# Responses to Reviewer #2

## Point #1

Comments on â??Bias-correcting individual inputs prior to combined calibration leads to more skillful forecasts of reference crop evapotranspirationâ?? by Yang et al. This study evaluated two calibration strategies for simulating reference crop evapotranspiration. The two strategies are (1) calibration directly applied to raw ETo forecast constructed with raw forecast of input variables; (2) bias-correcting input variables. The bias-correcting algorithm has been proved to be more feasible. Although this study is of significance, improvements and revision can make the study stronger and more compelling.

**Response: We appreciate the reviewer's insightful suggestions and comments on the manuscript. We address concerns from the reviewer carefully and improve the manuscript accordingly.**

## Point #2

Core of my concerns is the results presentation and discussion, many sections are superficial; the results are simply described, more insightful explanation and discussion are needed. See below for my suggestion. A moderate revision can easily address these comments. So I suggest a moderate revision.

**Response: We appreciate the reviewer's constructive comments. We improved the analysis and presentations by (1) creating boxplots to summarize results plotted as maps to better demonstrate results quantitatively, (2) performing statistical analyses (t-test) when comparing results from different model runs, (3) providing more statistical information in the Results section, and (4) Comparing findings of this work with published investigations. We further explain these improvements in detail as follows:**

**(1) Adding boxplots to Results**

**We created boxplots for results shown as maps (Figures 1 to 8 in the main text). We combine these boxplots with maps for Figures 2-7, which have extra zoom for adding new subplots. For Figures 1 and 7, which already include many subplots, we present the corresponding boxplots in the Supplementary Material. We also update the main text accordingly. Please find the boxplots as follows:**

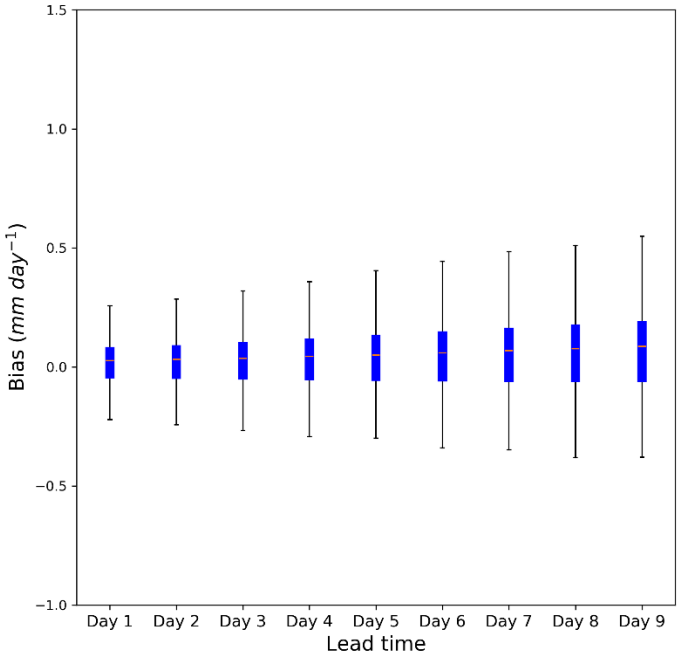



**Figure 2 Boxplot summarizing bias in calibrated ETo forecasts**

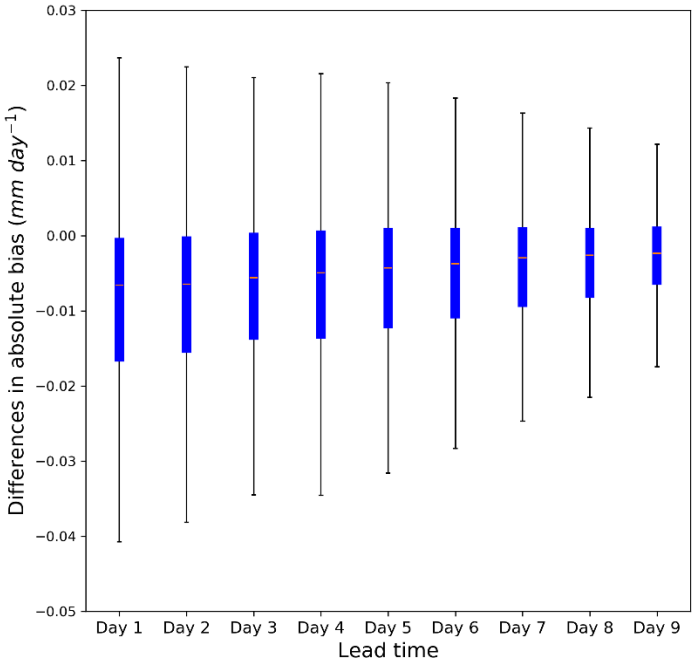
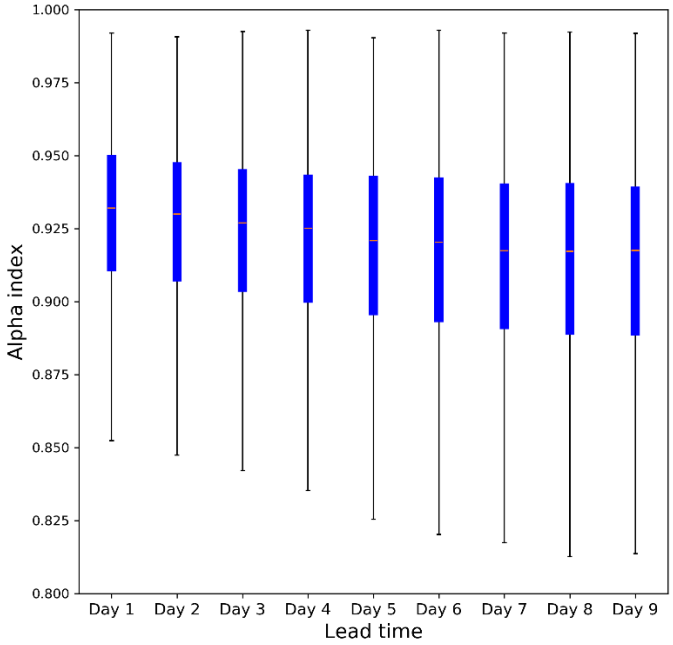
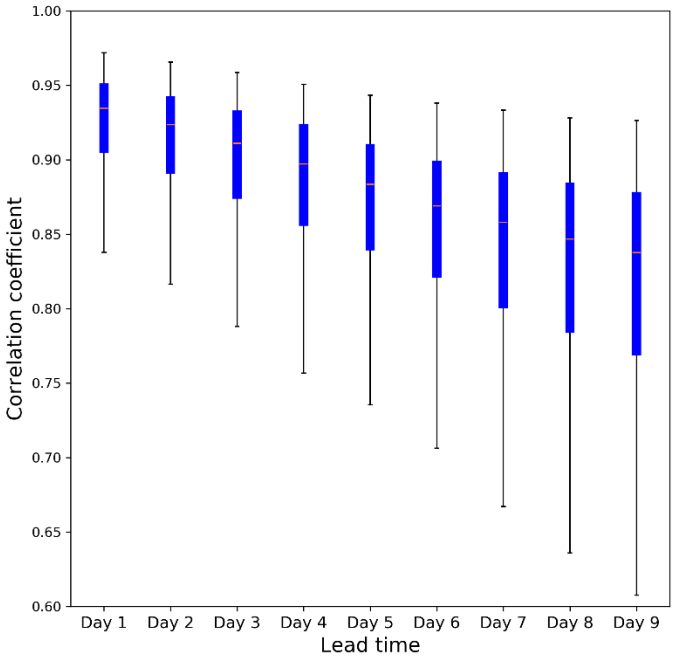



**Figure 3 Boxplot summarizing differences in absolute bias between calibrated ETo forecasts from Calibration 2 with Calibration 1**




41

42 **Figure 4 Boxplot summarizing the alpha index in the calibrated ETo forecasts**




43

44 **Figure 5 Boxplot summarizing correlation coefficient between calibrated ETo forecasts from**
45 **Calibration 2 and AWAP ETo data**

46

47

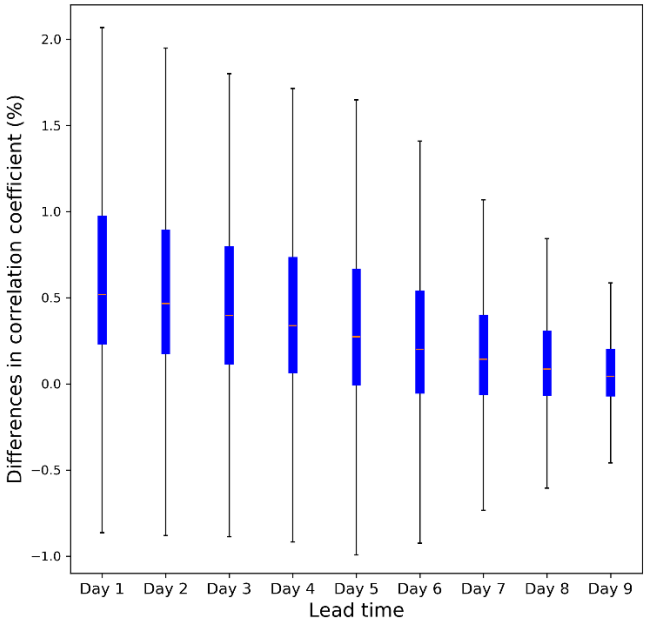



**Figure 6 Boxplot summarizing differences in the correlation coefficient (calibrated forecasts vs. AWAP ETo) between Calibrations 2 and 1**

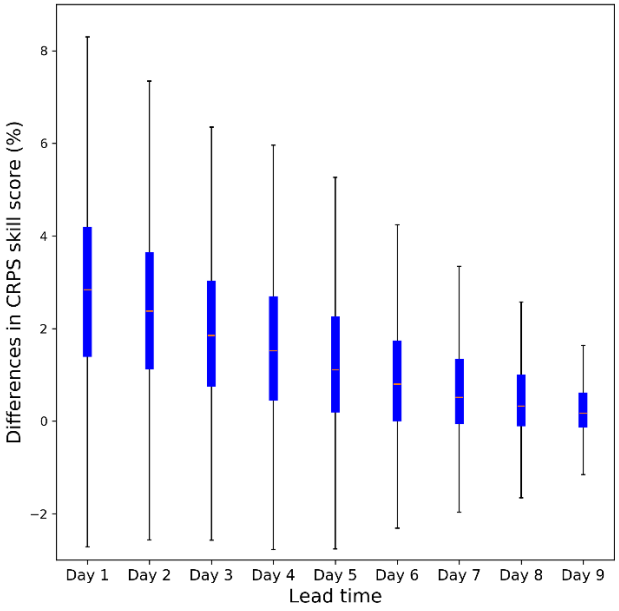



**Figure 8 Boxplot summarizing differences in CRPS skill scores between the calibrated forecast from Calibration 2 with those from Calibration 1**

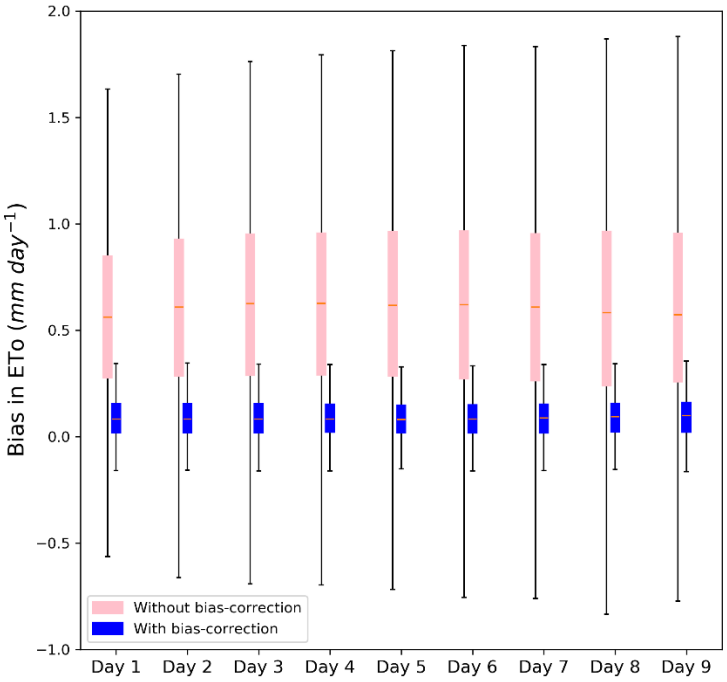



*Figure S12. Boxplot of biases in raw ETo forecasts constructed without bias-corrected input variables (pink) and correct inputs (blue)*

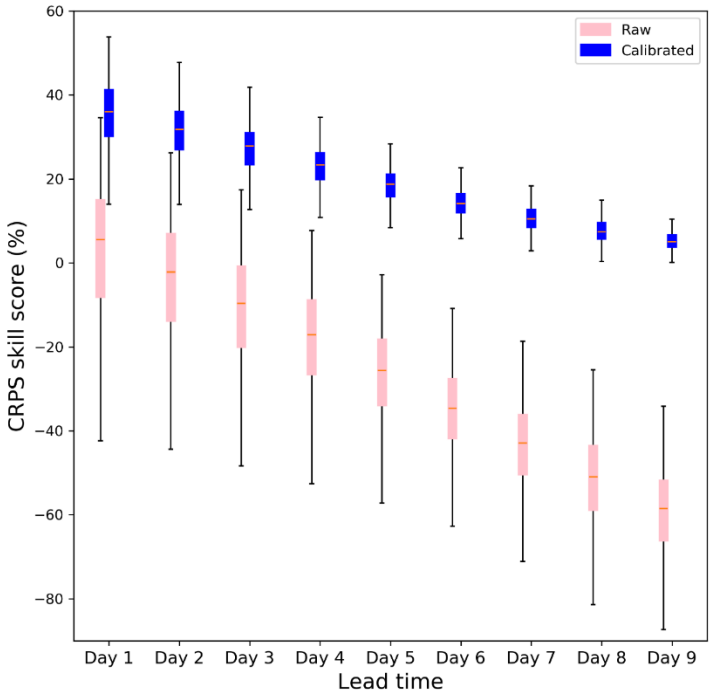



*Figure S14. Boxplot of CRPS skill score in raw (pink) and calibrated ETo forecasts (blue) from Calibration 2*

**（2）Conducting t-test to compare results from different Calibrations.**

**We conduct t-tests to further evaluate the performance of the two calibration strategies. Specifically, T-tests were conducted in the evaluation of bias, correlation coefficient, and CRPS skill score (figures 1, 2, 3, 6, 7, 8) of the raw or calibrated forecasts (Table S2). In addition, we also conducted t-tests (Table S1) to evaluate raw forecasts of the five input variables (Figures S2 to S6).**

**In the calculation of *t* statistics, we used the Spatial Degrees of Freedom (SDOF), rather than using the total grid cells in the study area, to account for the spatial correlation in the t-test. The SDOF is substantially smaller than total grid cells** (Toth, 1995)**.** Wang and Shen (1999) **investigated SDOF of GCM outputs and reported a range of 90-120, out of 738 grid cells for the southern hemisphere. In this study, we use 50 as the SDOF for our t-tests. Considering the large amount of total grid cells (281,622) in this study, we believe that 50 is a conservative estimate of SDOF for this investigation. We calculated the *t-statistics* and evaluate whether they are statistically significant using the SDOF of 50. Results of the t-tests (Tables S1 and S2) are added to the supplementary material.**

**Reference：**

Toth, Z.: Degrees of freedom in Northern Hemisphere circulation data, Tellus, Ser. A, 47 A(4), 457–472, doi:10.3402/tellusa.v47i4.11531, 1995.

Wang, X. and Shen, S. S.: Estimation of spatial degrees of freedom of a climate field, J. Clim., 12(5 I), 1280–1291, doi:10.1175/1520-0442(1999)012<1280:EOSDOF>2.0.CO;2, 1999.

**Table S1 Results of t-tests (*t-statistic*) for raw forecasts of input variables**

| Tests / Lead times | Test if bias in raw Tmax forecasts is different from zero (Figure S2) | Test if bias in raw Tmin forecasts is different from zero (Figure S3) | Test if bias in raw vapor pressure forecasts is different from zero (Figure S4) | Test if bias in raw solar radiation forecasts is different from zero (Figure S5) | Test if bias in raw wind speed forecasts is different from zero (Figure S6) |
|---|---|---|---|---|---|
| Day 1 | -8.96** | 1.66 | -3.18** | 11.83** | 16.04** |
| Day 2 | -8.16** | 2.65** | -3.43** | 11.39** | 16.50** |
| Day 3 | -8.19** | 2.68** | -3.77** | 11.81** | 16.57** |
| Day 4 | -8.12** | 2.56** | -4.05** | 12.17** | 16.56** |
| Day 5 | -7.87** | 2.41** | -4.09** | 12.45** | 16.45** |
| Day 6 | -7.70** | 2.27** | -4.21** | 11.88** | 16.45** |
| Day 7 | -7.73** | 2.22** | -4.33** | 10.81** | 16.29** |
| Day 8 | -7.70** | 2.17** | -4.30** | 11.41** | 16.56** |
| Day 9 | -7.44** | 2.20** | -4.18** | 11.95** | 16.82** |

**The Spatial Degrees of Freedom (SDOF) is 50 in the tests; ** indicates statistically significant differences at the 95% confidence interval.**

**Table S2 Results of t-tests (*t-statistic*) for performance evaluation**

| Tests / Lead times | Comparison of bias in raw ETo forecasts constructed with vs. without bias correction (Figure 1) | Test if bias in calibrated ETo forecasts from Calibration 2 (Figure 2) is different from zero | Test differences in absolute bias between calibrated ETo forecasts from Calibrations 2 and 1 (Figure 3) | Test difference in *r* between observations and calibrated ETo forecasts from Calibrations 2 and 1 (Figure 6) | Comparison of CRPS skill score between raw and calibrated ETo forecasts (Figure 7) | Test difference in CRPS skill score of calibrated ETo forecasts from Calibrations 2 and 1 (Figure 8) | Test difference in α-index between Calibrations 2 and 1 (Figure S8) | Test difference in CRPS skill scores between Calibrations 3 and 4 (Figure S10) |
|---|---|---|---|---|---|---|---|---|
| Day 1 | -9.76** | 1.80 | -4.08** | 5.73** | 27.59** | 11.53** | -0.54 | 11.53** |
| Day 2 | -9.86** | 1.91 | -3.93** | 4.93** | 29.03** | 10.86** | -1.47 | 10.86** |
| Day 3 | -9.86** | 2.07** | -3.68** | 4.43** | 31.14** | 9.77** | -1.81 | 9.77** |
| Day 4 | -9.81** | 2.27** | -3.54** | 4.01** | 33.77** | 8.58** | -1.17 | 8.58** |
| Day 5 | -9.71** | 2.40** | -3.36** | 3.75** | 38.11** | 7.16** | -2.09** | 7.16** |
| Day 6 | -9.54** | 2.60** | -3.37** | 3.17** | 42.59** | 6.44** | -1.28 | 6.44** |
| Day 7 | -9.34** | 2.76** | -3.26** | 2.69** | 44.38** | 6.15** | -1.99 | 6.15** |
| Day 8 | -9.04** | 2.98** | -3.13** | 2.32** | 45.57** | 5.85** | -1.57 | 5.85** |
| Day 9 | -9.21** | 3.13** | -2.91** | 1.85 | 51.91** | 5.05** | -1.70 | 5.05** |

**The Spatial Degrees of Freedom (SDOF) is 50 in the tests; ** indicates statistically significant differences at the 95% confidence interval.**

**(3) Improving the Results section**

**We add more specific information in describing the key findings of this study and introduce the results of the statistical analyses (Tables S1 and S2). Since we modified many sentences, we decide not to list them here. Please see details in the revised manuscript.**

**(4) Improving the Discussion section**

**We further compare the findings of this investigation with related studies in discussion:**

“In the ETo forecasting across 40 Australia, although raw ETo forecasts constructed with NWP outputs reasonably captured the magnitude and variability of ETo, forecast skills better than climatology were only found for the first 6 lead times (Perera et al., 2014). Our investigation suggests that statistical calibration could substantially improve forecast skills and outperform the climatology forecasts for all 9 lead times across Australia. The findings of this investigation agree well with the site scale short-term ETo forecasting based on GCM outputs (Zhao et al., 2019a) in improving forecast skills. Calibrated forecasts from Calibration 2 demonstrate similar skills as those of Zhao et al. (2019a). However, our calibration achieves the improvements using much shorter archived raw forecasts (3-year vs. 23-year) than Zhao et al. (2019a), thanks to the capability of SCC in calibrating short-archived forecasts (Wang et al., 2019). Calibrated forecasts from Calibration 2 also demonstrate comparable biases (0.32-0.95%) with calibrated ETo forecasts (0.49-0.63%) in the U.S. based on the Bayesian model averaging (BMA) model and weather forecasts from three NWP models during 2014-2016.”

## Point #3

Lines 11, fully implemented.

**Response: we change it to 'fully implemented '.**

## Point #4

Line 27, â□□divergentâ□□ emphasizes completely different assumption, you can just use replace it different to ensure a general term.

**Response: We replace the word ‘divergent’ with 'different'.**

## Point #5

Line 38, physical processes of the atmosphere, it is unclear, atmospheric circulation or atmospheric wind formation, or physical processes in the atmosphere

**Response: Thank you for the suggestion. We change the sentence as follows:**

"ETo is affected jointly by temperature, vapor pressure, solar radiation, and wind speed (Bachour et al., 2016; Luo et al., 2014). Prediction models using these weather variables as inputs allow for

representations of atmospheric dynamics and often produce reasonable ETo predictions (Torres et al., 2011)."

## Point #6

Section 3.1, 3.2, the authors described the results in the figures. However, most of those text are vague, please provide more specific (quantitative) information to support your statement. When you compare different results or method, it is better to report some statistic results (p value, r2, etc).

**Response: We conduct statistical analysis to quantify the difference between different model runs, and update the Results sections accordingly. Details of the t-tests could be found in our response to your comments point #1.**

## Point #7

for example, line line 223-225, you report the overprediction in Tmax, and underpredict in Tmin in different regions. If it is underprediction, what is the range of that underprediction, same for overprediction, are these different statistically significant? There are many similar issues in other sections.

**Response: We appreciate the reviewer's valuable suggestions. The reason we did not introduce errors in raw forecasts of the input variables in detail is that systematic errors in raw NWP forecasts have been well documented. Evaluation of the raw forecasts of the inputs is not the key information we want to deliver in this study. However, we agree with the reviewer that more statistical information is needed. We conduct statistical analysis to quantify errors in raw forecasts (Table S1), and update contents in Results accordingly. Statistical analyses could be found in our response to your comment #1. Here is the updated description of errors in raw forecasts:**

**"**Raw forecasts of the five input variables demonstrate significant inconsistencies with the corresponding AWAP data (Figures S2-S6). In most parts of Australia, daily maximum temperature (Tmax) forecasts are lower than AWAP data by 1-2 °C. Overpredictions in Tmax are only found in coastal areas of northwestern Australia. The daily minimum temperature (Tmin) is underpredicted by more than 1.5 °C in western and central parts of Australia by the raw forecasts, but is overpredicted by ca. 1 °C in eastern and southern Australia. Forecasted wind speed is higher than the reference data by more than 1m/s (or by 50%) in most parts of Australia. Similarly, raw solar radiation forecasts are about 5% higher than AWAP data across Australia. Vapor pressure is underpredicted in western and central regions by ca.14%, but is overpredicted by ca. 6% in coastal areas of south-eastern Australia by the raw forecasts. For each of the five variables, spatial patterns of biases in raw forecasts are consistent across the 9 lead times, demonstrating systematic errors in the raw NWP forecasts. According to our statistical test, overpredictions or unerpredictions in raw forecasts of the input variables are statistically significant *(P*<0.05) for most lead times (Table S1). **"**

## Point #8

In the discussion section, I would be willing to see a comparison with other studies with different algorithms for the ETo simulation. Some quantitative comparison to elucidate the better performance of the new bias-correction algorithm needs to be done. I believe it will prove the reliability of the new algorithm.

**Response: We appreciate the constructive comments. This is the first continental-scale ETo forecasting in Australia. Previous NWP/GCM-based ETo forecasting in Australia is conducted at the site scale. As a result, in the original manuscript, our evaluation was primarily focused on the comparison against observations. In this area of weather/climate forecasting, different calibration models, based on different statistical theories, have been developed and implemented. Previous comparisons suggest that the performance of these models varied with study areas, NWP models, and choice of evaluation metrics (Wilks, 2018), and there is no conclusion regarding which group of post-processing models has the best performance.**

**More importantly, rather than developing a new calibration model, this investigation is to evaluate the necessity of including an extra step before forecasts are calibrated. As we introduced in the maint ext, the objective of our investigations is to address a common challenge faced by NWP-based ETo forecasting, and we expect the calibration strategy developed in this study will benefit ETo forecast calibrations, no matter which statistical model is employed.**

**However, we agree with the reviewer that comparison of model performance with other models will help readers better understand the reliability of this work. We review previous studies and add the following content to Discussion (section 4.1):**

"In the ETo forecasting across 40 Australia, although raw ETo forecasts constructed with NWP outputs reasonably captured the magnitude and variability of ETo, forecast skills better than climatology were only found for the first 6 lead times (Perera et al., 2014). Our investigation suggests that statistical calibration could substantially improve forecast skills and outperform the climatology forecasts for all 9 lead times across Australia. The findings of this investigation agree well with the site scale short-term ETo forecasting based on GCM outputs (Zhao et al., 2019a) in improving forecast skills. Calibrated forecasts from Calibration 2 demonstrate similar skills as those of Zhao et al. (2019a). However, our calibration achieves the improvements using much shorter archived raw forecasts (3-year vs. 23-year) than Zhao et al. (2019a), thanks to the capability of SCC in calibrating short-archived forecasts (Wang et al., 2019). Calibrated forecasts from Calibration 2 also demonstrate comparable biases (0.32-0.95%) with calibrated ETo forecasts (0.49-0.63%) in the U.S. based on the Bayesian model averaging (BMA) model and weather forecasts from three NWP models during 2014-2016."

**In addition, we also highlight the importance of testing the proposed calibration strategy (strategy ii) based on other calibration models in the future in section 4.2:**

"Second, further investigations based on other calibration models are needed to validate the conclusions of this investigation. Our analyses based on two different methods (based on ETo anomalies vs. based on original ETo) find similar improvements in calibrated ETo forecasts following bias-correction of input

variables. Additional tests using other calibration models will be needed to evaluate whether the improvements will be achieved when the calibration is conducted with a different model."

**Reference:**

Wilks, D.S., 2018. Chapter 3. Univariate Ensemble Forecasting, in: Vannitsem, S., Wilks, D.S., Messner, J.W. (Eds.), Statistical Postprocessing of Ensemble Forecasts. pp. 49–89. https://doi.org/https://doi.org/10.1016/C2016-0-03244-8

## Point #9

Line 388, feasible or reliable ETo forecasting.

**Response: This paragraph has been rewritten. Please see the revised contents in our response to your comment #10.**

## Point #10

Line 390, short-term ETo forecasting provides highly valuable information for real-time decision making on water resource management and planning farming practices. This study proved the bias-correction approach is a feasible method for a more robust calibration of the NWP-based ETo forecasting.

**Response: We appreciate the reviewer's valuable suggestions. We remove redundant sentences and combine the last two paragraphs in the Conclusion section:**

**"**This investigation clearly suggests the necessity of improving input variables as part of the NWP-based ETo forecasting. With this extra step, the bias, correlation coefficient, and skills in the calibrated ETo forecasts are all improved, particularly for the short-lead-time forecasts. Further investigation indicates that the improvements tend to be independent of the calibration method applied to raw ETo forecasts. Forecasting the highly variable ETo is often challenging. Our investigation provides an effective calibration strategy for improving NWP-based ETo forecasting. As a result, we anticipate that future calibration of NWP-based ETo forecasts could benefit from adopting this strategy to produce skillful calibrated ETo forecasts. This strategy is also expected to be applicable to enhancing the forecasting of other integrated variables that are calculated using multiple NWP/GCM variables.**"**