# Reviewer 1:

Thank you etc.

***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?***

Thank you for your comment, and 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. This is referred to as a choice of *model guidance*.
2. Manually *edit* the model guidance dataset using the Graphic Forecast Editor (GFE) software package.

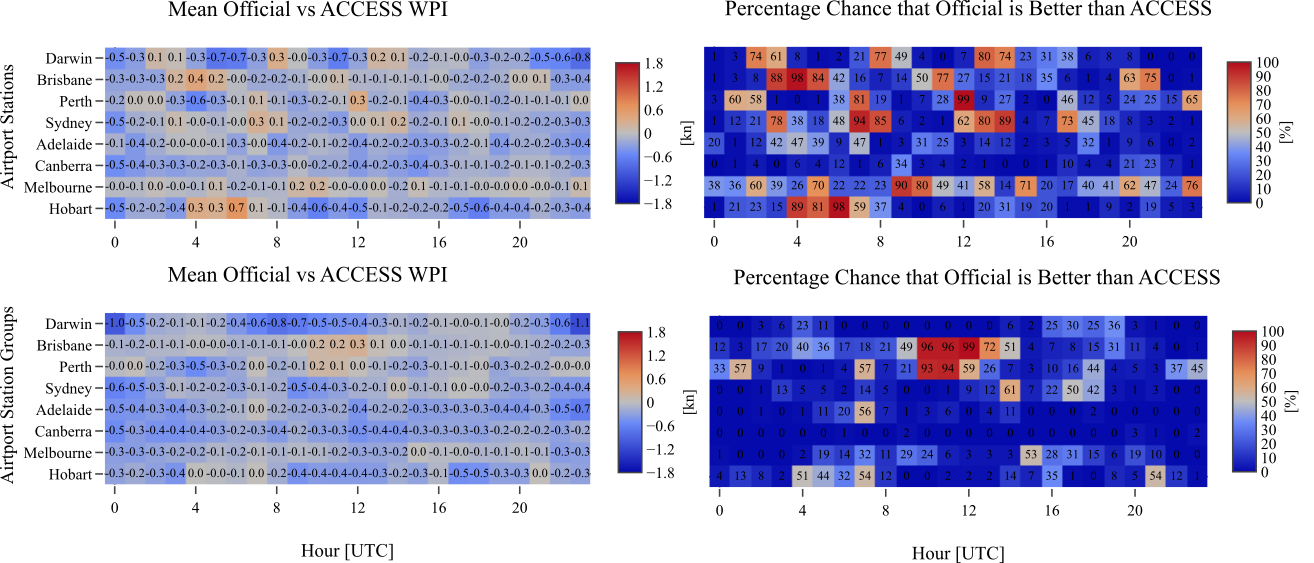
I view the “editing” step as distinct from the “choosing model guidance” step. Different forecasting centres across Australia have different practices for precisely how the forecasters choose model guidance and make edits, I don’t believe there are strict rule and regulations to guide this, just the managerial structures that exist within forecasting centres.

Data assimilation is a technique for ingesting observations into models during model initialisation. Operational forecasters in Australia do not run models themselves, and they therefore have limited input into decisions made by modellers, such as method of data assimilation, choice of parametrisation schemes and so forth.

I discussed these issues in the abstract, and throughout lines 31-41, and lines 51-66 of the original manuscript. I have edited lines 9-11 of the abstract to better distinguish between “choice of model guidance” and “edits”, also rewording the final sentence of the abstract to stay under the word limit. 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 “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 new introduction, to try and distinguish the types of decisions forecasters make, from modelling decisions such as 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 manuscript, but for the Official versus ACCESS comparison. The results are similar to Fig. 6 in that there are only a few times and locations where Official unambiguously outperforms ACCESS. At the city station group scale, and results are noisy for the individual airport stations. I noted this on lines 299-304 of the original manuscript (and lines 335-336 of the revised manuscript.) I consider it unnecessary to include these figures, as space requirements mean I can only present a subset of the results, and I think the more interesting comparison is why HRES, which unlike ACCESS is provided in three hourly time steps and not calibrated to Australian conditions, is still able to produce a better mean diurnal cycle (in the sense of the metrics defined in the paper) than the official forecast.

**

**Figure 1:** Analogous to Fig.~

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

Remember that 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, 5-95%, 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 197 to 214 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 consider it 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. However, I agree that probabilistic forecasts are desirable.

Note also 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 perturbations vectors), then considering means of this quantity. This is somewhat more sophisticated than a simple calculation of absolute error. There are innumerable ways I could have made the metric more complex, for instance by applying the Huber loss function to both terms in equation (1) before taking a difference, or to the overall expression, or to both the constituent terms and the overall expression. Such options grow extremely 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 forecasting practice beyond what is present in the conclusion (which has been revised slightly in the new manuscript.)

***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 125-142 of the new manuscript, with OCF results now present in most of the figures, with discussion of these results throughout the revised manuscript.

# 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 abscissa, 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).

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 abscissa. 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:

***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 – see below for details.

***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 very valid point. I did consider this issue, but reasoned that if I enforced constant areas I’d have inconsistent numbers of stations in each group. Such inconsistencies are an inherent limitation of station based verification, as station density and placement vary. Given the significant random turbulence at the station scale, I reasoned greater consistency could be achieved by enforcing equal numbers of stations rather than equal areas.

Regardless, presenting results at different locations is intended to elucidate physical differences in the mean diurnal cycle at different locations, and their effects on verification statistics, not to provide an objective “ranking” of forecasting centres in different states, or something similar. The mean diurnal cycle varies in amplitude significantly across different locations, and forecasters will therefore have different capacities to account for it due to these physical differences alone. Any attempt to objectively “rank” performance at different locations would also be misleading for these physical reasons. I have outlined these issues in a new sentence “ “ at line XX (line XX of the revised manuscript).

***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 certainly found it challenging to find the appropriate balance between 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 hope the revised manuscript makes my methods and results clearer, particularly the improved lay out and labelling of figures. My code is freely available online, and although I cannot circulate the Jive code it uses without the Bureau’s permission, I could add my code to the Bureau verification team’s server, where similar code is housed. This server is accessible to all Bureau employees, who would be able to run my code, and apply the methods to different stations, station groups or time periods.

***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 very sorry the manuscript was 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 simplified the way I use acronyms, and simplified the terminology I use. I have added new paragraphs to the methods section to better explain the appropriate way to view wind perturbations, and crucially, means of these perturbations.

***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 the figures were difficult to understand. I have amended all the figures, except Fig. 5 (Fig. 6 revised manuscript). 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 very 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 ECMWF 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 imply this is still true, or true across Australia, an error I should not have made.

I have therefore revised the manuscript substantially 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 , referencing 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 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 these things 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. These results are discussed in the revised manuscript, particularly on lines 396 to 405. However, on lines 403 to 405 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 140 to 142 of the revised manuscript I take care to point out that OCF produces lower errors in both wind speed and direction than any of the individual model datasets that comprise it.

I was aware of this diurnal 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 forecasts 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.

An obvious hurdle 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 kept: perhaps this could be done using GFE somehow.

***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 very valid point that I should have addressed directly in the original manuscript. The key idea is *not* to use metrics like DAE to verify performance on individual days (and I acknowledge the case study’s 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 mean differences in error between the underlying dataset.

To make this concrete, it is instructive to consider synthetic data. Consider data with an hourly time domain

***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?***

***- 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 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.***

***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.***

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

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

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

***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?***

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

***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.***

***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.***

***10. Figure 1:***

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

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

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

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

***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?***

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

***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).***

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