note that 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, so as to best resolve the land-sea breeze signal after spatial averaging (e.g. Vincent and Lane 2016). These eight station groups are referred to as the *coastal station groups*.

We also consider both daily and seasonal time scales. For daily time scales, we either consider just the individual airport stations, or modify the definition of WPI in equation (1) so that each

perturbation dataset is first spatially averaged over either the airport or coastal station groups. Confidence scores are calculated for the airport and coastal station groups in the same way as for the single 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).

For the seasonal scale comparison we define the *Climatological Wind Perturbation Index* (CWPI) by

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

where the over-bars denote temporal averages of the perturbations at a particular hour, across the three month time period. These temporally averaged perturbations represent the climatological diurnal wind cycle over the three month study period for each dataset. $\mathrm{CWPI}_{\mathrm{OE}}$ and $\mathrm{CWPI}_{\mathrm{EA}}$ are defined analogously. The three spatial scales are considered in the same way as for WPI, with the spatial average taken before the temporal average. Uncertainty in the CWPI is estimated through bootstrapping (Efron 1979). This is done by performing resampling with replacement on the underlying perturbation datasets, and calculating the CWPI multiple times using these resampled datasets. This provides a distribution of CWPI values, which analogously to with WPI, we treat as a sample from a random variable C, and use this to estimate $\Pr(C > 0)$.

Although the WPI and CWPI provide quantitive information on the accuracy of the diurnal cycle at different times of day, they do not provide much information on the structure of the diurnal wind cycles of each dataset. Gille et al. (2005) obtained summary statistics on the observed structure of the climatological diurnal wind cycles across the globe 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 the horizontal, 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 hourly wind data, 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. 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 nonlinear regression, performed using the `least_squares` function from the `scipy.optimize` python module (SciPy 2019).

Note Gille et al. (2005) fit equations (3) and (4) to the temporally averaged wind fields, so that (u_0, v_0) could be interpreted as the mean wind over the study's time period, and the remaining terms providing the climatological diurnal perturbations. In this study we fit equations (5) and (6) to the climatological perturbations themselves, with (u_0, v_0) now necessary to offset the asymmetry introduced by α, i.e. to ensure the time integral of the fitted perturbation values is approximately zero. Following Gille et al. (2005), the ellipse's orientation, i.e. the angle the semi-major axis of the ellipse makes with lines of latitude, as well as the ellipse's eccentricity are calculated algebraically, but the perturbation speed maximum, and the time at which this maximum is achieved, are instead obtained numerically.

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 (austral winter) 2018. First, absolute errors are compared on a daily basis using the Wind Perturbation Index (WPI) at three different spatial scales. Second, overall seasonal biases during this time period are assessed using the Climatological Wind Perturbation Index (CWPI), and by comparing structural indices derived from ellipses fitted to the climatological wind perturbations.

a. Daily Comparison

Figure 2 provides the WPI values and confidence scores for the coastal station groups for $\overline{\text{WPI}}_{\text{OA}}$, $\overline{\text{WPI}}_{\text{OE}}$ and $\overline{\text{WPI}}_{\text{EA}}$, which represent the the Official versus ACCESS, Official versus ECMWF, and ECMWF versus ACCESS comparisons, respectively. The results indicate that for the majority of station groups and hours, both the unedited ACCESS and ECMWF models outperform the Official forecast. The lowest $\overline{\text{WPI}}$ values occur at the NT station group at 23:00 and 00:00 UTC

for both $\overline{\text{WPI}}_{\text{OA}}$ and $\overline{\text{WPI}}_{\text{OE}}$. Although Official outperforms at least one of ACCESS or ECMWF 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, although the $\overline{\text{WPI}}$ values are comparatively low. ECMWF 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 WPI for the NT station group at 23:00 UTC. The time series shows significant temporal variability, with WPI frequently dropping below -2 kn. Figures 3 b) and c) show hodographs of the winds and wind perturbations, respectively, for the AWS observations, Official forecast, and ACCESS and ECMWF model datasets, at each hour UTC on the 3$^{\text{rd}}$ 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 ECMWF from 07:00 to 13:00 UTC, then unedited ACCESS from 15:00 to 21:00 UTC. The final two hours of the forecast show the Official winds acquiring a stronger east-northeasterly component than the other datasets. This rapid change is clearer in the perturbation hodograph shown in Fig. 3 c). At this time of year the prevailing winds throughout the NT are east-southeasterly, and 22:00 UTC corresponds to $\approx$ 08:30 local solar time (LST) in this region, so the rapid departure of the Official forecast from ACCESS at this time likely represents an edit made by a forecaster to capture boundary layer mixing processes.

Figure 5 a) shows the first ten values from wind soundings at Darwin Airport at 12:00 UTC on July 3$^{\text{rd}}$ and 00:00 UTC on July 4$^{\text{th}}$. In both instances the winds are indeed 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 south-

easterly component of the winds too much. Similar issues create low WPI scores on the 8th of June and 9th and 10th of July.

Figure 4 a) provides a time series of WPI for the South WA station group at 05:00 UTC. As with the NT station group there is significant temporal variability, with WPI frequently exceeding 1 kn. Figures 4 b) and c) provide hodographs of the winds and wind perturbations, respectively, on the 9th of June, which is an interesting example. The perturbation hodograph shows both ECMWF and ACCESS under-predicting the amplitude of the diurnal wind cycle on this day. In each dataset the 05:00 UTC perturbations are westerly to northwesterly, and given the orientation of the South WA coastline (see Fig. 1) and the fact that 05:00 UTC corresponds to around 13:00 local solar time (LST) in this region, the perturbations likely indicate boundary layer mixing processes.

Figure 5 shows wind soundings at Perth Airport, the nearest station to the South WA station group 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 perturbations from 04:00 to 07:00 UTC show stronger westerly perturbations than either ACCESS or ECMWF, improving the amplitude of Official's diurnal wind cycle. However, the AWS perturbations are more northerly than those of Official, 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 8th 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 WPI scores on

the 3rd of August, although in this case the Official forecast slightly improves both the magnitude and direction of the 05:00 UTC wind perturbations.

To contrast with the coastal station group results, Fig. 6 presents the $\overline{\text{WPI}}$ values and confidence scores for $\overline{\text{WPI}}_{\text{OE}}$, which represents the Official versus ECMWF comparison, for the airport stations, and airport station groups. 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. Official outperforms ECMWF at 01:00 and 02:00 UTC at both the Darwin airport station and the NT station group, although ECMWF outperforms Official between 08:00 and 14:00 UTC at Darwin and Brisbane airports, and the corresponding NT and QLD station groups, with the exception of the QLD station group at 12:00 UTC. ECMWF also outperforms Official at Hobart airport at almost all hours of the day, and at Adelaide and Canberra airports from 11:00 to 14:00 UTC.

For the remaining stations and times, Official only outperforms ECMWF at the Perth airport station at 06:00 UTC and the Melbourne airport station at 01:00 UTC, although in both cases WPI values are comparatively small in magnitude. Furthermore, in both cases there is no clear pattern to the $\overline{\text{WPI}}_{\text{OE}}$ values over the rest of the day. Note that the *multiplicity problem* (Wilks 2011, p. 178) requires care be taken before giving meaning to these two examples: i.e., given that we are calculating twenty four confidence scores for eight stations, then if WPI were uncorrelated across each station and hour we would expect to find $0.05 \times 24 \times 8 \approx 10$ instances where $P(W_{OE} > 0) \geq 95\%$, even if W_{OE} were in fact equal to zero.

For the airport station groups, ECMWF outperforms Official for the majority of station groups and times. The main exception is the Darwin airport station group, where Official outperforms ECMWF at 02:00 UTC, and there is ambiguity as to whether Official or ECMWF performs better at 01:00, 03:00 and 04:00 UTC, and from 15:00 to 22:00 UTC. In the analogous $\overline{\text{WPI}}_{\text{OA}}$ Official versus ACCESS comparisons (not shown), the airport station results are similarly noisy, although

the airport station group results are slightly more favourable to Official, with Official outperforming ACCESS from 10:00 to 12:00 UTC at the Brisbane station group, and fewer occasions overall where ACCESS outperforms Official than ECMWF does.

Figure 7 presents the $\overline{\text{WPI}}$ values and confidence scores for $\overline{\text{WPI}}_{\text{EA}}$, which represents the ECMWF versus ACCESS comparison, for the airport stations, and airport station groups. As with the Official versus ECMWF comparison in Fig. 6, the results for the airport stations are noisy, but more often than not show that ECMWF outperforms ACCESS. The results for the airport station group show ECMWF usually outperforms ACCESS, the main exceptions being the Darwin and Canberra airport station groups. Might be interesting to note that ACCESS-C+ does not run over Darwin or Canberra, possibly explaining the better performance of ACCESS there.

Naively, the fact that ECMWF generally outperforms ACCESS at these scales is surprising, as ACCESS runs at a higher spatiotemporal resolution than ECMWF, and is calibrated for Australian conditions. However, these results are not surprising in light of the mesoscale verification challenges discussed in section 1. The AWS data resolves motion with time scales as low as 10 minutes, and at arbitrarily small spatial scales: it therefore includes more unpredictable turbulence than either model dataset. Furthermore, because ACCESS runs at higher spatiotemporal resolutions than ECMWF, it includes additional scales of motion, and therefore adds additional variability to the wind fields. Unless the additional variability in ACCESS is perfectly correlated with observations, the average of $|\boldsymbol{u}_{\text{AWS}} - \boldsymbol{u}_{\text{A}}|$ will therefore increase, unless this additional variability is compensated for by a reduction in bias, i.e.$|\overline{\boldsymbol{u}}_{\text{AWS}} - \overline{\boldsymbol{u}}_{\text{A}}|$ decreases. These ideas are discussed in greater detail in section 4. Note finally that the results for the Official versus ECMWF comparison in Fig. 6 largely mirror those of the ECMWF versus ACCESS comparison in Fig. 7, suggesting that similar arguments apply to Official, as it as based on both ACCESS and ECMWF, as well as forecaster edits, which contribute additional variability.

b. Seasonal Comparison

Figure 8 provides the Climatological Wind Perturbation Index (CWPI) values and confidence scores for the coastal station groups for CWPI_{OA}, CWPI_{OE} and CWPI_{EA}, which represent the the Official versus ACCESS, Official versus ECMWF, and ECMWF versus ACCESS comparisons, respectively. At the NT station group Official outperforms both ACCESS and ECMWF at 03:00 UTC with confidence $\geq 93\%$. However, both ACCESS and ECMWF outperform Official at 23:00 and 00:00 UTC, consistent with the $\overline{\text{WPI}}$ results of Fig. 2. The NT station group results are discussed in more detail in section 4.

At the North WA station group at 01:00, 03:00 and 04:00, Official outperforms ACCESS with confidence scores of 77, 78 and 90%, respectively; Official also outperforms ECMWF at 01:00 and 02:00 UTC with confidence scores above 99%. Figure 9 a) shows that ECMWF's poor performance at 01:00 and 02:00 UTC is simply due to its linear interpolation at these times, whereas Official's very slight outperformance of ACCESS at 01:00, 03:00 and 04:00 is due to ACCESS's climatological diurnal cycle being slightly out of phase with that of the AWS observations, and the Official forecast correcting for this somewhat. Both Official and ECMWF slightly exaggerate the magnitude of the climatological sea-breeze, which peaks around 09:00 UTC, with ACCESS performing well in this respect.

At the South WA station group from 01:00 to 05:00 UTC, Official outperforms ECMWF with confidence scores of at least 88%. Figure 9 b) shows that ECMWF 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 Official, ACCESS and ECMWF noticeably 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, Official outperforms both ACCESS and ECMWF with confidence scores of at least least 95% and 75%, respectively. Figure 9 c) shows that these times correspond to "dimples" in the perturbation hodographs that are present in all four datasets. The Official hodograph closely resembles that of ACCESS, except for this dimple, which has been exaggerated relative to ACCESS. Don't know what is going on here. Figure 9 c) also shows that although ECMWF exaggerates the amplitude of the easterly sea-breeze perturbations, it captures the narrower shape of the AWS hodograph better than Official or ACCESS.

At the SA station group from 02:00 to 05:00 UTC and 09:00 to 12:00 UTC, Official outperforms both ACCESS and ECMWF, 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 Official only slightly outperforms ACCESS in the results of Figures 8 a) and b).

For constrast, Fig. 10 presents the CWPI values and confidence scores for $\mathrm{CWPI_{OE}}$, which represents the Official versus ECMWF comparison, for the airport stations, and airport station groups. These results show much greater similarity with the Official versus ECMWF comparisons at the coastal station groups shown in Figs. 8 c) and d), than do the analogous $\overline{\mathrm{WPI}}$ results in Fig. 6 and Figs. 2 c) and d). This likely because the temporal averaging has reduced the additional unpredictable variability in Official, revealing biases in Official and ECMWF that are partly shared across the three spatial scales. This point is discussed further in section 4. The analogous CWPI comparisons with ACCESS (not shown) are more ambiguous, although are generally more favourable for Official than those for $\overline{\mathrm{WPI}}$. For example Official outperforms both ACCESS and ECMWF at Darwin Airport from 02:00 to 03:00 and 15:00 to 17:00 UTC with at least 90% confidence. However, Official performs less well compared to ACCESS over the airport station groups,

with CWPI values close to zero for most times and station groups, but ACCESS now strongly outperforming Official over the Darwin Airport station group.

Note that 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 perturbations, as described in section 2. Figure 11 provides 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%. It also performs well at the airport station and airport station group scales, with a few exceptions, including the ACCESS and Official meridional perturbations at the Canberra airport station group, and the ECMWF zonal perturbations at Melbourne airport.

The ellipse fits are used to derive four descriptive quantities: the maximum perturbation speed, the eccentricity of the fitted ellipse, the angle the fitted ellipse's semi-major axis makes with lines of latitude, and the time maximum perturbation speed occurs. Figure 12 provides these four quantities for each dataset and location across the three spatial scales. A variety of structural differences are apparent at a number of locations and scales. For example, Fig. 12 a) shows that at Brisbane airport, the maximum AWS perturbation is at least 1 kn greater than Official, ACCESS and ECMWF, and Fig. 12 c) shows that the orientation of the AWS fitted ellipse is at least 20 degrees anticlockwise from the other datasets. Figures 13 a) and b) show hodographs of the Brisbane airport perturbation climatology and ellipse fit, 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, ECMWF and ACCESS sea-breezes approach more from the east-northeast. To check whether this just represents a direction bias of the Brisbane Airport station, Fig. 12 shows the climatological perturbations at the nearby Spitfire Channel station (see Fig. 1

for the location of this station, and other stations referred to in this section). 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. Thus, the directional bias in Official, ACCESS and ECMWF at these stations is likely genuine, and not just a consequence of biased AWS observations. Figure 1 shows there are two small islands to the east of Brisbane airport; the more northwesterly 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 Official, ACCESS or ECMWF.

Another example is the Hobart Airport station. Figure 12 c) shows that the ellipse fits for the AWS perturbations are oriented 31, 35 and 62 degrees anti-clockwise from the ECMWF, Official 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. However, figures 13 d) and e) show that the fit still captures orientations accurately, although it underestimates the maximum AWS perturbation. Figure 13 f) shows 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.

The South WA station group also provides an interesting example. Here the ACCESS and Official ellipse fits are oriented at least 49 degrees anti-clockwise from those of AWS and ECMWF, and the ECMWF 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 ECMWF

that the other datasets. A similar issue affects the VIC station group, explaining why 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 provide further examples. In these cases there are timing differences between the perturbation maximums of up to 8.2 hours. Figure 14 shows that these differences occur because for some datasets, the later north to northwesterly sea-breeze perturbations dominate the diurnal wind cycle, but for other datasets the earlier easterly to southeasterly boundary layer mixing effects dominate.

4. Discussion

The results of section 3 may have implications for forecasting practice. If the goal of land-sea breeze and boundary layer mixing edits is to reduce absolute errors in the following day's forecast of the surface wind fields, then a necessary (but not sufficient) condition for this to occur is for these edits to at least reduce the absolute errors in the diurnal component of the surface wind fields. However, the WPI comparisons in Figs. 2 and 6 suggest that this is only possible when absolute error is calculated at coarse spatial scales. If the Official forecast is based, at least partly, on an edited high resolution model guidance dataset like ACCESS, then due to the mesoscale verification issues discussed in section 1, the larger absolute errors associated with a higher resolution model completely mask the effect of the edits, with a lower resolution unedited model like ECMWF scoring better overall. While the CWPI results in Figs. 8 and 6 suggest that forecaster edits can improve the accuracy of diurnal wind cycles in a climatological sense, it is not clear if these improvements have operational significance.

To investigate these ideas further, consider first just the zonal components of the AWS and Official wind perturbations, denoted by u_{AWS} and u_{O} respectively. Considering just the values at a particular hour UTC, at a particular station, 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

$$\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{O}}) = \underbrace{\mathrm{var}\,(u_{\mathrm{AWS}}) + \mathrm{var}\,(u_{\mathrm{O}}) - 2 \cdot \mathrm{covar}\,(u_{\mathrm{AWS}}, u_{\mathrm{O}})}_{\mathrm{var}(u_{\mathrm{AWS}} - u_{\mathrm{O}})} + \underbrace{(\overline{u}_{\mathrm{AWS}} - \overline{u}_{\mathrm{O}})^2}_{\text{squared bias}} \tag{8}$$

where var, covar and over-bars denote the sample variance, covariance and mean respectively. The first three terms are the total variance of $u_{\mathrm{AWS}} - u_{\mathrm{O}}$, whereas the last term is the square of the bias between u_{AWS} and u_{O}. Note that the mean square error $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{O}})$ is closely related to $\overline{\mathrm{WPI}}$, which is the difference between the mean absolute error of Official and AWS, and a model guidance dataset and AWS. Similarly, the CWPI is closely related to the squared bias component $(\overline{u}_{\mathrm{AWS}} - \overline{u}_{\mathrm{O}})^2$ of the mean square error. Equation (8) can also be applied to wind perturbations that have first been spatially averaged over a station group, and to $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{E}})$ and $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{A}})$, where u_{E} and u_{A} are the ECMWF and ACCESS zonal perturbations, respectively.

Figure 15 shows each term in the mean square error decomposition of equation 8 for both $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{O}})$ and $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{E}})$, for Darwin Airport, the Darwin station group, and the NT station group. This region provides an interesting case study because Fig. 6 shows that Official has some skill at both Darwin Airport and over Darwin Airport station groups, in contrast to most other locations. At Darwin Airport, $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{O}})$ exceeds $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{E}})$ from 04:00 to 16:00 UTC due to higher total variance, whereas outside of these times $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{E}})$ exceeds $\mathrm{mse}\,(u_{\mathrm{AWS}}, u_{\mathrm{O}})$ due to larger bias. The higher total variance of $u_{\mathrm{AWS}} - u_{\mathrm{O}}$ occurs because $\mathrm{var}\,(u_{\mathrm{O}}) > \mathrm{var}\,(u_{\mathrm{E}})$. This additional variability is mostly random from 04:00 to 14:00 UTC, i.e. u_{O} is not sufficiently correlated with u_{AWS} at these times for the additional variability of u_{O} to produce a reduction in mean square error. Thus, while the bias between Official and AWS is lower, or about the same, as that between ECMWF and AWS, the higher random variability of Official results in higher mean square error for most of the day. Figure 16 shows similar conclusions can be drawn for the meridional perturbations at Darwin Airport, although in this case $\mathrm{var}\,(u_{\mathrm{O}}) > \mathrm{var}\,(u_{\mathrm{E}})$

for the entire day. Most of the difference between the WPI and CWPI scores for the Official versus ECMWF comparison at Darwin Airport in Figures 6 and 10, respectively, can be explained through the different mean square error and bias terms for the zonal perturbations alone.

Figure 14 a) shows that ECMWF's climatological perturbations at Darwin Airport underestimate the easterly perturbations from 00:00 to 03:00 UTC, which are presumably associated with boundary layer mixing processes. Official does a better job of resolving these easterly perturbations, but is generally outperformed by ECMWF in resolving the northerly sea-breeze perturbations. Similar points can be made for the Darwin and NT coastal station groups. While spatial averaging reduces a portion of the unpredictable variability in Official, Official also often has larger meridional biases at these scales compared to ECMWF. Figures 14 and 12 show that these biases can be explained in terms of amplitude and orientation differences between Official, ECMWF and AWS. Figures analogous to Figs. 15 and 16, but for other locations around Australia, show similar results, but generally without large biases in the Official forecast at the coarser scales like those present in the meridional perturbations over the Darwin Airport station group and NT coastal station group.

These examples illustrate the idea that the additional unpredictable variability introduced by a higher resolution edited forecast needs to be "paid for" by a reduction in bias, otherwise the net result will just be an increase in error. One obvious way to reduce the influence of unpredictable variability at higher resolutions is to move to an ensemble forecasting approach. However, computational resources typically require a choice between either a single model of higher resolution, or an ensemble of models at lower resolution, and there is a long and spirited debate in the literature about the relative merits of each (Brooks and Doswell III 1993). If ensemble forecasting is not possible, careful thought must be given to precisely what scales of motion the Official forecast is intended to represent. If the end user doesn't care about the "realism" of the forecast and simply wants the lowest errors, and if daily errors are larger with a higher resolution, edited forecast,

than with a coarse model guidance product, there may be an argument for smoothing or filtering the higher resolution forecast before it is provided to the end user, assuming of course the higher resolution forecast actually reduces biases.

Furthermore, it is unclear if the Official forecast's wind fields are intended to be regarded as predictions of the actual winds at a specific location, or a predicted Reynold's average, and if so, at what scale. If we assume the BoM's Official wind forecast intends to represent variability at hourly timescales, and horizontal scales less than 50 kms, then sea-breeze and boundary layer mixing edits appear to have little effect, because the intended reduction in error is washed out by the unpredictable turbulent variability at these scales, and lower errors can be achieved simply by using a coarser resolution unedited model forecast. However, some users may be more interested in whether the variability of the forecast wind field matches those of observations, than in whether the forecast minimises absolute error, and a higher resolution edited forecast will likely perform better than a coarse resolution model in this regard.

Ambiguity in what the Official forecast represents, and the resulting verification challenge, points to an important role for human forecasters: post-processing and editing model data so that the forecast consistently represents what is of interest to individual users. One barrier to this is that the Official forecast is provided to the national public as a whole, which includes diverse users with different representation needs. Part of the solution may therefore be the development of a secondary economy, either within national weather services like the BoM, or in the private sector, where human forecasters ensure their forecast products consistently represent the "filtered version of reality" of interest to the end user. This would then help address the representation problem as it applies to the verification of operational forecasts, as the individual user's representation needs could then be taken as the intended representation of the forecast, and appropriate verification metrics chosen accordingly.

5. Conclusion

In this paper we have presented a method for verifying the diurnal component of wind forecasts issued to the public, 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 two temporal scales, and three spatial scales, but the method is immediately generalisable to other scales.

When the method is applied to Australian forecast data, the results indicate that when the Official, edited forecast, is assessed on a daily basis, it only produces lower absolute errors in the diurnal wind cycle at very coarse spatial scales of at least 500 km. Even at these scales, the improvements are isolated to particular times of day, and only apply at some locations. Furthermore, while the Official forecast can outperform the two most commonly used model guidance products ACCESS and ECMWF in the sense of absolute error, it rarely outperforms both simultaneously, suggesting that forecaster skill lies more in making the choice of model guidance than in making edits. When the Official forecast is assessed on a seasonal basis, i.e. the average or climatological diurnal wind cycle is assessed, the Official forecast performs better than when assessed on a daily basis, particularly at the single station spatial scale. However, its performance is not overwhelming, as it struggles to unambiguously produce lower absolute error than ACCESS.

An alternative to calculating absolute errors is to assess the realism of structural features of the atmosphere, and following Gille et al. (2005), we do this in an objective way by fitting ellipses to hodographs of the climatological diurnal wind cycles, and deriving structural metrics from the ellipses. In the Australian context 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.

Future research could extend this study in multiple directions. An important goal would be to identify precisely the spatial scale at which the Official forecast can produce lower absolute error on a daily basis than a coarse model like ECMWF: for Australia over the time period considered, our study shows that this occurs somewhere between our airport station group scale (50 - 200 km), and coastal station group scale (1000 - 2000 km). Another interesting question is whether the diurnal component of the Official forecast can outperform a climatological diurnal cycle calculated from previous observations.

In summary, we have shown that forecaster edits can reduce errors in the diurnal cycle of surface winds, but only at very large spatial scales, or in a climatological sense. This scale sensitivity suggests careful thought needs to be given to how the representation problem applies to the verification of operational forecasts. A consistent answer may prove challenging for national forecasting centres to reach, due to the diverse representation needs of forecast users, and in the Australian context, due to the hybrid nature of the Official 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 Tasked Automation team. (Link to Jive homepage or GitHub page? Thanks are due to Michael Foley for providing support at the Bureau of Meteorology's Melbourne office, and to Craig Bishop for some helpful conversations. The code written for this study is freely available online (Short 2019).

References

Abkar, M., A. Sharifi, and F. Porté-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.

Brooks, H. E., and C. A. Doswell III, 1993: New technology and numerical weather prediction — a wasted opportunity? *Weather*, **48 (6)**, 173–177, doi:10.1002/j.1477-8696.1993.tb05877.x.

Bureau of Meteorology, 2010: Operational implementation of the ACCESS numerical weather prediction systems. Tech. rep., Bureau of Meteorology, Melbourne, Victoria. [Available online at http://www.bom.gov.au/australia/charts/bulletins/apob83.pdf].

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*. No. 4, IFS Documentation, European Center for Medium Range Weather Forecasting, [Available online at https://www.ecmwf.int/node/18714].

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.

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

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

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

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

Lynch, K. J., D. J. Brayshaw, and A. Charlton-Perez, 2014: Verification of European subseasonal wind speed forecasts. *Monthly Weather Review*, **142 (8)**, 2978–2990, doi:10.1175/MWR-D-13-00341.1.

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

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

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

SciPy, 2019: Optimization and root finding (scipy.optimize). SciPy, [Available online at https://docs.scipy.org/doc/scipy/reference/optimize.html].

Short, E., 2019: eshort0401/forecast_verification_paper. GitHub, [Available online at https://github.com/eshort0401/forecast_verification_paper].

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.

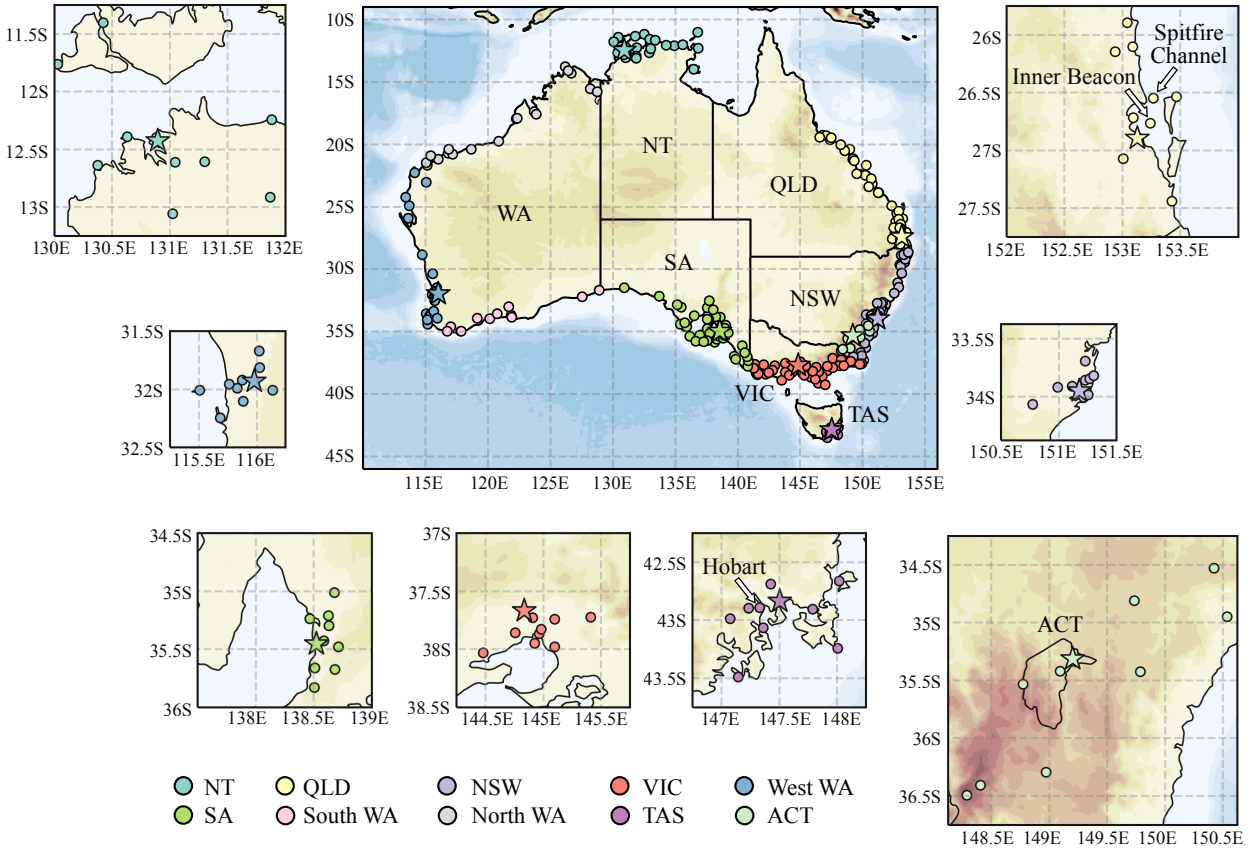
Zwiers, F. W., and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *Journal of Climate*, **8 (2)**, 336–351, doi:10.1175/1520-0442(1995)008⟨0336:TSCIAI⟩2.0.CO;2.

LIST OF TABLES

Airport	Austral Summer	Austral Winter
Darwin	6.3 kn	6.2 kn
Brisbane	8.6 kn	7.0 kn
Perth	11.3 kn	7.9 kn
Sydney	12.2 kn	10.2 kn
Adelaide	9.5 kn	10.3 kn
Canberra	7.4 kn	7.9 kn
Melbourne	10.0 kn	12.1 kn
Hobart	10.0 kn	8.7 kn

723 TABLE 1. Average 10 m wind speeds for austral winter (June, July August) 2018, and austral summer (De-
724 cember, January, February) 2017/18 across the eight Australian capital city airport weather stations.

LIST OF FIGURES



768 FIG. 1. Locations of the automatic weather stations used in this study. Stars indicate capital city airport
769 stations. Height and depth shading intervals every 200 and 1000 m, respectively.

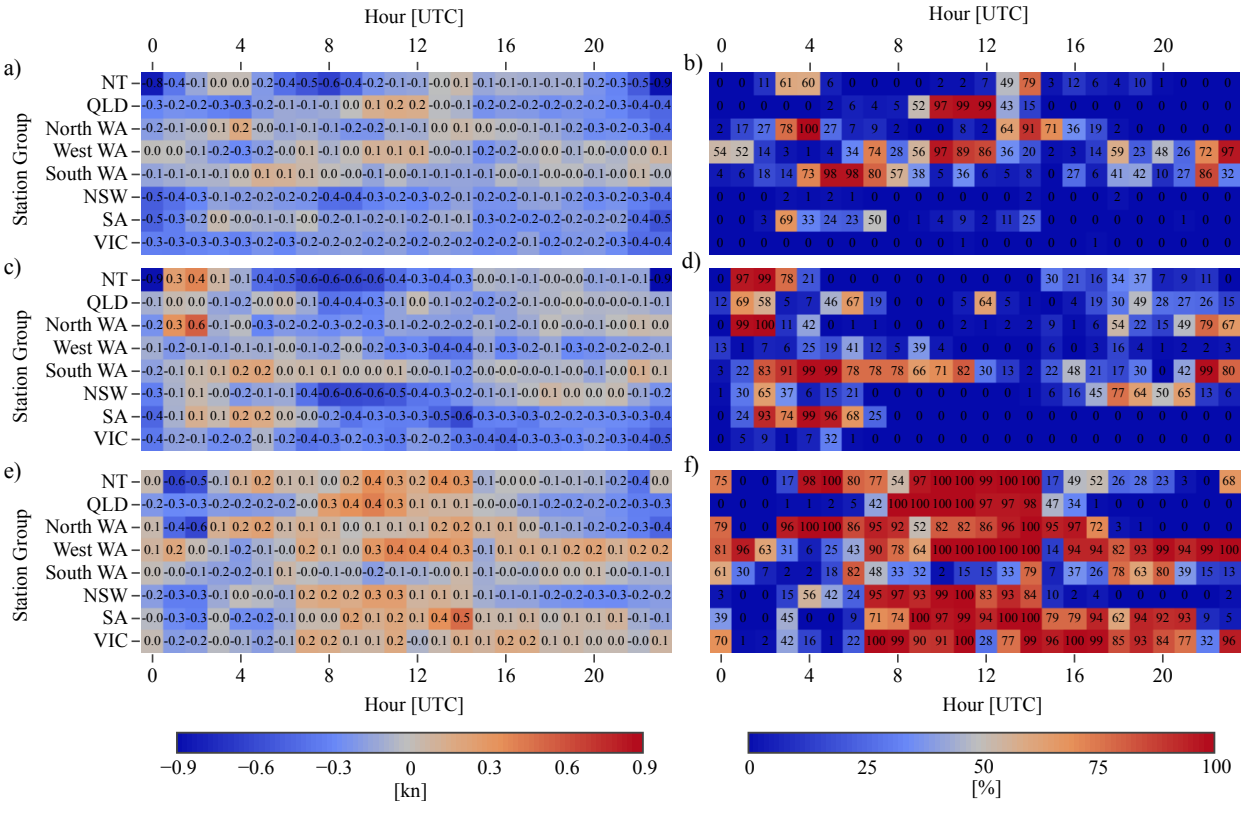
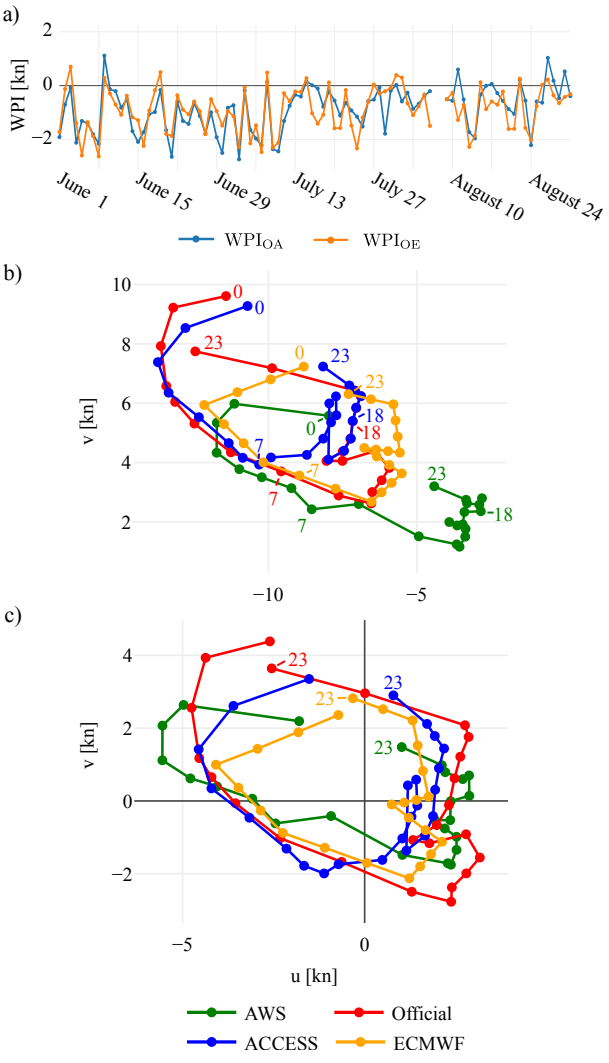
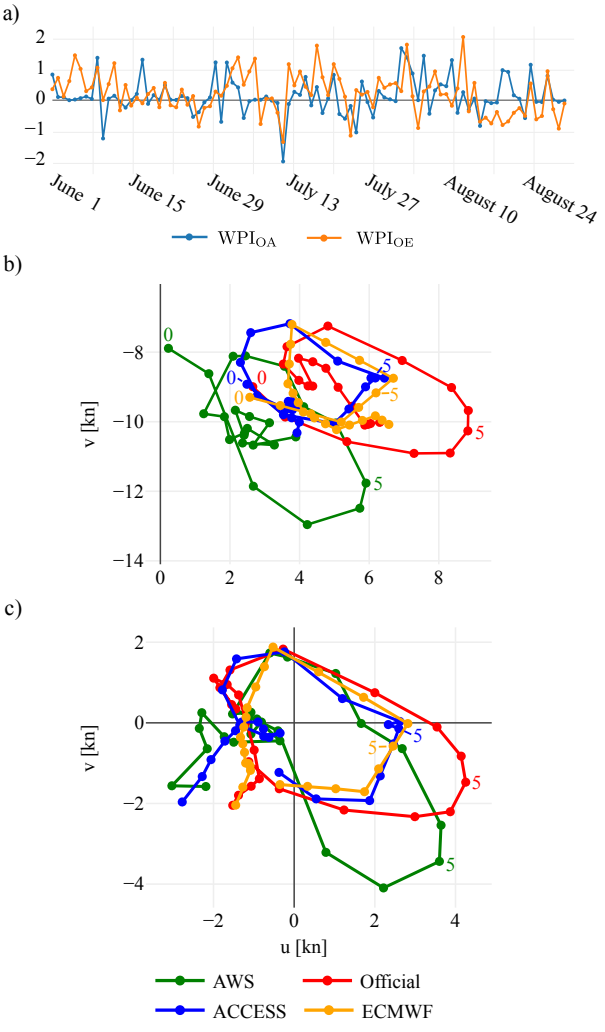


FIG. 2. Heatmaps of $\overline{\text{WPI}}$ values, a), c), e), and confidence scores, b), d), f), for each coastal station group and hour of the day: Official versus ACCESS, a) and b), Official versus ECMWF, c) and d), ECMWF versus ACCESS, e) and f). Positive $\overline{\text{WPI}}$ values indicate that the former dataset in each pair is on average $\overline{\text{WPI}}$ kn closer to observations than the latter dataset is. Confidence scores provide the probability the population or "true" value of $\overline{\text{WPI}}$ is greater than zero.



775 FIG. 3. Time series, a), of $\overline{\text{WPI}}_{\text{OA}}$ and $\overline{\text{WPI}}_{\text{OE}}$ for the NT station group at 23:00 UTC, and b), hodographs
776 showing hourly changes in winds, and c), wind perturbations, at the NT station group on the 3rd of July 2018.



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

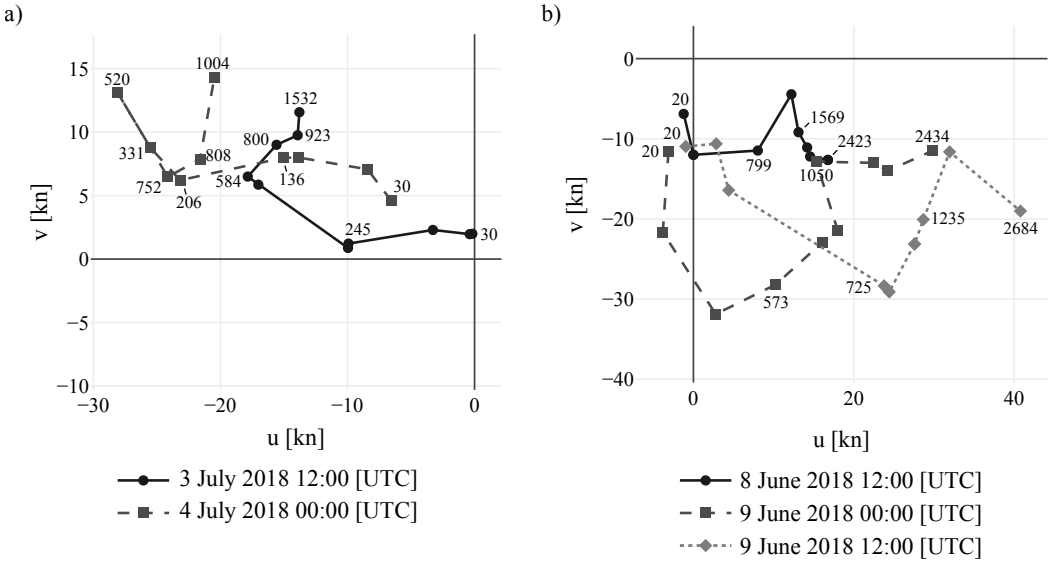
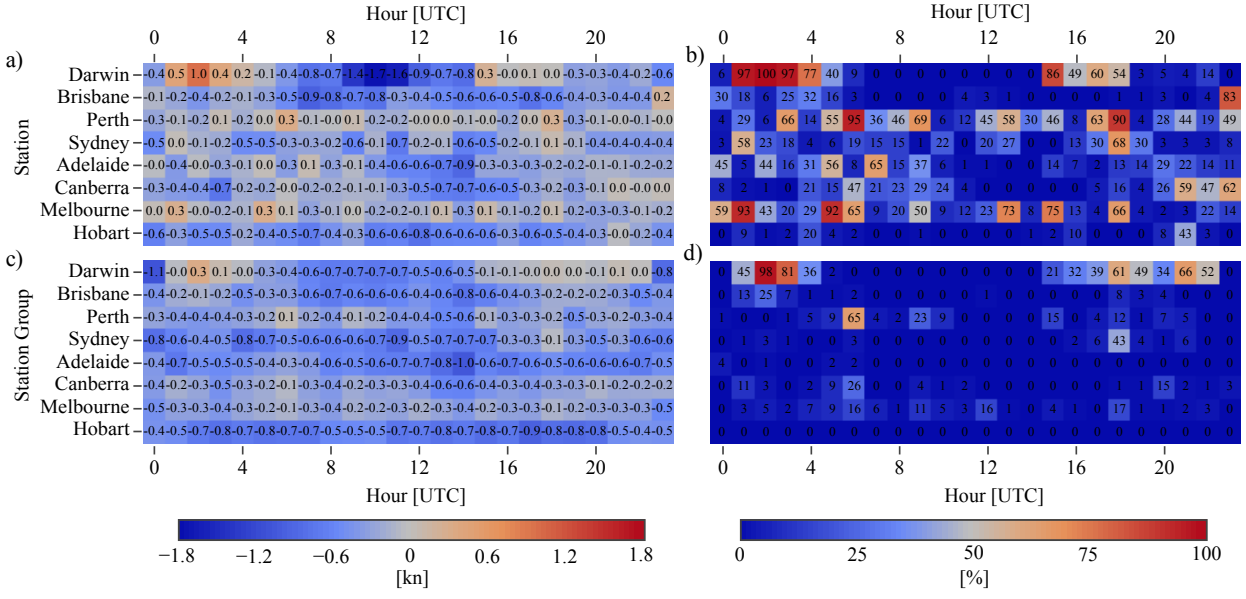


FIG. 5. Wind soundings at, a), Darwin Airport, and b), Perth Airport.



779 FIG. 6. The $\overline{\text{WPI}}_{\text{OE}}$ values, a) and c), and confidence scores, b) and d), for the airport stations, a) and b), and
780 airport station groups, c) and d).

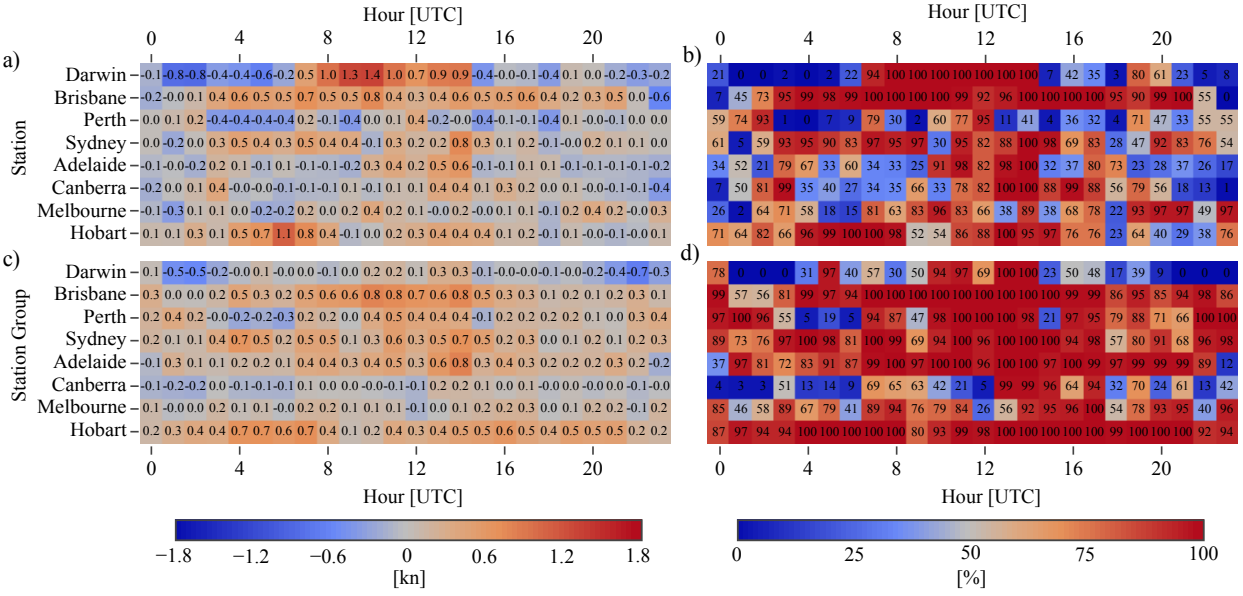


FIG. 7. As in Fig. 6, but for the $\overline{\mathrm{WPI}}_{\mathrm{EA}}$ values and confidence scores.

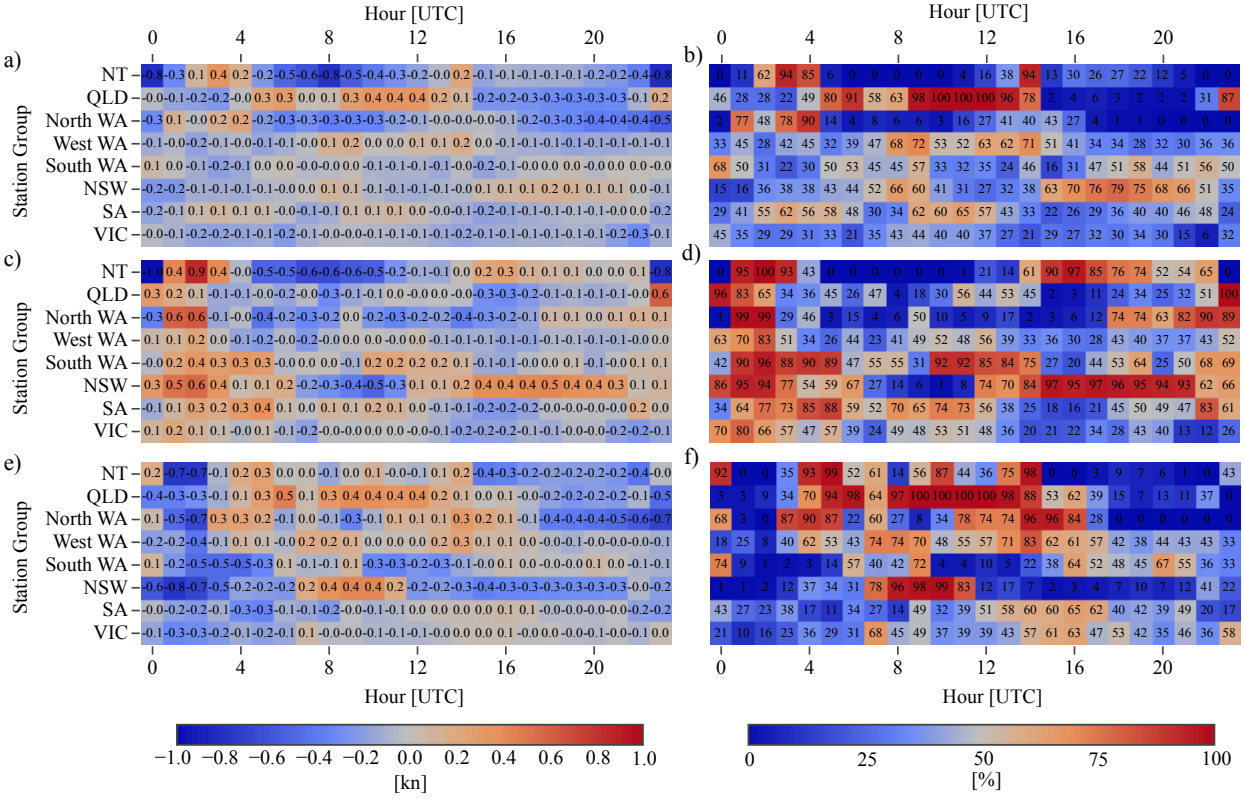
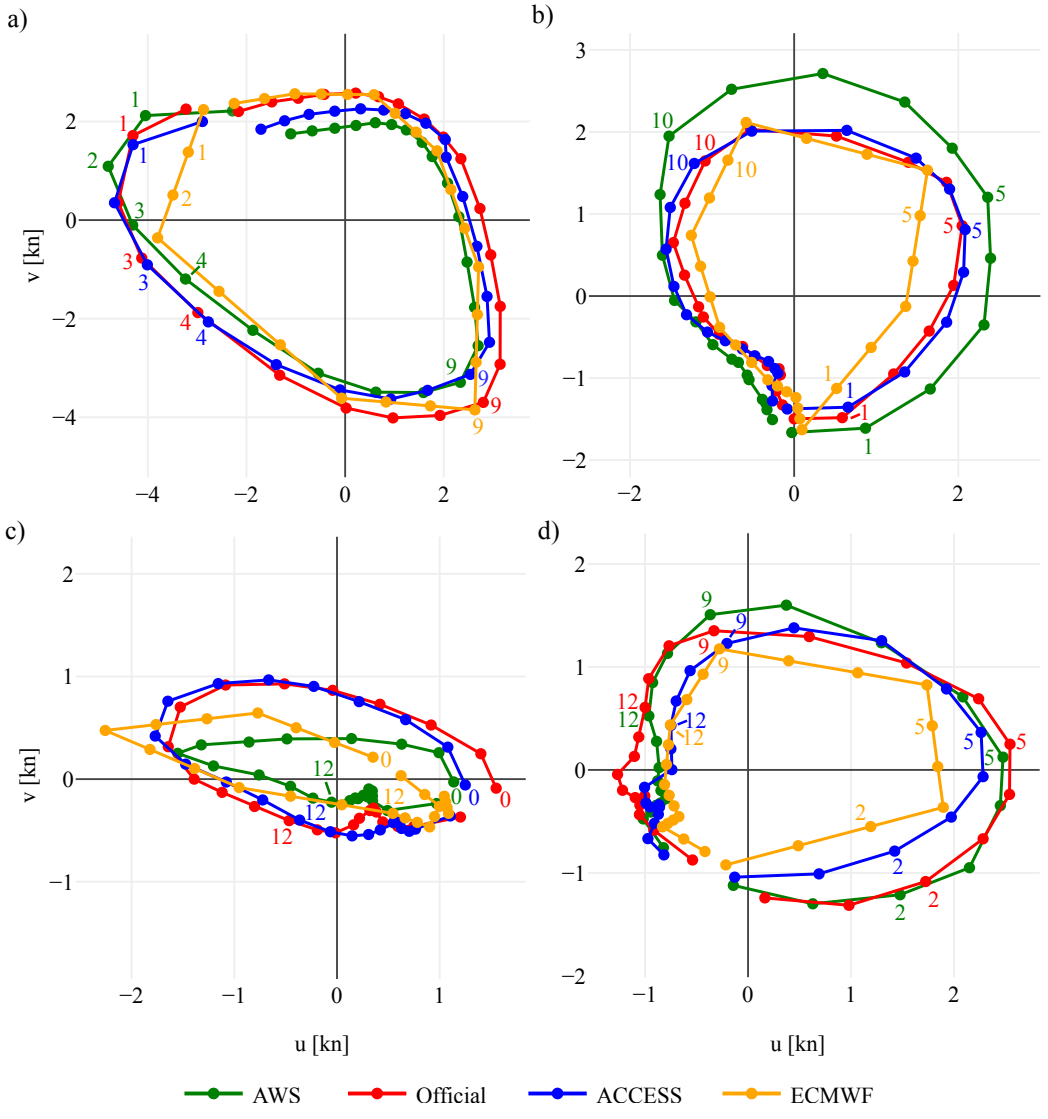


FIG. 8. As in Fig. 2, but for the CWPI values and confidence scores.



781 FIG. 9. Average wind perturbations over June, July and August 2018 for the, a), North WA, b) South WA, c)
782 NSW and d), SA coastal station groups.

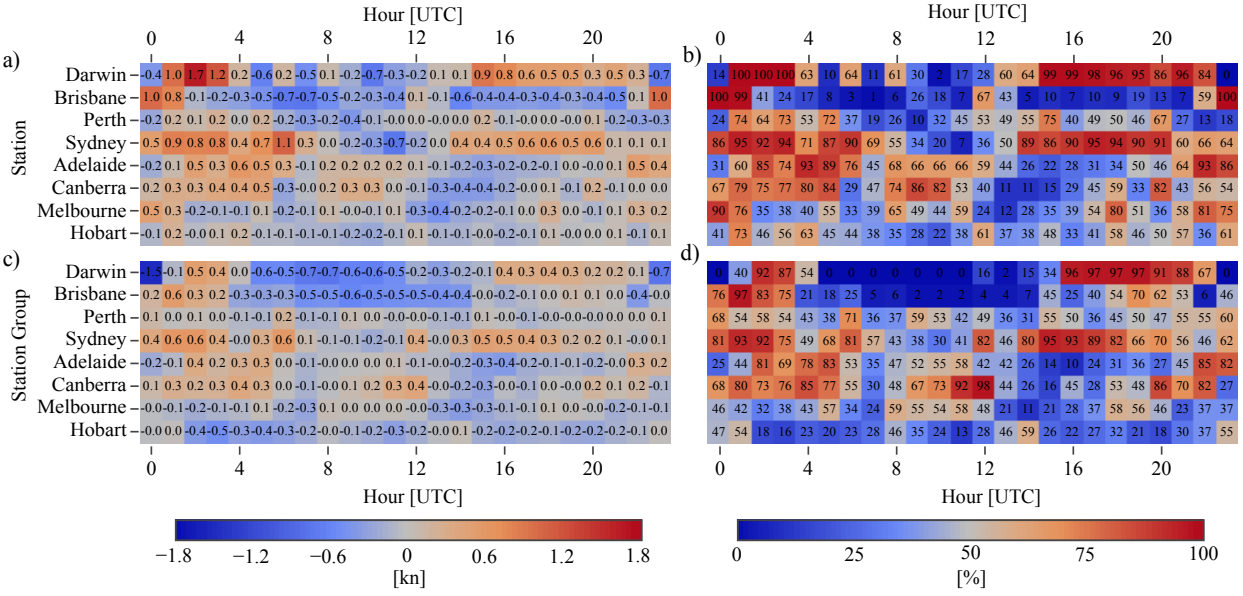


FIG. 10. As in Fig. 6, but for the CWPI values and confidence scores.

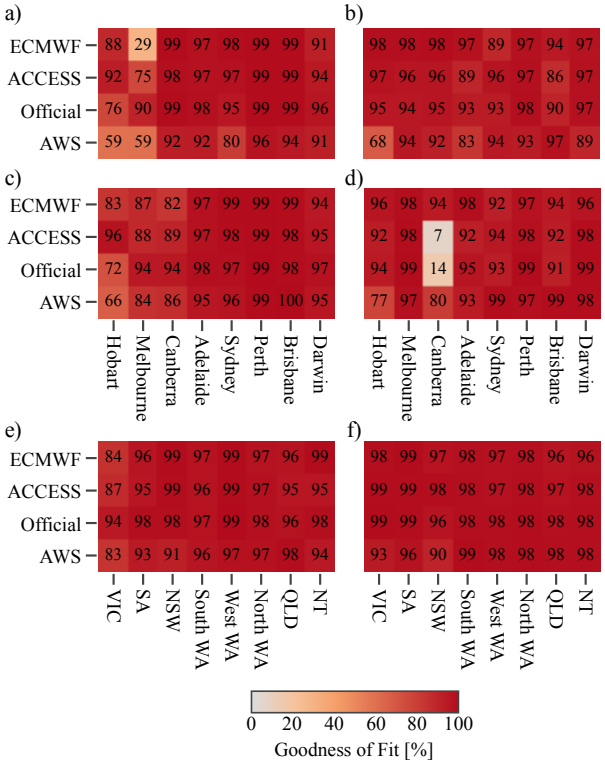


FIG. 11. R^2 values as percentages for the fit of equation (5) to the zonal perturbations, a), c) and e), and equation (6) to the meridional perturbations, b), d) and f), for the airport stations, a) and b), airport station groups, c) and d), and coastal station groups, e) and f).

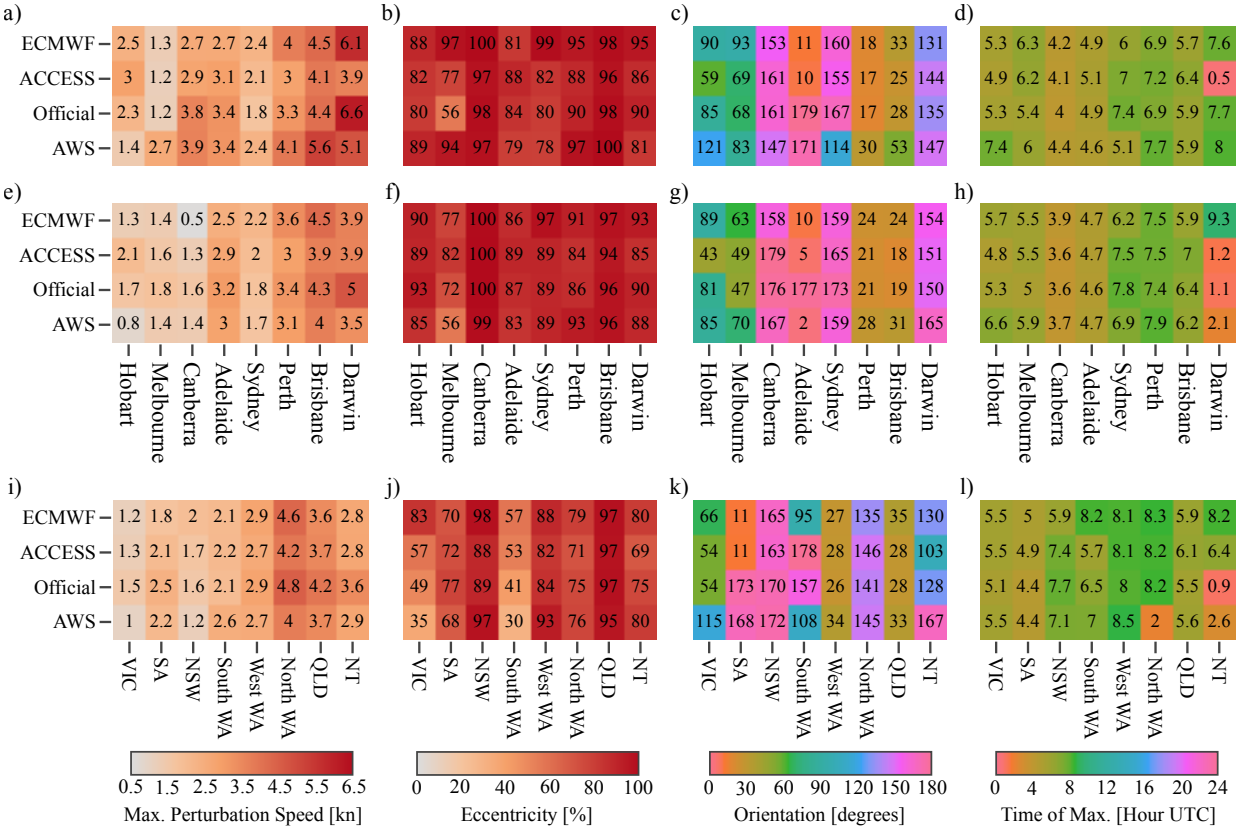
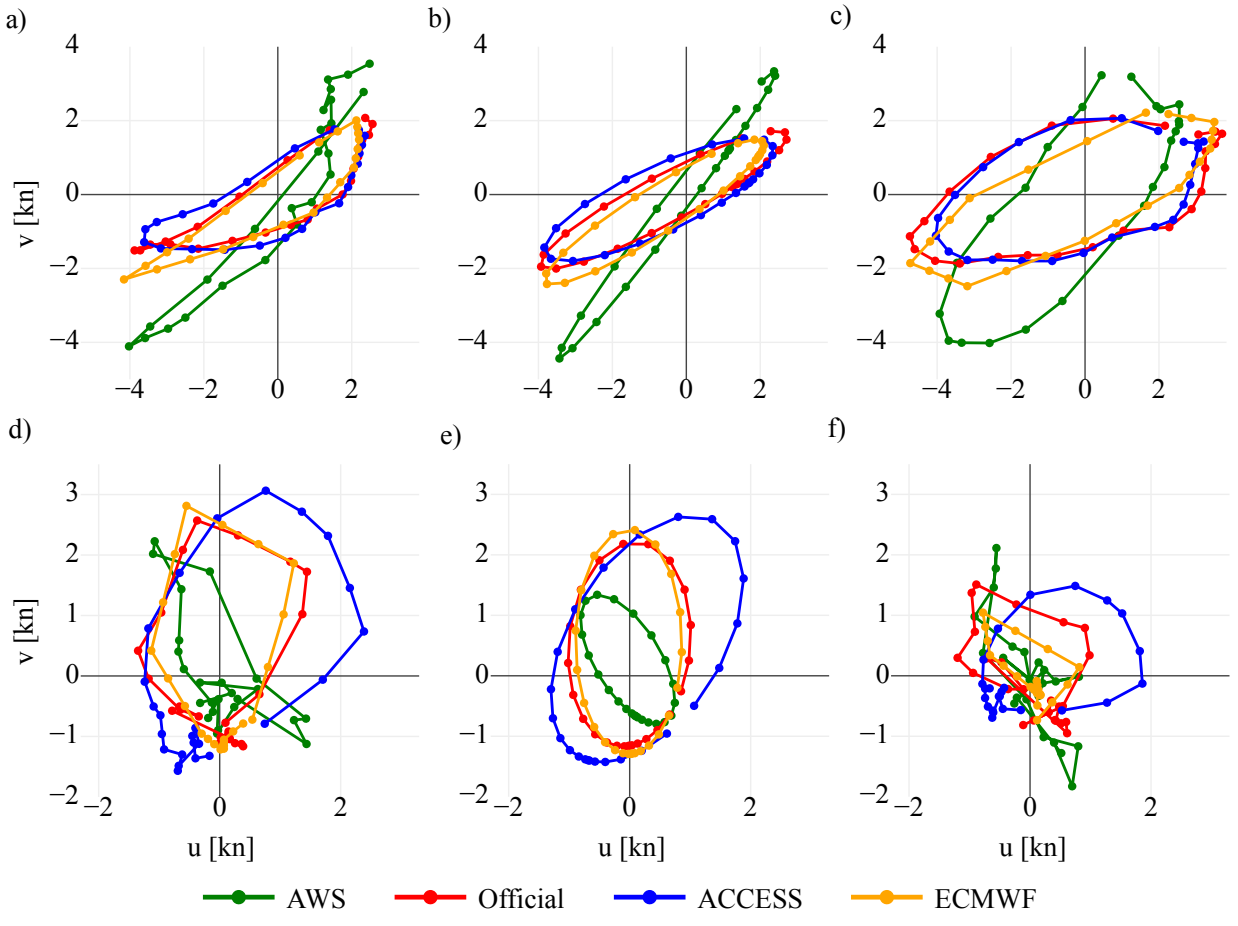
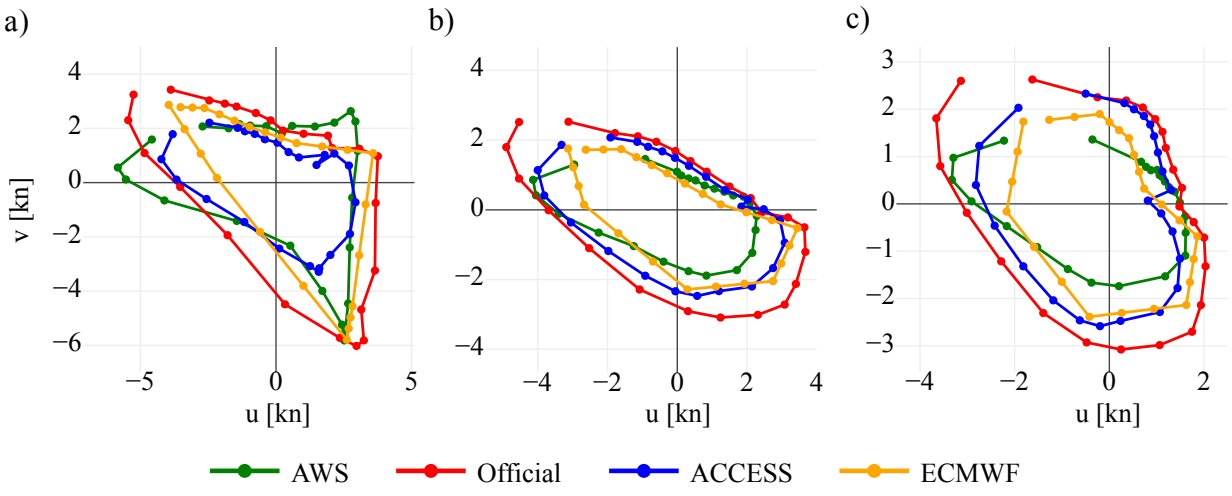


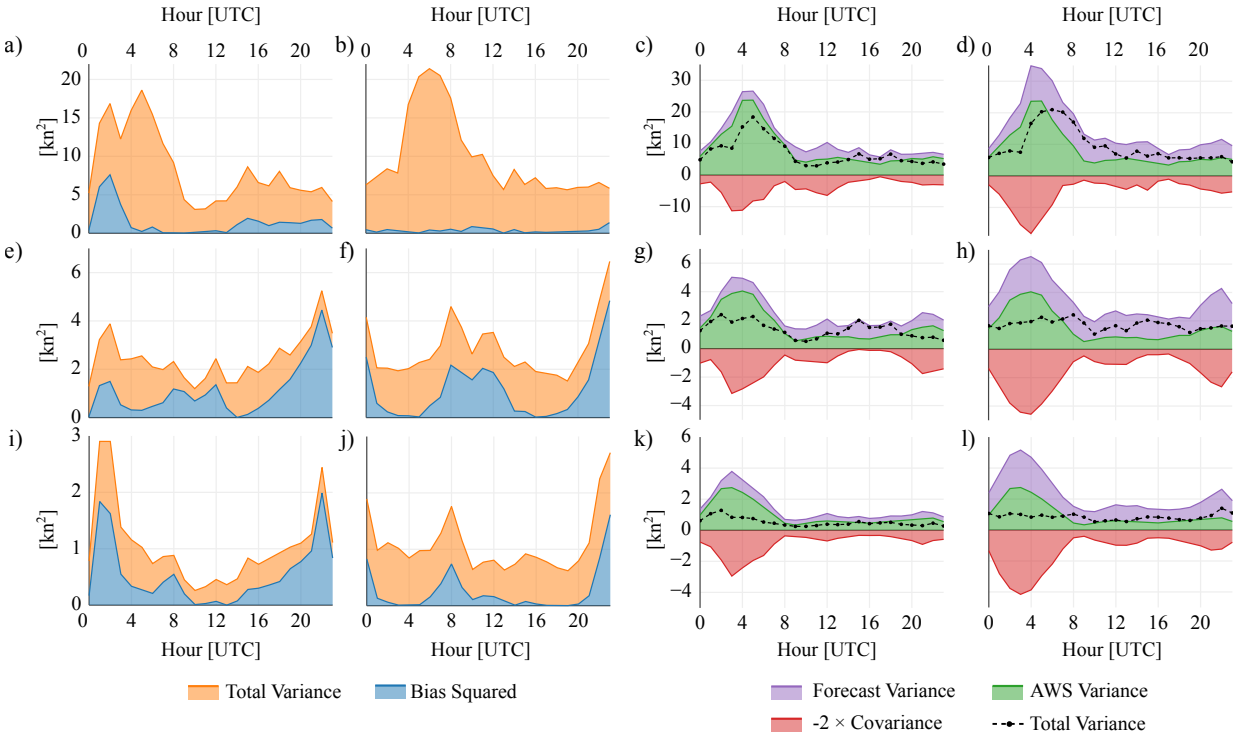
FIG. 12. Metrics derived from fitting the elliptical equations (5) and (6) to the climatological perturbations: maximum perturbation speed, a), e) and i), eccentricity, b), f) and j), orientation, c), g) and k), and time of maximum perturbation, d), h) and l), for the airport stations, a) to d), airport station groups, e) to h), and coastal station groups, i) to l).



790 FIG. 13. Hodographs of the climatological perturbations at Brisbane and Hobart airports, a) and d), and the as-
791 sociated ellipse fits, b) and e). For comparison, c) and f) provide hodographs of the climatological perturbations
792 at Spitfire Channel and Hobart (city), respectively.



793 FIG. 14. Hodographs of the climatological perturbations at, a), Darwin Airport, b) the Darwin Airport station
794 group, and c), the NT coastal station group.



795 FIG. 15. Mean square error between the ECMWF and AWS zonal perturbations $\overline{(u_{\mathrm{AWS}} - u_{\mathrm{E}})^2}$ decomposed
796 into the total variance $\mathrm{var}(u_{\mathrm{AWS}} - u_{\mathrm{E}})$ and squared bias $(\overline{u}_{\mathrm{AWS}} - \overline{u}_{\mathrm{E}})^2$ terms of equation (8), a), e) and i), and
797 analogously for the mean square error between the Official and AWS zonal perturbations $\overline{(u_{\mathrm{AWS}} - u_{\mathrm{O}})^2}$, b),
798 f) and j). Also, the total variance term $\mathrm{var}(u_{\mathrm{AWS}} - u_{\mathrm{E}})$ decomposed into the $\mathrm{var}(u_{\mathrm{AWS}})$, $\mathrm{var}(u_{\mathrm{E}})$ and $-2 \cdot$
799 $\mathrm{covar}(u_{\mathrm{AWS}}, u_{\mathrm{E}})$ terms, c), g) and k), and analogously for $\mathrm{var}(u_{\mathrm{AWS}} - u_{\mathrm{O}})$, d), h) and l). Decompositions given
800 for Darwin Airport, a) to d), the Darwin Airport station group, e) to h), and the NT coastal station group, i) to l).

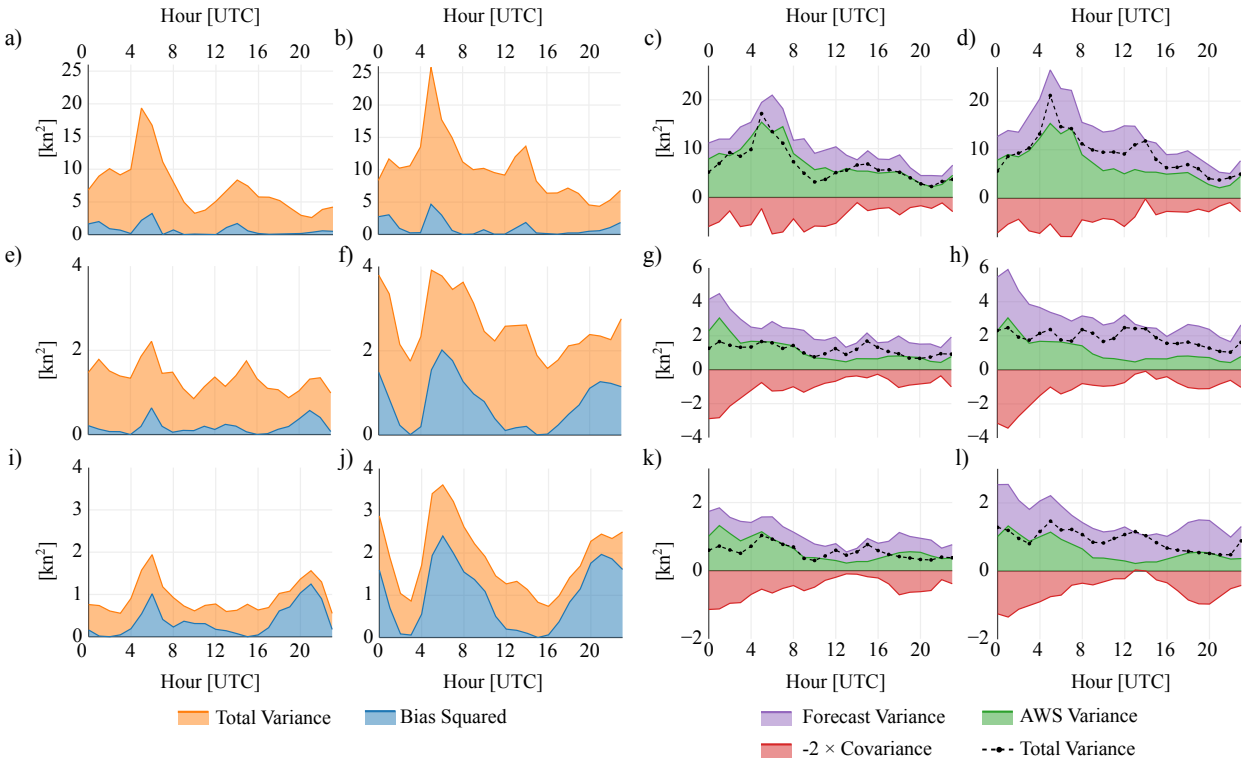


FIG. 16. As in Fig. 15, but for the meridional perturbations.