# Verifying Operational Forecasts of Land-Sea Breeze and Boundary Layer Mixing Processes: Response to Reviewers

### Ewan Short

# General Comments:

1. I would like to thank all Reviewers for their helpful comments which I believe have improved the manuscript considerably. I have addressed each of the comments and modified the manuscript accordingly. In addition to the responses to Reviewer comments below, I would like to note two additional changes.

First, in the original manuscript I described the “coastal station groups” as consisting of stations within 150 km of the nearest coastline: I meant to say within 100 km of the nearest coastline. This has been changed throughout the manuscript.

Second, I discovered a minor inconsistency in my code, which has been corrected. To calculate perturbations over the three month JJA 2018 period, I started with data for the 31/05/2018 to 01/09/2018 (inclusive), subtracted the 24 hour centred running means, then restricted the resulting perturbation datasets to 01/06/2018 to 31/08/2018 (inclusive). In a few places I discovered this restriction step was not working, resulting in the perturbations for 31/05/2018 (for the last twelve hours, where perturbations are defined) and 01/09/2018 (for the first 12 hours) being included in subsequent calculations. I have corrected this, and some of the numbers in the “scorecard” type figures have changed slightly as a result. These changes are not large enough to affect any of my arguments or conclusions.

# Reviewer 1:

Thank you very much for your helpful suggestions, particularly your comments regarding machine learning and A.I.

***1. From the explanation on "Edits", I hardly understand what kind of "edit" is this to improve the model performance! The author completely missed to explain clearly, what are Edits? Edits is only choosing the base model forecast or there are many more rules and regulations to guide the daily forecasts. How does Edits differ from Data assimilation?***

I am sorry the description of the editing processes was unclear. To construct the Bureau’s official forecast dataset, Australian forecasters use a two-step process.

1. Choose a model dataset to base the official forecast dataset on – the model dataset comes directly from the raw output from one of the available operational weather prediction models, which have (of course) already benefited from data assimilation through the initialisation and cycling processes. This first step is referred to as a choice of *model guidance*.

Manually *edit* the model guidance dataset using the Graphical Forecast Editor (GFE) software package. This manual editing is exactly that – they use a graphical program to subjectively change the automated forecast based on their own personal experience and known model errors, etc.

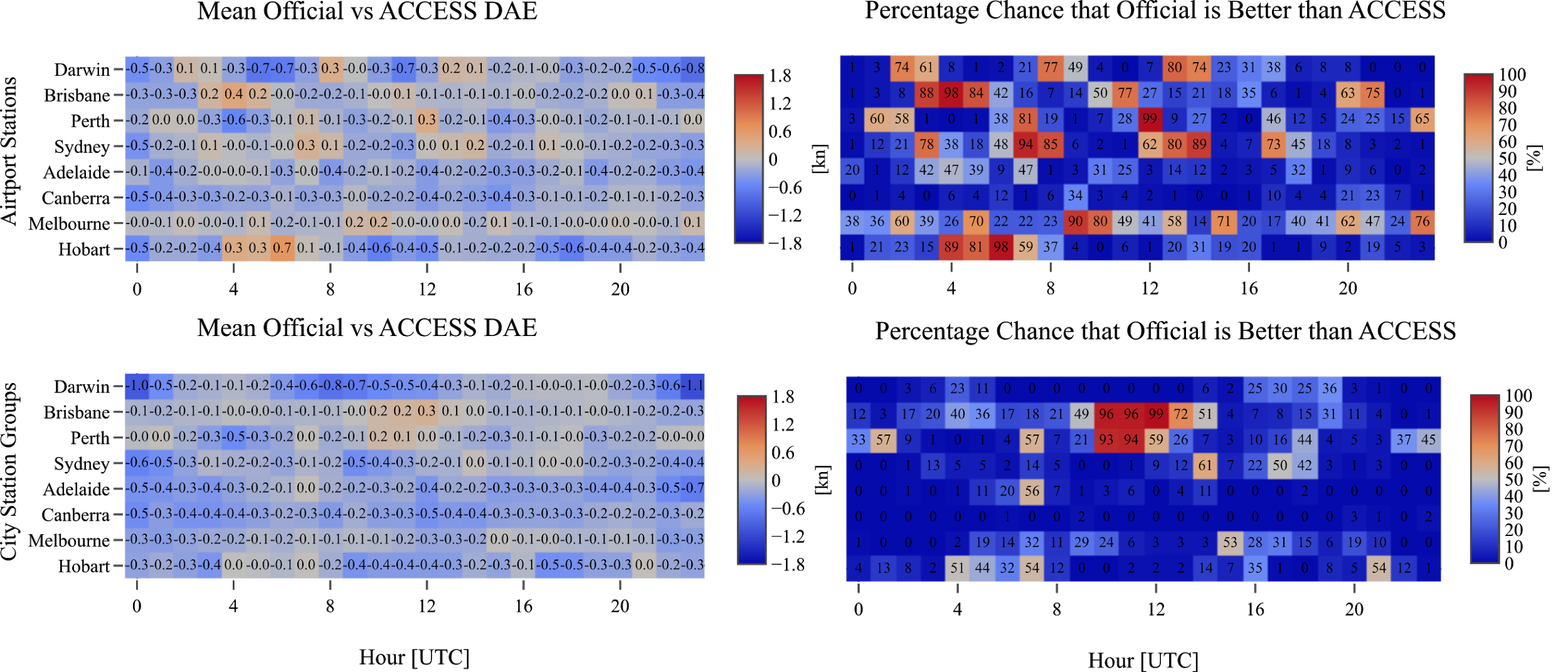
Different forecasting centres across Australia have different practices for precisely how the forecasters choose model guidance and make edits, and I don’t believe there are formal “rules and regulations” to guide this, just the routine structures and decision making processes that exist within each forecasting centre.

In the original manuscript I discussed edits in the abstract, and on lines 31-41, and lines 51-66. I have edited lines 9-11 of the abstract (same lines in revised manuscript) to better distinguish between “choice of model guidance” and “edits”. I have also re-written sections of the introduction for clarity, and restructured the order of paragraphs to ensure I finish discussing edits before moving on. I have also added the sentence “Australian forecasters themselves are rarely directly involved in model setup or post-processing, modelling is instead performed by other teams either within the BoM or internationally” to the second paragraph of the revised manuscript’s introduction (lines 36-38 new manuscript), to better distinguish the decisions forecasters make, from modelling decisions, such as choosing the method of data assimilation.

***2. Similar to Fig 6, how does Official vs ACCESS perform for Airport station and City Stations?***

Figure 1 (below) shows results analogous to Fig. 6 of the original manuscript, but for the Official versus ACCESS comparison. The results are similar to Fig. 6 in that at the city station group scale there are only a few times and locations where the official forecast unambiguously outperforms ACCESS. As with HRES, results are noisy for the individual airport stations. I noted this on lines 299-302 of the original manuscript (and lines 340-341 of the revised manuscript.)

I considered it unnecessary to include these figures, as space requirements meant I could only present a subset of the results, and I felt the more interesting comparison was between the official forecast and HRES. Following your helpful comments about machine learning and A.I. (as well as comments of other reviewers), I have decided to include results for the BoM’s gridded Operational Consensus Forecast (see response to comment 4 below) in the revised manuscript, and to make space I have further de-emphasised the ACCESS results somewhat.



**Figure 1:** Analogous to Fig. 6 in the original manuscript, but for the official forecast versus ACCESS mean DAE results.

***Official Vs HRES shows very low confidence over most of the locations.***

The confidence score provides the probability that the population or “true” mean difference of absolute errors is greater than zero. A score near 0% implies *high* confidence that this mean is less than zero, whereas a score near 100% implies *high* confidence it is greater than zero. A score between, for instance, 10-90%, implies low confidence either way. This is described in the caption to Fig. 2 and on lines 177 to 186 of the original manuscript (and Fig. 3 and lines 199 to 216 of revised manuscript). Out of context the term “confidence” may be confusing, but read in the context of the caption and the description in the text I believe it to be appropriate.

***DAE (Difference in Absolute error) is a very simple skill score and been obsolete. What about the skill scores of forecasts based on probabilistic forecasts, e.g Ranked Probability Skill Score (RPSS)?! It is observed that most of the time probabilistic skills are superior than dynamical skill.***

Unfortunately the Bureau doesn’t issue probabilistic wind forecasts, so metrics like the Ranked Probability Skill Score cannot be used in this study. However, I agree that probabilistic forecasts are desirable.

Note that although absolute error of a scalar field is indeed a very simple measure of skill, in this study I am dealing with the difference of absolute errors (with absolute error the Euclidean distance between wind perturbation vectors), then considering means of this quantity, with statistical methods used to determine whether differences in the mean error of each dataset are statistically significant. Measuring error is certainly challenging at mesoscales, which is why I pursued a “fuzzy verification” approach, aggregating data over different spatial scales before considering absolute errors. I believe this is somewhat more sophisticated than a simple calculation of absolute error of a scalar field.

There are innumerable ways I could have made the metric more complex, for instance by applying the Huber loss function to the terms in equation (1) before taking a difference, or to the overall expression. Such options grow rapidly, and given that I hadn’t seen anything else in the literature similar to what I was doing, I decided to start simple, and leave it to others to experiment with more complex variations of the metric.

***3. HRES and ACCESS carry lesser error and higher confidence (Fig 7) and compared to Edits. A method which is inferior in the forecast, in general been rejected. This research work shows the forecasts skills are very random in nature. Thus, with the facts shown in the illustrations, it is a bit hard to trust in these forecasts.***

Crucially, it is just the *diurnal component* of the wind forecasts that are being assessed, not the overall wind fields. Furthermore, the metrics I consider assess just *some aspects* of the wind forecasts. For these reasons I am reluctant to make broad assertions/recommendations about weather forecasting practice, or the value of different techniques and models, beyond what is presented in the conclusion section.

***4. Knowing the "edits" are constrained to and much depended on the expertise of the human, who performs the edits. I think, Machine learning and AI will be superior options to choose a base model! However, the error of model bias and random variability in the forecast by edits will retain questionable?***

The most substantial change I have made to the manuscript is to include results on the Bureau of Meteorology’s gridded Operational Consensus Forecast (OCF) model guidance product. This product represents the Bureau’s goal of automating the choice of model guidance, and provides forecasters with an objective alternative to their subjective evaluation of different model guidance products. Description of OCF can be found on lines 125-143 of the new manuscript, with OCF results now present in most of the figures. OCF results are discussed throughout the revised manuscript, but particularly on lines 404-414 (see tracked changes document for other changes).

# Reviewer 2:

1. Thank you for your kind and encouraging comments!

***In Figs. 2, 6-8, 10, it might be very helpful to convert the abscissa to Local Time, rather than UTC. Additionally, the groupings along the ordinate should be arranged from geographic locations in the east near the top, those in central Australia in the middle, and those in western Australia at the bottom. This follows the natural march of the sun across the sky during the day.***

***In its present form, the figures require the reader to mentally translate UTC to local time for each of the geographical groupings, particularly for those readers who reside outside of Australia. And the choice or order of geographic groupings is seemingly random, with Brisbane (far east) plotted 1 row above Perth (far west). This must be contributing to the "noisiness" in the plots, which the author frequently notes throughout the text.***

***I am motivated by the desire for the reader to able to instinctively determine the points on those plots where sunrise and sunset occur, and for the rows along the ordinate to be logically arranged by time zone, following the march of the sun. Doing so will make these plots far more interpretable, and far more accessible. Also, I have the sense that by re-ordering the abscissa to local time is likely to reduce the "noisiness" in the plots.***

This is an excellent suggestion. For context, my initial choice of UTC and layout were motivated by a few factors. First, during my time at the Darwin forecasting centre in July 2018, I noticed that forecasters seemed to think and work in terms of UTC, and so figured my results might be more intelligible to them if I used UTC in place of my preferred local solar time (LST). Second, because the coastal station groups are averages over many degrees of longitude, reporting results in LST would require choosing an “average LST” value for each group. Finally, to plot results on a single LST horizontal-axis, different LST values for each location would have to be interpolated, or approximately matched with a single set of LST values; interpolating risks additional and inconsistent smoothing of data at different locations, and a “nearest-neigbour” type matching would result in a loss of temporal precision. As to my choice of row layout, this was done based on approximate latitude of each group, as one of my initial questions with this work was to see whether latitude played a role in verification outcomes, given the significant effect of latitude on land-sea breeze dynamics (e.g. Rotunno 1983).

However, I entirely agree that given how the results turned out, it makes much more sense to order rows by longitude, and this has been done. Furthermore, I agree that providing an indication of LST is vital, but for the reasons given above I am uncomfortable translating all results onto a common, approximate, LST horizontal axis. As a compromise, I have instead opted to display both UTC and LST on the relevant plots, with Perth LST displayed along the top line, and Brisbane LST displayed along the bottom line. I believe this is sufficient to give readers an immediate sense of the solar heating cycle when interpreting these figures. Furthermore, I have also included LST in Fig. 14 (Fig. 15 of the revised manuscript.)

# Reviewer 3:

Thank you very much for your constructive comments, which I believe have improved the manuscript substantially. Please see below for a detailed response.

***I would like to thank the author for submitting his manuscript for review. The manuscript contains some interesting concepts, but at this stage, I do not believe it is ready for publication in Weather and Forecasting. My reasons for this recommendation are outlined below.***

***a. Does the paper fit within the stated scope of the journal?***

***Yes***

***b. Does the paper identify a gap in scientific knowledge and add new knowledge to the overall body of scientific understanding?***

***The paper does identify the importance of verifying human edited forecasts.***

***c. Is the paper free of errors in logic?***

***Yes.***

***d. Do the conclusions follow from the evidence?***

***I have some concerns about the methodology. Until these concerns are addressed, I do not know how much weight to put on the results from which the conclusions are drawn.***

These concerns are well articulated and valid. I should have anticipated and addressed them in the original manuscript. Please see the response to major comments 2 and 3 below.

***e. Are alternative explanations explored as appropriate?***

***Yes.***

***f. Are biases, limitations and assumptions clearly stated, and uncertainty quantified?***

***I am not sure the author is aware of the limitations in his methodology. For example, the spatial domains he defines are of widely varying sizes, and I think this may well make comparison of the results from different locations difficult.***

This is a valid point. Please see the response to major comment 3 below.

***g. Is the methodology explained in sufficient detail so that the paper's scientific conclusions could be tested by others?***

***I think the methodology probably could be reproduced by others. But the author could have made his descriptions of the method much easier for readers to understand.***

This is a good point. I attempted to strike a balance between length, detail, completeness, and readability with this manuscript. I have reworked many parts of the text (see tracked changes document), and simplified or improved all of the figures except Fig. 5 (Fig. 7 of revised manuscript) which shows the vertical soundings. I have also included a schematic in the methods section illustrating the fundamental ideas.

Regarding reproducibility, I believe the revised manuscript makes my methods and results clearer, particularly the improved layout and labelling of figures. My code is also freely available online (with the details of how to obtain it within the manuscript).

***h. Is previous work and current understanding cited and represented correctly?***

***Mostly. But the author either ignores or is unaware of the current forecaster practise of using a gridded consensus forecast as the starting point for manual edits.***

An excellent point. Please see the response to major comment 1 below.

***i. Is information conveyed clearly enough to be understood by the typical reader.***

***I think this is a major weakness in the paper. The method and results could be conveyed much more clearly. This paper is difficult to read.***

I am sorry you found the manuscript difficult to read. I have reworked almost all the figures, and added a new schematic to the methods section to illustrate the essential ideas regarding perturbations and DAE. I have produced a better labelled map. I have simplified the way I use acronyms, and simplified the terminology I use. I have also added a new paragraph to the methods section (lines 246 - 255 of revised manuscript) to better explain the appropriate way to view wind perturbations, DAE, and their means.

***j. Are all the figures and tables necessary, appropriate, legible and annotated (as appropriate)?***

***No. I have provided a minimum set of improvements for the figures in the comments below.***

I apologise you found the figures difficult to understand. I have amended all the figures, except Fig. 5 (Fig. 6 revised manuscript), which show the vertical soundings. Colormaps in the “scorecard” type figures have been changed, locations reordered based on longitude, and indications of the local solar time (LST) provided. I have reduced the number of examples presented, and better subdivided figures based on station group. Figures are now labelled more extensively, providing better indications of location and station group.

***Major comments***

***1. The author correctly identifies the procedures that Bureau of Meteorology forecasters use to prepare their official forecasts. First gridded guidance is loaded into the Graphical Forecast Editor (GFE), and then the forecaster makes manual edits to the guidance. It is correct that the forecasters have access to several NWP models for initial guidance, including the ACCESS-R and ECMWF models. However, additionally, there is a consensus forecast available called Gridded OCF (Operational Consensus Forecasts). This is the guidance which forecasters are now expected to use as their first guess in most situations. The Gridded OCF system is described in several "Operations Bulletins" available at:***

***http://www.bom.gov.au/australia/charts/bulletins/nmoc\_bulletin.shtml***

***In particular, Bulletin 91 may be of interest to the author:***

***http://www.bom.gov.au/australia/charts/bulletins/apob91.pdf***

***For the period of the author's study (mid 2018), forecasters may have used ACCESS-R and ECMWF guidance, but they will also have utilised the Gridded OCF guidance. Unfortunately, it is not clear if there are records of which guidance the forecasters used for any particular official forecast. Given that forecasters now mainly rely on Gridded OCF guidance, it is disappointing that the author didn't include this data in his study.***

A valid point. For context, this study began as an internship at the Darwin Regional Forecasting Centre in July 2018. During that time the forecasters I spoke with told me the HRES and ACCESS models were their most common model guidance choices for winds, and that OCF wasn’t commonly used for winds. At a few points in the original manuscript, I implied this is still true, or true across Australia, which I should not have done.

I have therefore revised the manuscript to incorporate the OCF results, which are now present in almost all the revised figures, and are discussed throughout the manuscript (see tracked changes document). A summary of the OCF method is presented on lines 125-143 of the new manuscript. I reference the foundational work of Woodcock and Engel (2005), Engel and Ebert (2007), and Bureau Operations Bulletins numbers 60, 74, 91 and 113. To make space I have removed figures that compare HRES to ACCESS, and reduced the number of specific examples I discuss. Because I drafted the original manuscript so that my arguments and conclusions would be appropriate when considering just a subset of the model guidance products available to forecasters, the overall conclusions have changed little with the inclusion of OCF.

However, OCF exhibits amplitude biases in its mean diurnal cycle (subject to how I define and measure mean diurnal cycles in the study), particularly at the individual airport station scale. These results are interesting in their own right, and could be useful if OCF is now the go-to choice of model guidance for winds. These results are discussed in the revised manuscript, particularly on lines 404 to 411. However, on lines 411 to 414 I am careful to emphasise that just because OCF’s mean diurnal wind cycle, as I define it in the study, is suppressed, this does *not* mean that OCF’s overall wind speeds or directions are biased. On lines 141 to 143 of the revised manuscript I take care to point out that OCF produces lower errors in both wind speed and direction of the overall wind field than all of the individual model datasets that comprise it.

I was aware of this mean diurnal cycle amplitude result for OCF, at least for Darwin, very early in the process of conducting this research. Fig. 10 a) of the revised manuscript was one of the first figures I produced during my internship in Darwin. I reasoned that OCF’s suppressed mean diurnal cycle was one reason forecasters may not favour it for winds, at least in northern Australia where diurnal processes are more significant. This speculation, combined with the added complexity I believed considering OCF would add to an already complex manuscript, was why I neglected it from my original draft.

A challenge in conducting this research has been that records of which model guidance was used on a given day, and what types of edits were performed and why, were not kept, or at least not available to me while working with the BoM’s verification team. I believe verification studies like the one I have attempted here could be made much easier if such records were available.

***2. The difference of absolute errors metric (pages 7 and 8)***

***This metric is interesting. However, I have some concerns about its applicability to a vector wind field. In meteorology, wind fields commonly contain sharp discontinuities at fronts. It is not uncommon for the wind direction to change by very large amounts in a period of minutes.***

***The author attempts to identify diurnal cycles by subtracting a twenty four hour centered running mean from the observed or modeled wind. But consider this hypothetical situation: the first twelve hours of winds are northerlies at a constant speed, and the next twelve hours are southerlies with the same speed. The twenty four hour vector mean will be zero, and the perturbations at each hour will be quite large. I question whether in this situation the perturbation is of use for identifying diurnal cycles.***

***The above considerations aside, the difference of absolute errors metric may still be of value. The author may find that it would be possible to produce a shorter paper focusing solely on the applicability of this metric for weather forecast verification.***

This is a valid point that I should have addressed directly in the original manuscript. The idea is *not* to use metrics like DAE to verify performance on individual days (and I acknowledge the NT and south WA case studies I presented early in the results section of the original manuscript may have been misleading in this regard), but to consider means of such metrics, with statistical methods employed to assess whether positive or negative mean values actually represent differences in mean error between the underlying datasets. Taking hourly means of perturbations or of DAE significantly reduces the impact of non-diurnal processes, like sudden changes in background winds. This was also why the “case study” discussions in the results section considered the overall wind field in conjunction with the perturbations and the DAE scores. Moreover, while it is certainly true that sudden changes in background winds induce large perturbations, provided such sudden changes occur at similar times in each dataset, they largely cancel in DAE. This was part of the motivation for considering DAE, and restricting the study to lead day 1 forecasts.

To make these points concrete, it is instructive to look at how the metric behaves with synthetic data. Consider data with an hourly time domain over a three month period, like that of the study: , 24\*92. Suppose a hypothetical automatic weather station (AWS) has recorded zonal wind data of the form,

,

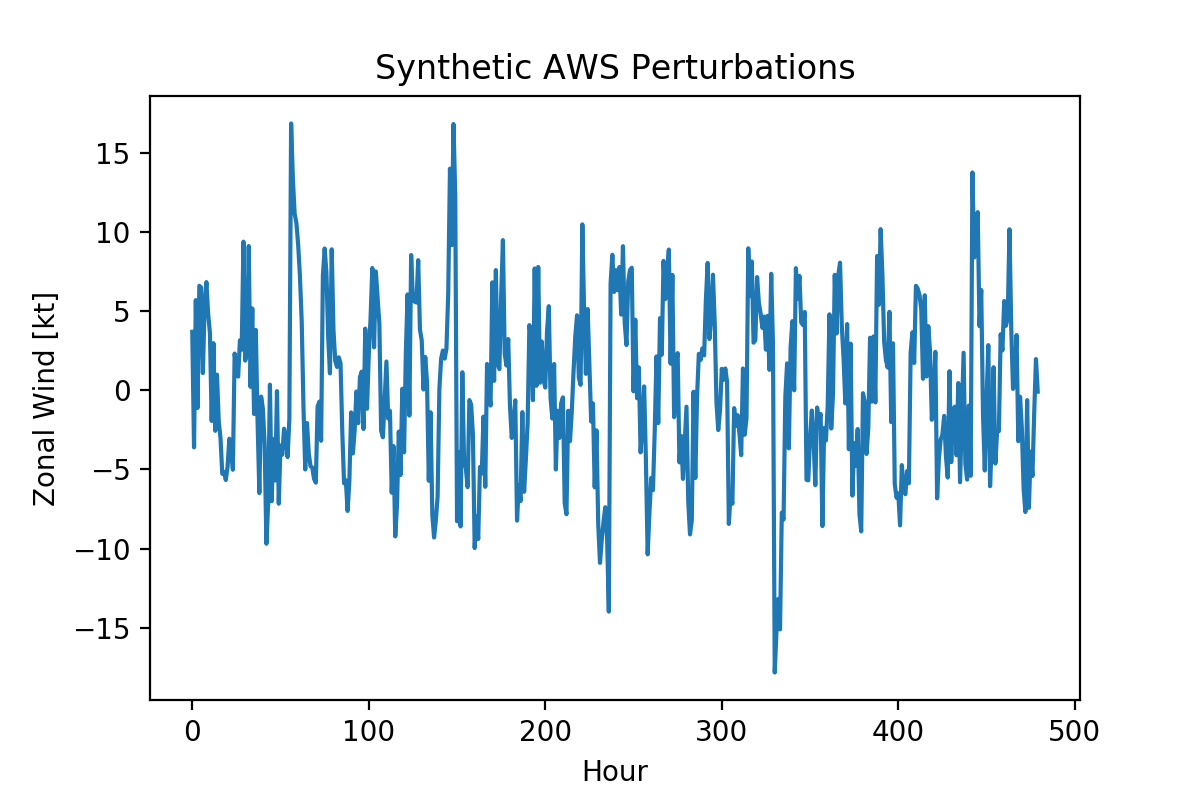
where , , and denote the “diurnal”, “noise” and “synoptic” parts of the signal respectively. Suppose the noise signal is sampled from a uniform distribution across the range , and that the diurnal signal is given by,

,

where denotes the amplitude of the diurnal signal, and is the phase offset in hours. Suppose we define so that it represents sudden reversals in the background wind direction, as in the example provided by the reviewer. Suppose flips between , with the number of hours between direction changes sampled independently from a normal distribution with a mean of hours, and a standard deviation of hours (in the unlikely event a number less than 1 is sampled, replace it by 1).

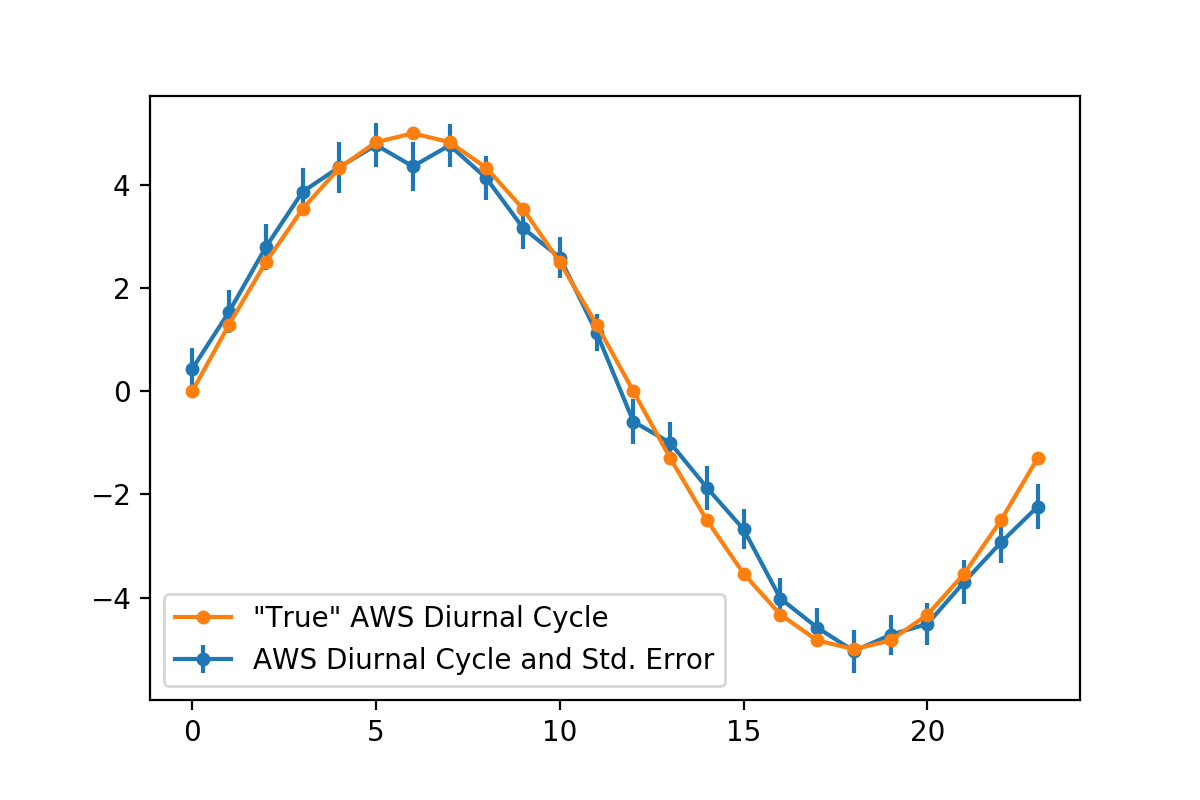
As an initial example, let , kt, kt, and let hours and hours. With defined, we can then calculate perturbations by subtracting a 24 hour centred running mean (discarding the first and last 12 hours where perturbations are undefined). The first 20 days of these synthetic perturbations are shown in Fig. 2.

**Figure 2:** First 20 days synthetic AWS perturbation data for the sudden background wind shifts example.



As the reviewer correctly points out, perturbations are large when the background wind changes suddenly. Despite this, when we average data from each hour of the day we recover the original diurnal signal. This is shown in Fig. 3 below. This idea underlies the “differences of biases” metric defined on lines 207-210 of the original manuscript (lines 237-239 of revised manuscript).

**Figure 3:** “True” diurnal cycle given by *d*(*t*) (orange), and approximate diurnal cycle resolved by taking means of the perturbations at each hour (blue).

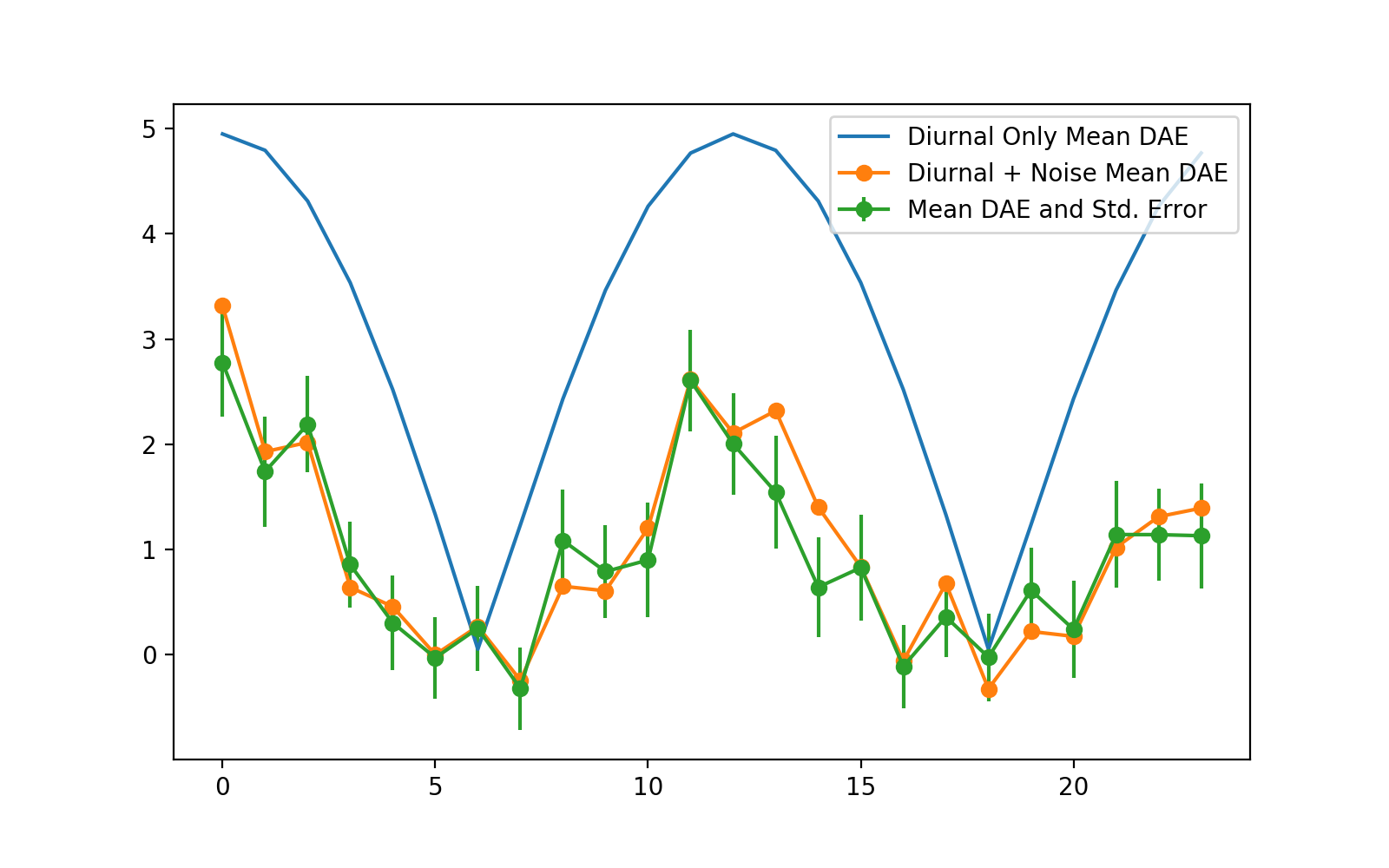
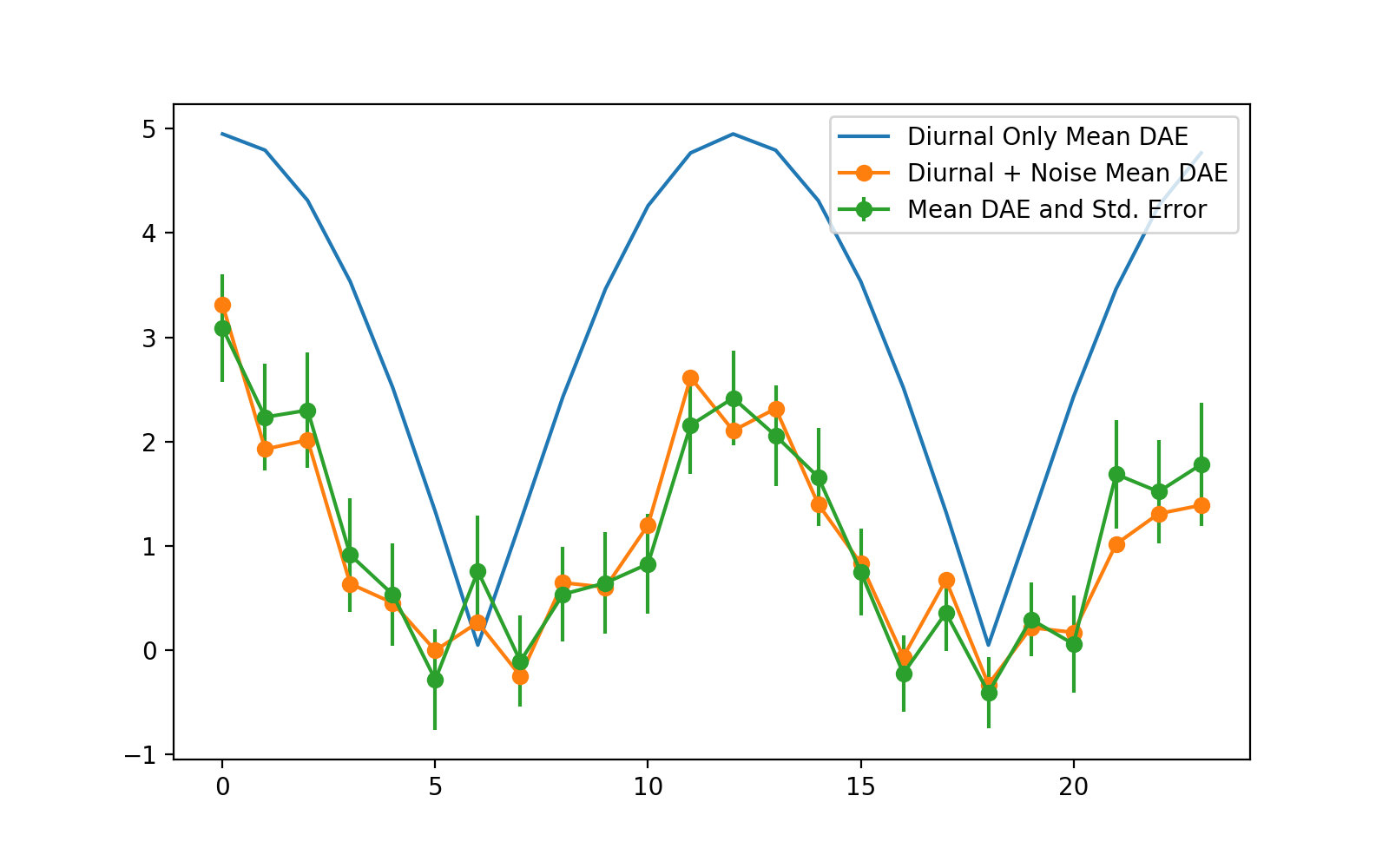


Hour

Mean Perturbation [kt]

To study the effect of these sudden wind reversals on the mean DAE, consider now synthetic official forecast and model guidance zonal wind data, denoted and , respectively, and generated analogously to that for above. Suppose has the same parameters as , but it’s own independently generated noise signal. Suppose the phase of the diurnal component of the synthetic model guidance data is offset two hours from that of AWS and the official forecast, i.e. hours, and that its amplitude is 2 kt higher, i.e. =7 kt. As with , assume has its own independently generated noise signal. Use the same random timing between background wind direction changes obtained for for both and but add random errors to these timings for both and , independently sampled from a normal distribution with a mean of zero, and a standard deviation of 2 hours. This is intended to simulate a situation where all datasets produce a sudden change in large-scale winds at roughly the same time, but with some timing errors.

**Figure 4:** Mean DAE scores calculated from full perturbation datasets (green), just the diurnal and noise components (orange), and just the diurnal components (blue). Results when “synoptic” wind reversal timing errors have, a), standard deviations of 2 hours, and b), standard deviations of 48 hours.



a)

b)

Hour

Mean DAE [kt]

Mean DAE [kt]

Fig. 4 a) shows the mean DAE scores from the synthetic datasets (green curve), as well as the scores that result when applied to perturbations calculated from just the diurnal components of each dataset, (blue curve), and the sum of diurnal and noise components, (orange curve). Note that the overall mean DAE, and the mean DAE with the “synoptic” signals removed (i.e. no sudden changes in background wind), are very similar. Recall that the DAE is defined

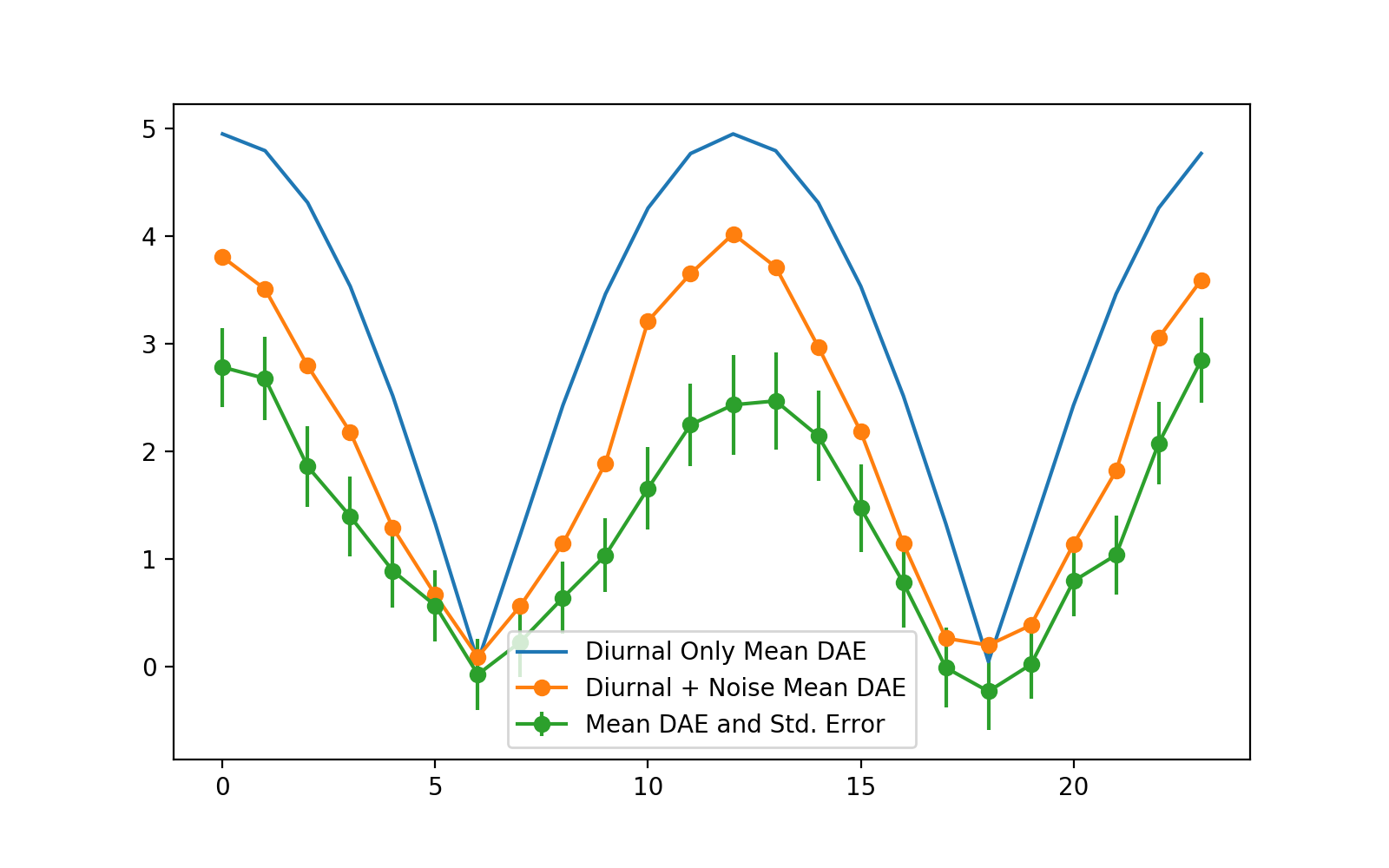
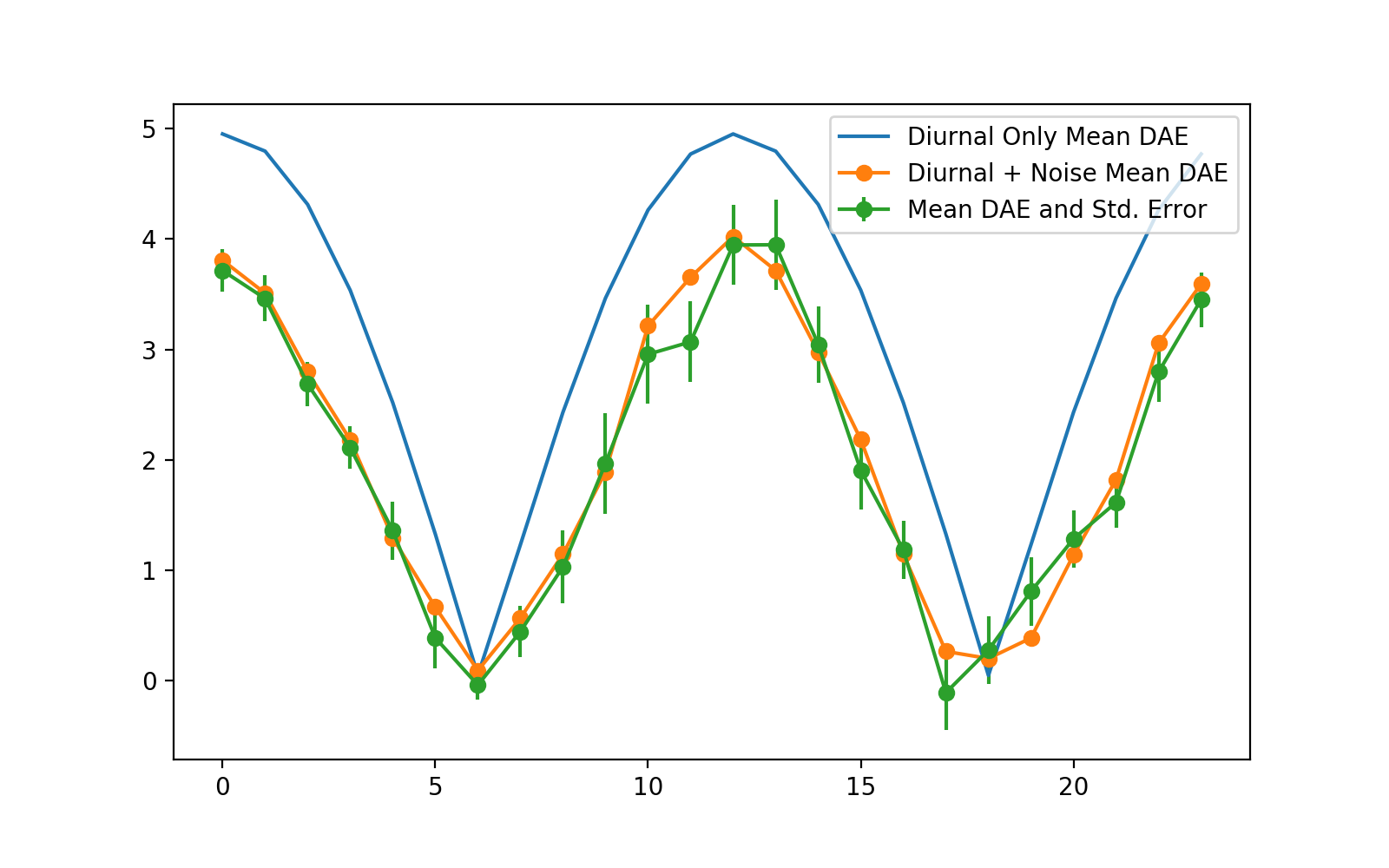
DAE =

While sudden wind reversals will certainly produce large perturbation values, provided these reversals occur at similar times in each dataset, they largely cancel out: this was part of the motivation for considering DAE in this study, and focusing solely on lead day one forecasts.

Furthermore, Figure 4 b) shows the analogous results to Fig. 4 a), but after adjusting the sudden wind shift timing errors for the official forecast and model guidance synthetic datasets to have standard deviations of 48 hours, so that the wind shifts occur at completely different times for each dataset. Note the effect on mean DAE is minimal, because in this example the noise terms dominate the “synoptic” terms when the mean of DAE is taken.

If we repeat the above experiment, but with the amplitudes of the noise terms in each dataset reduced to 2 kt, we obtain the mean DAE results shown in Figure 5: here the increased “synoptic” timing errors have a greater effect on mean DAE due to the reduced amplitude of the noise signal. In Fig. 5 b), the wind shift timing errors have a standard deviation of 48 hours, much larger than what we might expect for real synoptic frontal wind shifts at lead day 1. Despite this, the overall mean DAE (green curve) is still of the same sign as the mean DAE with no sudden wind shifts (orange curve), and evolves in a very similar way, implying it still contains useful information about errors in the actual diurnal signals of each dataset, not just the timing errors of the sudden wind direction reversals. This extreme example illustrates mean DAE may actually be more robust than I originally thought, and might be appropriate to apply beyond lead day 1.

**Figure 5:** As in Fig. 4, but with the amplitude of the noise terms in each synthetic dataset reduced to 2 kt.



a)

b)

Hour

Mean DAE [kt]

Mean DAE [kt]

Note the code used to generate the examples in this response is available online. All the parameters can be freely varied to examine how this affects the difference between hourly means of perturbations and the “true” underlying diurnal cycle (Fig. 3 above), and mean DAE (Figs. 4 and 5 above). The code also allows other choices of the “synoptic” signal to be tested, such as taking to be a sin curve.

While these examples are limited, in particular they only consider purely zonal winds, I believe they adequately illustrate the fundamental ideas, and could be extended into an independent study that would be very interesting in its own right. I note this on lines 560 to 563 of the revised manuscript.

I provide a summary of the ideas discussed in this response on lines 246 to 255 of the revised manuscript, as well as a reference to the code used to produce the above examples. I have also added a new sentence at lines 308 to 310 of the revised manuscript, reinforcing the limitations of studying DAE on specific days (i.e. looking at DAE rather than its mean), and have softened the language throughout the subsequent “case study” discussion (see tracked changes document).

Finally, I acknowledge that the concept of a “diurnal cycle” makes less sense on individual days than when temporal means of perturbations are considered. Certainly the perturbations on a single day do not necessarily represent a diurnal “cycle”. I have therefore replaced “diurnal cycle” with “diurnal signal” where appropriate throughout the manuscript (see tracked changes document).

***3. Coarser spatial scales (city station groups and coastal station groups)***

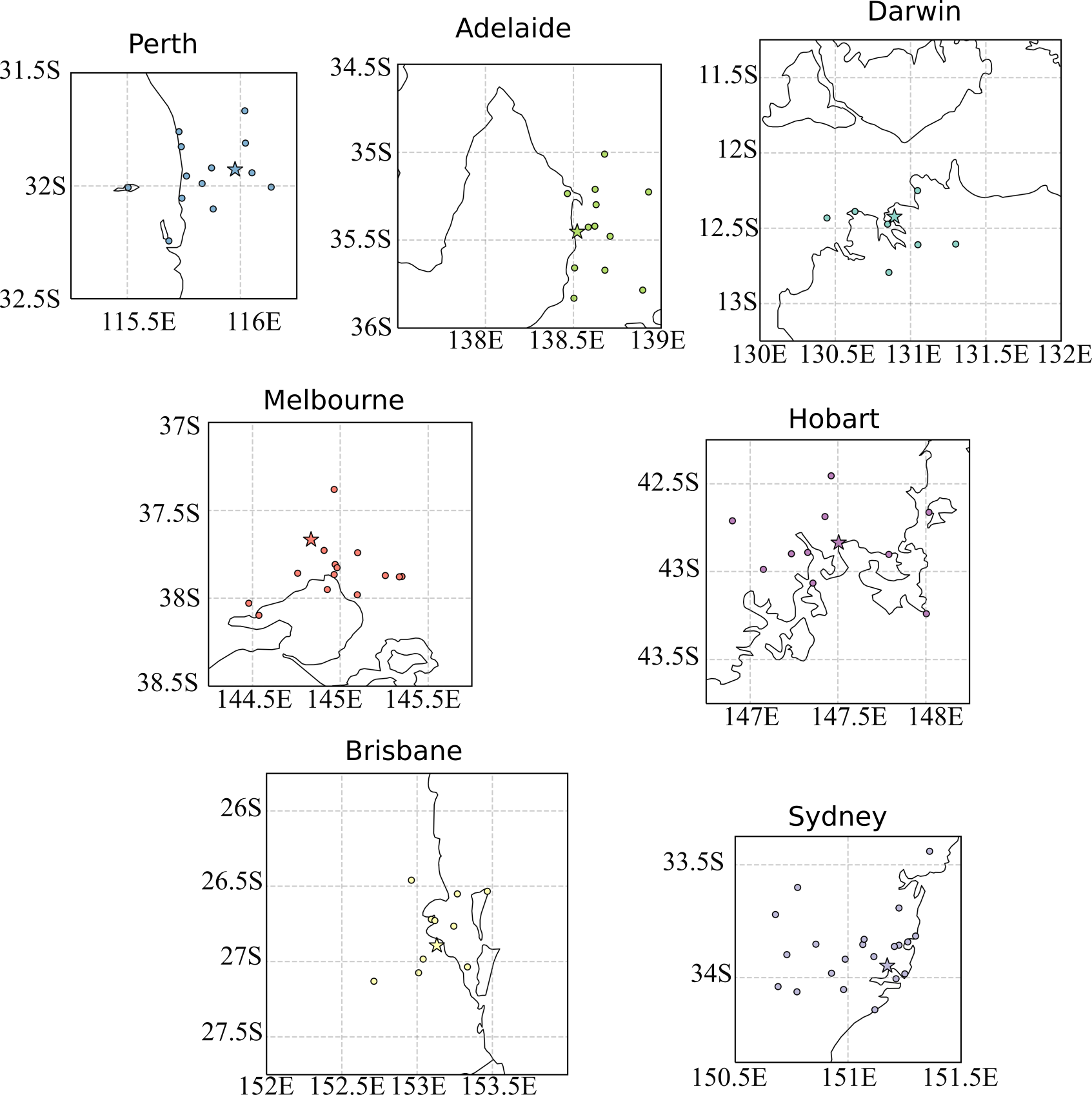
***The author looks at a number of spatial scales. The city station groups comprise the ten stations closest to each capital city. The coastal station groups comprise all stations within 150 km of the nearest coastline. I have the following concerns:***

***- The area of the city spatial groups varies quite a lot. The largest city spatial group is three or four times the area of the smallest group. Is comparison of statistics from these differing sized groups valid?***

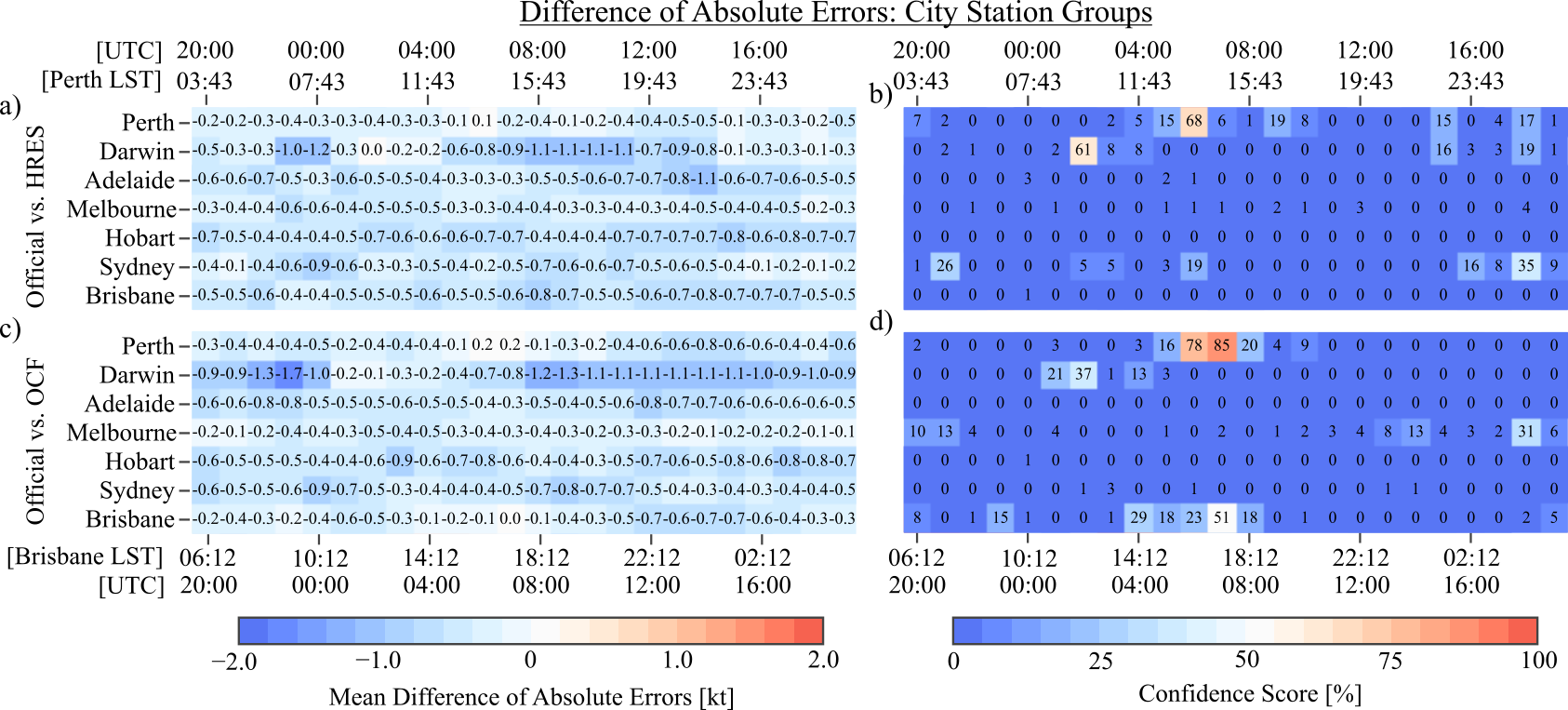
A very good point. I did consider this issue, but reasoned that if I enforced approximately constant areas in each station group, I’d have inconsistent numbers of stations in each group. To then get consistent numbers of stations in each group, I’d have to arbitrarily choose a subset of the stations in groups containing more stations, to match the number of stations in the group with the least.

To test the effect of alternative choices for the city station groups, I have produced new results where the “city station groups” are redefined to include all stations within 100 km x 100 km boxes centred on each capital city airport (see Fig. 6 below). The results for the mean DAE comparisons are presented in Figure 7 below: they are basically the same as those in Figs. 6 c) and d) of the original manuscript, and Fig 7 of the revised manuscript, with model guidance outperforming the official forecast for the overwhelming majority of station groups and times. This has been summarised on lines 347 to 350 of the revised manuscript.

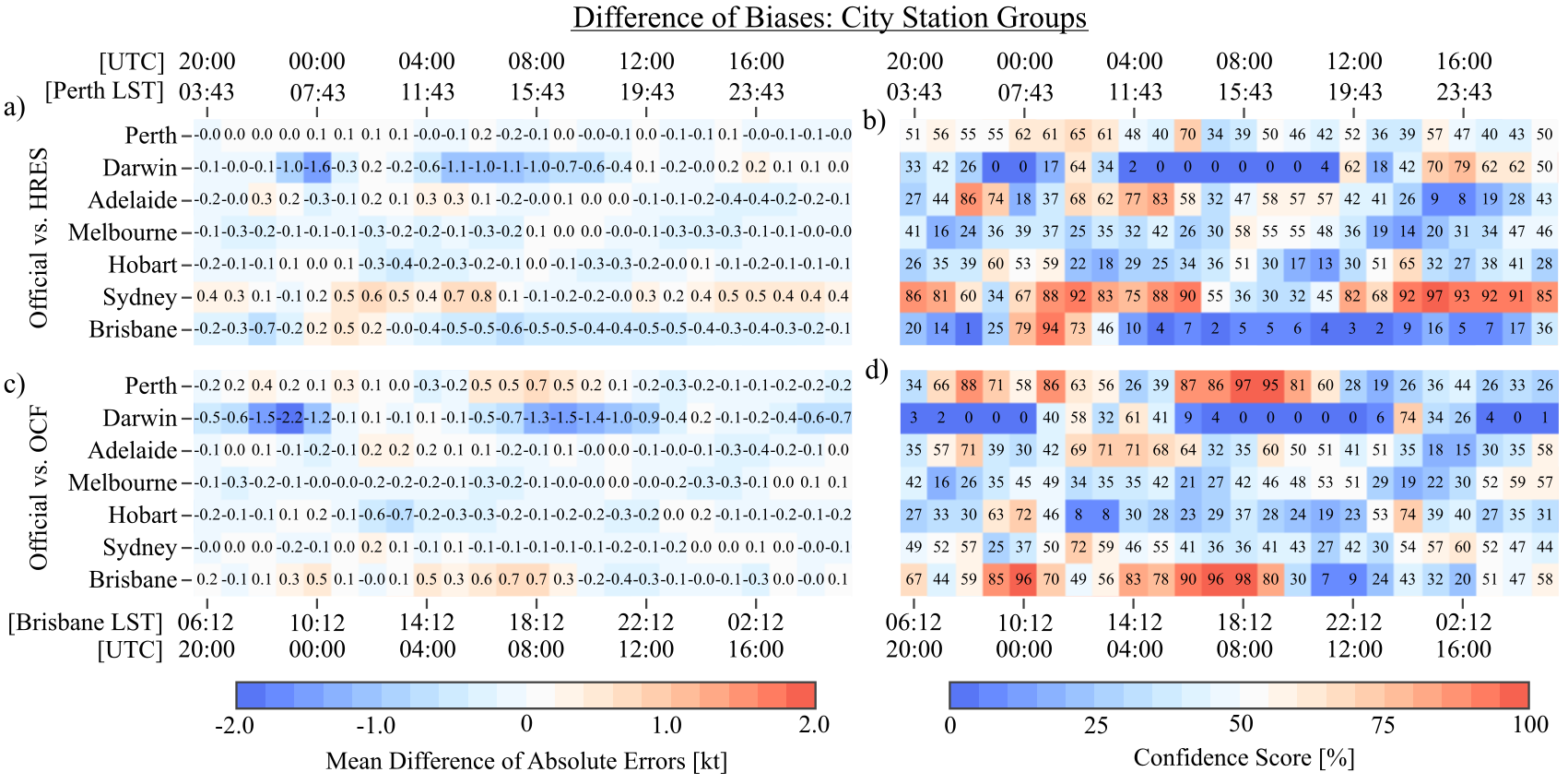
Figure 8 provides the analogous DB (difference of biases) comparisons for the alternative city station groups: the results are very similar to those of Figs. 10 c) and d) of the original manuscript and Fig. 12 of the revised manuscript. For both DB and mean DAE, the largest differences appear to be at the Darwin city station group. Here, the new definition results in the Tiwi Island (large islands north of Darwin airport) stations being dropped, likely explaining the changes at this station group.



**Figure 6:** Map of the alternative “city station groups”, defined as all the stations within 100 km boxes centred on each capital city airport.



**Figure 7:** Alternative city station group mean DAE results, where the alternative city station groups are formed by taking all stations within a 100 km x 100 km box centred on each capital city airport. Compare this with Fig. 7 of the revised manuscript.



**Figure 8:** As in Fig. 7 above, but for the difference of bias (DB) results. Compare this figure with Fig. 12 of the revised manuscript.

Importantly, comparing results at different locations is *not* intended as a “ranking” of the performance of forecasting centers at different states, or anything similar, but merely as an exploration of the physical differences in the diurnal wind cycles at different locations, and how this is reflected in the various metrics considered in the manuscript. In addition to inconsistencies in area, or number of stations in station groups, diurnal wind cycles are of significantly different amplitudes in different areas, making it much more difficult in some regions for forecasters/models to simulate it. I have added sentences at lines 282 to 286 of the revised manuscript summarising this point.

***- It is debatable if Canberra can be considered a coastal area. It is too far from the sea to be subject to sea breezes.***

I have removed the Canberra Airport and Canberra City station group results from the manuscript.

***I believe the above concerns need to be adequately addressed before the results can be published.***

***Minor comments and typos***

***There are numerous minor concerns with the paper. Some are listed below:***

***1. Line 98 and elsewhere: I believe the ACCESS model the author is using is ACCESS-R. There are other configurations (a global ACCESS-G and high resolution ACCESS-city models as well, so it is important to specify which model configuration is being used.***

I have changed the sentence “ACCESS is a nested model: in this study we consider the component covering the Australian region...” on lines 108-109 of the original manuscript to “ACCESS is a nested model: in this study I consider just the ACCESS-R component covering the Australian region…” on lines 114-115 of the revised manuscript.

***2. Lines 110-111: Although the ACCESS time steps may be in the order of 5 minutes, forecasters only get to see hourly data. It may pay to note this point.***

I agree, but believe I have done so adequately in the discussion on lines 132-137 of the original manuscript (lines 148-157 of the revised manuscript.)

***3. Line 128: I think upscaled should be downscaled***

Here I use “upscaled” in the sense of Ebert (2008) to refer to the regridding of high resolution model data, like ACCESS-C+ (<http://www.bom.gov.au/australia/charts/about/about_access.shtml>), onto a coarser grid. Given I hadn’t discussed ACCESS-C+ I agree this is confusing. I have therefore changed the phrase “...which then interpolates or coarsens the model data onto a standard…” on line 121 of the original manuscript, to “...which then downscales the model data, or in the case of high spatial resolution mesoscale model guidance, coarsens the model data, onto a standard…” (line 146-147 new manuscript).

***4. Line 131-132, last sentence: A reference describing the standard approach the BoM takes would be helpful.***

The text here should probably read “This is the standard approach the BoM takes when comparing the performance of the official forecast to unedited model guidance.” I have made this change, and provided a reference (lines 157-159 revised manuscript.)

***5. Pages 7 and 8: A diagram may help readers more easily understand how the perturbations are calculated.***

Figure 1 of the revised manuscript is now a schematic illustrating the essential ideas of perturbations, and how to calculate DAE.

***6. Line 147 and elsewhere: u appears to refer to the wind vector. But in meteorology, the standard usage of u is for the west-east component of the wind vector. Perhaps a different symbol could be used to avoid confusion?***

This was typeset as a bold italicised ***u*** in the original manuscript, but I agree it is hard to distinguish from the zonal component *u*. Following AMS guidelines (<https://www.ametsoc.org/ams/index.cfm/publications/authors/journal-and-bams-authors/formatting-and-manuscript-components/mathematical-formulas-units-and-time-and-date/>) I have changed the vector wind symbol to an upright bolded roman **u**, which is easier to distinguish from *u.*

***7. Line 151: "means" → "arithmetic means".***

This has been changed. See line 183-184 of the revised manuscript.

***8. Page 10 and elsewhere: A map with place names would be helpful. Remember that the audience of Weather and Forecasting is international, so international readers will be less familiar with Australian geography.***

I have reformatted the map in Fig. 1 of the original manuscript (Fig. 2 of revised manuscript) to include full names of Australian states, and location of Gulf of Carpentaria, as well as labels of the different city station groups.

***9. Multiple places in the results section: Rather than mention UTC after all dates and times, at the beginning of the section you could mention that all dates and times are in UTC.***

This has been done: see lines 291-292 of the revised manuscript.

***10. Figure 1:***

***1. No scale for the height colours***

***2. Inadequate place name labelling***

***3. State name abbreviations (WA, ACT etc) need defining.***

These issues have been addressed: see Fig. 2 of the revised manuscript.

***11. Figure 2: The right hand side panels are difficult to read. The black numbers are hidden by the dark blue colouring.***

Colormaps for all the “scorecard” type figures throughout the manuscript have been changed so that shading is not as dark.

***12. Various figures and other places in the text: Wind speed units of knots are used. Although knots are commonly used in aviation (where the ICAO abbreviation is kt, i.e. different from the ISO abbreviation of kn), perhaps these units should be converted to m/s?***

When I was Darwin, forecasters appeared to think and work in terms of knots, and (from memory) wind fields were given in GFE in terms of knots. Also, the Bureau’s internal “Jive” verification suite uses knots. I stuck with knots because I was motivated by increased accessibility of the results to forecasters, and because I’d seen knots used in other wind verification studies. However, the papers in Weather and Forecasting I’ve seen that have used knots (e.g. Tyner et al. 2015)

use the “kt” abbreviation, so I have changed “kn” to “kt” throughout the manuscript.

***13. Figures 6-8 and elsewhere: Same dark colour problem as in Figure 2.***

Changed: see response to minor comment 11.

***14. Figure 9 (and elsewhere): Where panels in a plot refer to different places, rather than labelling the panels a, b, c and d and then mentioning what these refer to in the caption, put a heading on each panel (e.g. Northern Territory, South Western Australia and so on).***

This has been done. I have also added headings and subheadings to many other figures, particularly the “scorecard” type figures, to better communicate the results and assist readability. (For some reason I thought such headings/subheadings were not permitted in AMS publications.)

***15. Figure 11: The colour scale here is not very helpful; most of the boxes are the same red colour.***

Shading has been removed from this figure.

### References:

Ebert, E.E. (2008), Fuzzy verification of high‐resolution gridded forecasts: a review and proposed framework. Met. Apps, **15**, 51-64. doi:[10.1002/met.25](https://doi.org/10.1002/met.25)

Rotunno, R., 1983: [On the Linear Theory of the Land and Sea Breeze.](https://journals.ametsoc.org/doi/abs/10.1175/1520-0469(1983)040<1999%3AOTLTOT>2.0.CO%3B2) *J. Atmos. Sci.,* **40**, 1999–2009, <https://doi.org/10.1175/1520-0469(1983)040<1999:OTLTOT>2.0.CO;2>

Tyner, B., A. Aiyyer, J. Blaes, and D.R. Hawkins, 2015: [An Examination of Wind Decay, Sustained Wind Speed Forecasts, and Gust Factors for Recent Tropical Cyclones in the Mid-Atlantic Region of the United States.](https://journals-ametsoc-org.ezp.lib.unimelb.edu.au/doi/abs/10.1175/WAF-D-13-00125.1) *Wea. Forecasting,* **30**, 153–176, <https://doi-org.ezp.lib.unimelb.edu.au/10.1175/WAF-D-13-00125.1>