Authors' response to Anonymous Referee #2 interactive comment on "Daily temperature records from a mesonet in the foothills of the Canadian Rocky Mountains, 2005–2010" by Wendy H. Wood et al.

Author comments are italicized and prefixed with AC:

This paper needs a good comma clean-out!

AC: We scrutinized our commas during the revisions and hope to have addressed some of the concerns here. Part of this may be U.S. vs. U.K. English. We trust that the Editorial office will keep us on track for any remaining troublesome commas.

Pg. 1 line 23 change "and" to "or"

AC: replaced as suggested

Pg. 1 line 24 change "influencs that regulates" to regulators

AC: replaced as suggested

pg. 2, line 4: Why do the temperature "need" to be distributed? Also consider expanding the description of weather and climate models. Some nested models can simulate at less than a km. I think you are intending to implicate GCM's here?

AC: revised, p.2, ll.3-6 – now clarified to reflect that we were thinking of GCMs here; references added for much-cited examples from CMIP5, ECMWF reanalyses, and a recent paper from the 'high-resolution' GFDL model. These GCMs are operating at resolutions of 0.5 to 2 degrees (~200 km).

pg. 2, lines 8-9 The network isn't really two dimensional given that the observations follow the topography. Because the next sentence describes the station distribution simply strike "two dimensional"

AC: removed as suggested

pg 2, line 9: indicate that these are near-surface observations or, if you are feeling bold, say they are 2- m air temperatures and humidity. For precipitation this is less important.

AC: Added as suggested. We prefer the use of near-surface as sensors were setup at different heights as indicated in section 2 (p.3, ll.18-20).

Pg 2, line 14: "spatial patterns of temperature structure" sounds redundant unless you indicate that the structure of interest is vertical lapse rates in land surface temperature. This is made clearer in the following sentence, but should be indicated here.
AC: Removed "structure" as the intent with the overall study was to examine temperature in general and lapse rate was one application

Pg 2, line 18: strike "backcountry" unless you want to indicate some state of environmental disturbance.
AC: Replaced with "remote" (p.2 l.19).

Pg 2, line 29: decapitalize “City”
AC: Done as suggested (p.2 l.30).

Pg. 3, line 6-7: How was the terrain variability determined? Was the mountain/prairie classification done manually or through some numeric means?
AC: Classification was done using a formula based on absolute elevation and elevation variability within a 250m buffer of the site.

Pg. 3, line 30 make read “...altitudes up to 3500 m, and straddle the border between British Columbia and Alberta in this region with a northwest to southeast alignment.” Or, keep boundary if you are referring to the climatological separation of prairie climates and those further west.
AC: Altered as suggested (p.2 l.31).

Pg. 3 first full paragraph. Quantifying the various landcover types by area in this region might be useful, or refer to your table to give the reader a sense of land cover predominance and how the stations are situated within.
AC: Text altered to reference Table 1 (p.3 l.4).

Pg. 3 line 24: strike “fabulously”
AC: Removed.

Pg. 3 third full paragraph: In a region dominated by rectilinear property lines, transportation infrastructure and whatever else a true grid might result in a systematic bias in the sampling. Furthermore, the Rocky Mountains have a strong linear structure in some locations. Although these aren't orthogonal or parallel to the grid, this could introduce an interaction between the sampling and the topography if a true 10 km x 15 km grid were established. It would be very difficult to show, but the less regular southern portion of the array might be the better network. Secondly, a grid isn't a random sampling. A true random sampling would involve sampling from the range of lat and lon in your region and placing instruments at those sample locations. Obviously, this isn't actionable or even mathematically showable, but it's a concern in designing a network such as this one. A grid would be okay id the landscape being sampled were random, but it's not in your region.

AC: We have replaced “random” with representative and added a reference where it was determined that the sites were statistically representative of the overall terrain characteristics of the area (p.3 ll.27-29).

Same paragraph: was any effort made to distribute the stations in the vertical in locales with steep topography? i.e. every couple hundred metres or so?
AC: This was not done as the purpose of this study was to identify regional patterns rather than local vertical structure. We do have other studies where we have looked in more detail at the vertical temperature structure within a given valley/slope, but these will be reported elsewhere as the objectives and datasets differ.

Pg. 3, line 32: Clarify "shootings". Were sensors shot at? Were participants shot at? This is alarming and maybe out of place in the paper.
AC: Replaced with "vandalism". It did appear as if some sensors had been used for target practice (p.4 l.3).

Pg. 4, line 4: can you indicate in Fig. 1 which transect was retained?
AC: Added text to the paragraph indicating the line retained (p.4 l.7).

Pg. 4, first para in sec 3.1: If I understand correctly, the sensors recorded instantaneous temperature at the top of the hour in the field while the calibration was done against hourly aggregates. This is a potentially significant problem. The hourly averaging will dampen the magnitude of solar influence on observations by mixing observations taken under calm spells with those taken under breezier periods. A single hourly observation has a luck-of-the-draw chance of being taken either during calm or breezy, cloudy or clear conditions. It's hard to say how big of a difference this distinction could make without calibrating specifically for it. Because the calibration differs from the observational approach, this weakens the utility of the calibration.
AC: Correct – this is a good point, thank you. Paragraphs 4-6 in section 3.1 discuss the solar warming effect under calm conditions, but without additional wind and solar measurements in the field, this cannot be corrected for. We may have understated the maximum effect on daily maximum temperatures. We have added a paragraph (p.5, l.4) describing results based on instantaneous hourly measurements, which are directly comparable with the field implementation and provide a better estimate of the uncertainty in the field data. The average daily mean difference using hourly measurements was -0.2°C, compared with -0.1°C using aggregated hourly measurements.

Pg. 5, lines 6-7 make read "...are the greatest due to solar heading, and can be problematic..."
AC: Modified as suggested (p.5 l.17).

Pg. 5, line 8: I don't think the Georges and Kaser reference is the right one to use here – there must be other references or documents that have documented temperature offsets in unventilated shields in more typical, non-cryospheric conditions. Their study is over a tropical glacier which would have very extreme solar radiation conditions. For example, direct sensor heating is difficult to imagine in a low albedo condition like a boreal forest or even open grassland. During winter in the Rocky Mountain foothills where albedo might be high, the solar insolation will be low and for a short duration due to short days so, again, not likely comparable to a tropical glacier.
AC: Reference replaced with Huwald, H., Higgins, C. W., Boldi, M., Bou-Zeid, E., Lehning, M. and Parlange, M. B. 2009. Albedo effect on radiative errors in air temperature measurements. *Water Resources Research (p.5 l.16).*

Pg. 5, lines 13-15: The averaging is also described earlier in the section. Combine those two efforts.
AC: Some text removed in this paragraph (p.5 ll.15-21).

Pg. 5 on calibration: Was overcooling investigated? This occurs when an instrument, again unventilated, is exposed to conditions of low incoming longwave radiation. This can lead to cooling of the sensor of the sensor + radiation shield and underestimation of temperature. Ventilation prevents this. The negative bias early in the study period shown in table 2 suggests this, but the later calibrations argue against this. A mention of this possibility is still warranted.
AC:Thank you for suggesting this. We did not investigate this, but have added a sentence indicating this could be a possibility (p.4 ll. 8-10).

Pg. 6, line 32: What does "queries were used to delete" mean? Is this a true database query or something else? The non SQL informed reader would wonder how asking or querying the data would result in a deletion...
AC: Rephrased sentence and removed "queries" (p.7 ll.8-9).

Pg. 7, line 21 strike "real". I understand the intent here but, a real error sounds strange. Could say "...one or more hours are typically errors."
AC: Replaced with "actual" (p.7 l.25).

Pg. 7, line 27: Can you indicate what the magnitude of the adjustments to the Calgary extreme values was? Given the large elevation range, the adjustment may need to have been large
AC: The extreme value test has been removed. On review of the qc tests we found that the spike test was identifying all extreme value violations.

Pg. 7, line 32: insert "temperature" after diurnal.
AC: Added as suggested (p.8 l.5).

Pg. 8, line 1: strike "in general" here as I'm sure snow burial wasn't a general issue to the network. Further here, how were the blocks of flagged data defined when the sensor was found buried? Is this different somehow or did it yield the same results as detecting burial in the data alone?
AC: Text removed as suggested. Blocks of data were identified during the visual review with neighbouring sensor data (p.8 ll. 4-8).

Pg. 8, line 13: make clear if boundary sites are those at the edges of the network. Or are they boundary to some other feature?
AC: Text modified to read "sites at the edge of the survey" (p.8 ll.19-20).

Pg. 8, line 20: I'm curious about the rationale for removing the extreme from a given day's grouping of observations in a neighbourhood. Given the relatively small number of observations that a group of ~10 stations gives, subtracting two will have an effect on the standard deviation calculated, supressing it, and thereby making rejection of potentially good data more likely? Were any sensitivity tests done on this (and other) QC procedures

to see what the effect of restricting the data in this way might have been? I could also invision a case where a grouping of stations happens to be in conecert by chance thus making the standard deviation very small and the resulting 5 sigma threshold for rejection a fairly mundane value.
AC: Sensitivity tests were done to determine an appropriate method. Removing the extremes does reduce the standard deviation and increase the likelihood of removal, but this seemed like a more conservative choice. Keeping the extremes increases the standard deviation and reduces the likelihood of removal, making the inclusion of erroneous data more likely. This was a subjective test where records flagged were also visually reviewed, and we looked at 3, 4 or 5 sigma for this – based on visual checks for 'reasonableness' of data, with and without retention of extreme values. We chose the method that performed best.

Pg. 9: It might be valuable to indicate the failure rates for the various tests that you applied to indicate which tests held the greatest utility. I'd see this as an optional addition to the paper.
AC: Commonly, when sensors malfunctioned, it was for an extended period of time and the bad data was readily identifiable in field checks, therefore this manual check had the highest exclusion rate. Of the automatic tests, the neighbourhood consistency check was the most effective at identifying shorter periods of bad data.

Pg. 9, line 15: Instead of "poor performance" maybe more specifically say "spatial or temporal discontinuities in interpolated of modelled temperature surfaces"
AC: Text has been modified as suggested (p.9 ll.22-23).

Pg. 10, line 7ish: give a list of the variables used in the DFA if it's relatively brief. Or, be more explicit in your reference to Table 4 so it's clear that the variables may be found therein "The weather variables used in the analysis are listed in Table 4".
AC: There was an extensive testing process to select the variables, which needs a more detailed description (Wood, 2017) to give an adequate treatment. Table 4 does not fully describe the weather variables, but more indicates overall conditions for each weather type. We added a short sentence on this and reference to Wood (2017), p.10, ll.10-11.

Pg 10, first paragraph in sec 4.1: The last three sentences here contain slightly redundant information. Is it not true that the utilization of anomalies precludes a need to deseasonalize the data?
AC: The paragraph has been rewritten. We used anomalies as the method to deseasonalize the data (p.10 ll.11-14).

Pg 10, lines 10-15: The manual classification leaves the question regarding discussion of the DFA and application: I'm having a bit of a hard time following how the groups based on one year of Calgary Airport data are then extended back in time. From there I'm having a hard time following how that was used to select station pairs for subsequent gap filling. There need to be a couple of paragraphs explaining this method after line 15 on this page.
AC: Again this is described in detail in Wood (2017). The period used for creating the discriminant functions does not overlap with the FCA data collection period. However,

we had additional data for this period which we believe allowed us to do a better manual classification. The manual classification period and the FCA period are only a few years apart, so average conditions are not expected to have changed much and the discriminant functions should still perform well. The DFA technique was really just about characterizing what a given weather type looks like in our region, based on daily weather conditions. There is an implicit assumption that a given weather type, e.g. a chinook or a cold, dry (cP) air mass, will have similar weather anomalies from one year to the next. The DFA classification can then be applied across time, as long as mean conditions do not change much. There is a second implicit assumption that the dominant weather-system types all occurred at least once in the calibration period (2013-2014), as per our seed groups. Exotic weather types during the FCA period will not be captured through this approach. but our interest here is the most common regional weather systems. We have added a brief discussion of how DFA is applied in our context, p.10, ll.16-28.

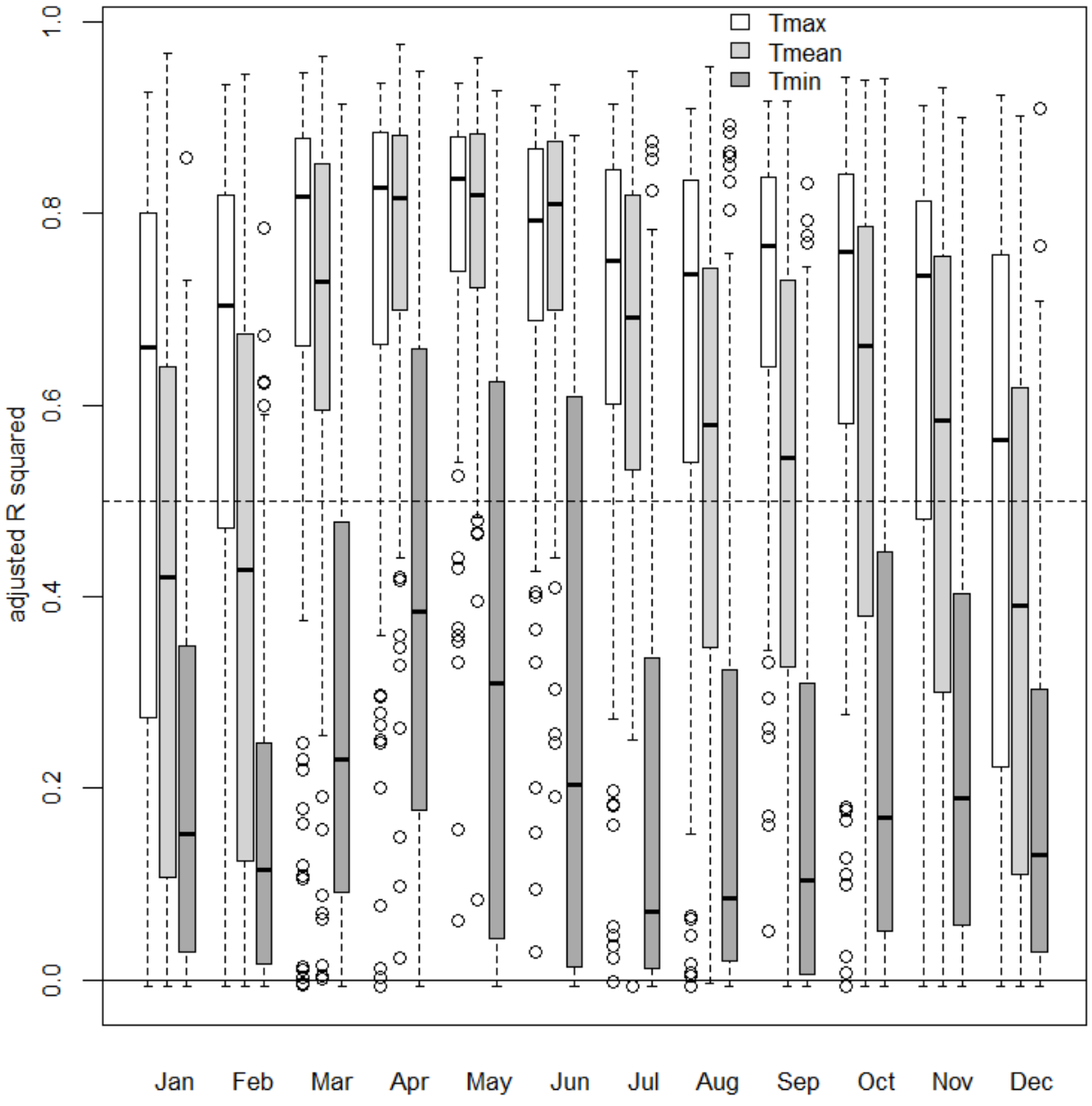
Page 10, lines 26-27: make the Tvar abbreviations match between the text and the table such that the table has Tmin, Tmax, and Td or the text has Tmin, Tmax and Tmean.
AC: Text and tables have been modified to consistently use Tmean.

Page 11, lines 5-6: Was the relationship with elevation or lack thereof numerically determined or eyeballed? To my eye, it looks like the tmin data do show a tendency toward smaller errors at high elevation although with greater variability.
AC: Relationship was statistically determined using correlations. Reviewing the results we see we missed the weaker relationship for Tmin, text has been updated (p.11 ll.19-20).

Page 11, line 13: So, the lapse rate analysis was only done over the mountain region? Were the results indeterminate for the prairies due to lack of elevation variability?
AC: Yes, only mountain sites were used. The reason was to have a more even distribution of sites through the full elevation range. By including prairie sites, lower elevations are over represented and can skew the lapse rate. Lapse rates using prairies alone were non significant

Page 11, line 15-16: Some indication of the coherence of the lapse rates would be interesting as well such as the R2 for each fitted regression. A similar plot to figure 5 could be made showing the scatter in R2 values. This would immediately indicate how frequently a linear model is a good one for temperature in the mountains here.

AC: Agreed that this is useful information. We have attached a figure here for the reviewer's interest, and added a short summary discussion to the text, p.11, l.31.



Page 11. I'm curious if the calculation of average monthly lapse rates is associative. Do you get the same value if you lump all of the month's data into a single lapse rate calculation as you would if you calculate the average of daily average lapse rates as you have here?
AC: Yes values are the same whether lapse rates are calculated for aggregated monthly data or daily and aggregated to a monthly value.

Pge 11, line 23: make read "There are significant difference in lapse rate between minimum..."

AC: Text modified as suggested (p.12 l.9).

Page 12: Just a comment of interest. It's Interesting about the seasonally steeper lapse rates in spring and fall. I know spring is the season for rain and more cyclonic stors in the front ranges of The Rockies. I wonder if this has to do with cloudiness at all? It's telling that tmin inversions are almost as frequent in the summer as they are in winter and they are minimal during the transitional seasons. This would align with a notion of greater IR loss in the clear sky periods of summer. It's surprising that there are any tmax inversions in summer. An indication of the quality of a linear fit might help determine if the numeric inversions are at all meaningfull or if a negative value happened to be the mean among a confused dataset.

AC: Certainly the Tmax summer inversion is anomalous and is associated with a low R^2, thus indicating a weak temperature/elevation relationship as opposed to a consistent inversion. Summer Tmin inversions are common during overnight clear sky conditions.