This paper needs a good comma clean-out!

Pg. 1 line 23 change “and” to “or”

Pg. 1 line 24 change “influencs that regulates” to regulators

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?

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”

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.

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.

Pg 2, line 18: strike “backcountry” unless you want to indicate some state of environmental disturbance.

Pg 2, line 29: decapitalize “City”

Pg. 3, line 6-7: How was the terrain variability determined? Was the mountain/prairie classification done manually or through some numeric means?

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.

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.

Pg. 3 line 24: strike “fabulously”

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.

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?

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.

Pg. 4, line 4: can you indicate in Fig. 1 which transect was retained?

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.

Pg. 5, lines 6-7 make read “...are the greatest due to solar heading, and can be problematic...”

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.

Pg. 5, lines 13-15: The averaging is also described earlier in the section. Combine those two efforts.

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.

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...

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.”

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

Pg. 7, line 32: insert “temperature” after diurnal.

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?

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?

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.

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.

Pg. 9, line 15: Instead of “poor performance” maybe more specifically say “spatial or temporal discontinuities in interpolated of modelled temperature surfaces”

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”.

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?

Pg 10, lines 10-15: The manual classification leaves the question regarding

Pg 10 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 needs to be acouple of paragraphs explaining this method after line 15 on thos page.

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.

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.

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?

Page 11, line 15-16: Some indication of the coherence of the lapse rates would be interesting as well such as the R^2 for each fitted regression. A similar plot to figure 5 could be made showing the scatter in

R^2 values. This would immediately indicate how frequently a linear model is a good one for temperature in the mountains here.

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?

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

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.