Author's Response to Anonymous Referee #1

We would like to thank Referee #1 for taking the time to review and offer feedback on this manuscript. Our responses to the comments and actions taken are listed below in red.

RC: P2, L15: In this paragraph, a short comment of how these stations are maintained after snowfall, or when they might get ice around, would be beneficial.

AC: This paragraph is more of an introduction to the SPICE sites and intercomparison period so a discussion of site maintenance may not be appropriate here. However, to answer the question, site maintenance was the responsibility of the site host and many sites kept service logs that were made available to improve data quality control. These logs usually note gauge servicing and instrument malfunctions.

Action: The following sentence has been added to the Methods sections to clarify how site maintenance issues were identified and handled in data quality control: "Where available, service logs were provided by site hosts to assist in data quality control and the identification of outliers due to servicing (e.g. rapid drops or increases in precipitation gauge bucket weights) or maintenance (e.g. instrument malfunctions or other human interventions that may impact the data)."

RC: P2, L28: What is a Precipitation detector?

AC: For SPICE, this was typically an optical disdrometer that is capable of detecting the occurrence of even light precipitation events at relative high temporal resolution (1-min in this case).

Action: Added the statement to describe "precipitation detector" as "typically an optical disdrometer capable of identifying the occurrence of even light precipitation.

RC: P3, L6: A reference after "demonstrating that the unshielded catch from both SUT types were very similar." would help in this sentence.

AC: This statement appears in the same sentence that references Kochendorfer et al. (2017b). The same reference applies to the latter statement in the sentence.

Action: To clarify, the sentence was changed to read "The transfer functions presented in K2017b were developed…, after the authors demonstrated that the unshielded catch…"

RC: P4, L9-11: I think that a more detailed explanation of this sentence would help to complete the idea. For example, answering what makes these evaluations "marginally independent"?

AC: We will modify this paragraph to clarify.

Action: The paragraph now states: "As discussed above, the data from all eight sites in Table 1, including data from multiple gauges of the same configuration at each of the sites, were combined to fit the transfer function models. This model, developed by pooling the data from multiple sites, was then applied to each individual gauge at each site. By applying a model developed from data collected at

multiple sites to individual gauges at each individual site, the authors maintained some independence between the model development and the evaluation."

RC: P4, L17: I would include this sentence as the main result in the abstract.

Action: The relevant sentence in the abstract now reads: "Due to the short intercomparison period, the dataset was not sufficiently large to develop and evaluate transfer functions using independent precipitation measurements, although on average the adjustments were effective at reducing the bias in unshielded gauges from -33.4% to 1.1%."

RC: P4, L29: What do you mean with "more natural circumstances"?

AC: We meant in situations more typical of those experienced by a user collecting data in the field. We will revise this sentence to reflect this.

Action: Sentence was changed to say: "The methodology used during SPICE for developing and evaluating the transfer functions used only a subset of the observed data (the SEDS), and although this was a robust methodology for developing transfer functions, it did not provide a comprehensive evaluation of the adjustments under circumstances more typical of users collecting precipitation data in the field where the data is less filtered to remove smaller amounts."

RC: P4, L36: This has to be consistent with P4, L9-11.

AC: We think that the revision on P4, L9-11 now sufficiently explains the issue such that no further action is required here.

Action: none taken

RC: P5, L3: I think the word "continued" should be in present tense.

AC: agreed

Action: "continued" change to "continues"

RC: P5, L7: I suggest completing the end of the sentence as: "...it will also isolate the performance to adjust snow measurements"

AC: agreed

Action: sentence modified as suggested

RC: P9, L15: I have a question here, why not using a standardized RMSE value, as it highly depends on the precipitation rates?

AC: There is advantages and disadvantaged to either. Standardizing the RMSE by using precipitation rate would allow better intercomparison between sites but information regarding the absolute magnitude of the error at each individual site would be lost. The most important interpretation here is the change (or

lack of) in RMSE with adjustment. Therefore, we think that using RMSE and cautioning the interpretation between sites, combined with the use of the other metrics, is the appropriate action.

Action: none taken

RC: P10, L1: Could you please add a number or a percentage to quantify the adjective"insubstantial"?

Action: We quantified this as < 0.005 mm for the single-Alter and as 0.05 mm for the unshielded gauges.

RC: P9, L11-15; P10, L5 L7-8; P10, L28-32: these sentences describe the statistical parameters used to analyze the data. I suggest moving them to the methodology section.

AC: Agreed

Action: The description and justification of the metrics have been moved to section 2.2 Performance metrics under Methods

RC: P12, L10-11, The number of events that are not captured by the single-Alter shielded in XBK is over 50%, do you have any comments to propose a better monitoring system in this location? Or how we could possibly integrate these cases to the adjusted precipitation?

AC: Fortunately, many of those event are small events which reduces the impact on total precipitation. However, they are still significant given the percentage of the reference precipitation amount not measured by the non-reference gauges. Currently, the most effective way to reduce this error is to increase the catch efficiency of the non-reference gauge by adding shielding. This is captured in the sentence: "Since more shielding (e.g. double-Alter) generally means a higher catch (Watson et al., 2008; Smith, 2009; Rasmussen et al., 2012; Kochendorfer et al., 2017b), more shielding would also reduce the number of unmeasured events." The use of disdrometers could also increase detection of small events in cold and windy conditons.

Action: Added the following sentence: "Perhaps another option to increase detection of small events in cold and windy locations would be the use of optical disdrometers paired with the conventional accumulating gauges."

RC: Table 1; Please, specify the meaning of "gh" in the table caption, or as a footnote.

Action: done

RC: Figures 3 and 4, I suggest increasing the panel size and/or font size.

AC: That's a good suggestion and we will make sure those are larger for final production. For review, we wanted to try to get them all on the same page.

Author's Response to Anonymous Referee #2

We would like to thank Referee #2 for their thorough review of this manuscript. We appreciate the detailed comments and suggestions. Our responses to the comments and actions taken are listed below in red.

Major comments:

RC: In the abstract and section 1.1, the authors discuss using liquid, mixed phase and solid precipitation. A possible useful addition to Table 1 would be statistics on the percentage of data that was liquid or mixed phase precipitation and the percentage of the data that was solid precipitation for each site. This could potentially help explain some of the results you see in the other tables when comparing all precipitation phases to snow only.

AC: We agree that this could be useful information. We calculated the percentage of precipitation events for all phases (solid/mixed/rain) for the two seasons combined that were used in this analysis using the temperature thresholds noted in this paper and in Kochendorfer et al. (2017b). They are as follows:

Site	% Precip Events by Phase solid/mixed/rain (%)	% Total Precip by Phase solid/mixed/rain (%)	Approximate Change in RTC from All to Snow
XBK	59/16/25	43/16/40	-10%
CAR	39/32/29	28/33/39	+3%
CCR	91/8/1	85/15/1	0%
FOR	31/43/26	28/49/23	-20%
HKL	47/38/15	44/35/21	-12%
MAR	44/33/23	32/45/23	+12%
SOD	54/39/8	49/42/8	+2%
WFJ	79/19/2	78/21/1	+2%

There were really no surprises in these numbers. We believe what the referee was alluding to in their comment was that we should expect that the biggest changes in Relative Total Catch (from all precip to solid precipitation) should occur at sites with the smaller percentage of solid precip. This would support the hypothesis that the higher total catch of rain and mixed events is masking the errors associated with adjusting only solid precipitation. The largest drop in RTC does occur at FMG (~-20%) which also has the lowest percentage of solid precipitation events during the season (31%). However, the next highest drops in RTC occur at HKL (~-12%) and XBK (~-10%) and both of these sites exhibit a higher percentage of solid precipitation events during the study period (47% and 59% respectively). Alternatively, we could look at the precipitation totals for each type (second column in table above). Although the percentage value for total snow is generally lower than for event occurrence, it doesn't change the interpretation.

The relative amount of snow as compared to rain/mixed could possibly contribute to the difference in the statistics for FOR, but doesn't explain the differences at the other sites. We believe that the addition of a statement in the discussion referring to this will not improve the clarity of interpretation.

Action: Added columns in Table 1 to show the relative total precipitation by phase for each site. Added a reference to this information in Section 2.2.

RC: The authors show analysis using the wind data collected from 10m heights and gauge height and discuss many issues related to measuring the wind speeds. In the discussion starting on page 13, line 20, as well as the first part of the conclusions section, no mention is made of the possible impacts of wind variations over the 30-minute averaging period. For the windier sites used in the study, how variable were the winds over the 30 minutes and what role might that have played in some of the results? I agree with the discussions and results the authors have expressed regarding some of the other wind-related issues but this should also be addressed.

AC: This very issue was addressed in previous work during the WMO-SPICE project and was published both in Wolff et al. (2015) and Nitu et al. (2018). Wolff et al. looked at variations in both 30-minute temperature and wind speed at HKL (one of the windier SPICE sites) and concluded that removing points with large variability within the 30-minute period did not significantly change the results.

Action: The following sentences were added to the end of the discussion paragraph that formerly started on page 13, line 20: "Additional uncertainty related to wind speed may be attributed to the variability within the 30-min mean period. Although this wasn't included in the current analysis, previous work by Wolff et al. (2015) and Nitu et al. (2018) at HKL showed that the impact of high frequency variability in the wind speed over 30-min periods on transfer functions was negligible."

RC: Regarding the GEONOR gauges used in the study, the authors state they were GEONOR T-200B3 gauges.   I believe this implies a three-wire GEONOR gauge.   If that is the case, how were the data from the three wires used? Were they averaged? This should be discussed and/or clarified in the introduction or methods sections.

AC: The three Geonor wires were averaged before filtering.

Action: The methodology was revised to reflect this

RC: With regards to Fig 1, equation 2 shows a collection efficiency of 1.1 at a wind speed of 0, but I don't see a clear explanation for this. There is some discussion on P4, L1 that this might be indicative of rain at air temps > 2C, but equation 2 does not use temperature so I'm a bit confused as to why temperature is mentioned there and why eqn 2 would result in a catch efficiency > 1 at 0 m/s. Some clarification and additional explanation would be useful here.

AC: This is a good question and should have been addressed in the paper. Simply, the Equation 2 catch efficiency was greater than 1.0 because of the specific empirical fit to the WMO-SPICE catch efficiency data that produced the smallest error. It is due to the shape of the data that the equation was fit to, rather than a physical cause. This function was not originally published in this present manuscript, but

rather in K2017B. For this analysis, any catch efficiency that was calculated as > 1 was automatically reset to 1. This occurred relatively infrequently, since the mean wind speed during precipitation has to be less than 0.5 m s^{-1} and therefore did not significantly affect the results.

The discussion on P4, L1 does not explain why the catch efficiency for Eq. 2 is greater than 1.0. This text simply explains how liquid precipitation, classified as precipitation occurring when Tair > 2 deg C, should be handled according to K2017b.

Action: The following has been added to the manuscript in the methodology section (2.2): "Figure 2 suggests that Eq. 2 can exceed a catch efficiency of 1 at low wind speeds. There is no obvious physical explanation for the portion of the Eq. 2 catch efficiency function (originally published by K2017b) that is > 1.0 and this is related to the empirical fit of the catch efficiency curve to the original SPICE data. For this current assessment, calculated catch efficiencies > 1 were infrequent and occurrences were automatically set to 1."

RC: In section 3.2 and 3.3, there is mention of potential shadowing of the wind sensor at the MAR site. Was anyone at the site contacted and asked for pictures of the sensor at the site to confirm this (assuming none of the authors have visited or work at this site)? This would seem to be an easy thing to do without supposition. Could this also be a calibration issue with the wind sensor at that site?

AC: Co-author J. Kochendorfer is quite familiar with the MAR site and has done extensive work during the SPICE project on the wind measurements. This was outlined in Kochendorfer et al. 2017a. During the development of the Kochendorfer et al. (2017a) transfer functions, the ratio of the 30-min gauge-height wind speed to the 30-min 10 m height wind speed was plotted as a function of wind direction, and it was quite clear that the gauge-height anemometer was affected by obstacles in some wind sectors.

Action: P8, L31 has been changed from, "wind speed measurements may be shadowed…", to, "wind speed measurements were shadowed…". This statement was not based on supposition.

Minor comments:

RC: Throughout the manuscript: There is inconsistent use of the hyphen in the phrase "single Alter shielded". In some places, there are two hyphens, in others, just one. I believe it should be "single Alter-shielded".

Action: The wording in the text should now consistently be "single Alter-shielded"

RC: P1, L17 – Intercomparison is hyphenated here but nowhere else. The official SPICE title appears to not have the hyphen.

AC: Inter-Comparison was supposed to have been hyphenated so that the acronym SPICE made more sense, but it appears to have been dropped in the literature along the way. This should be consistent

Action: updated the text without the hyphen in Intercomparison.

RC: P1, L20 – The term "windshield" is typically used to describe the front glass on a car. In the case of precipitation measurement, I believe "wind shield" is more commonly used.

AC: agreed

Action: changed the text so that "wind shield" is consistently used.

RC: P1, L25-27, The sentence on these lines is somewhat awkwardly written.

Action: Changed to "Performance is assessed in terms of relative total catch (RTC), root mean square error (RMSE), Pearson correlation (r), and Nash-Sutcliffe Efficiency (NSE). Metrics are reported for combined precipitation types, and for snow only."

RC: P1, L29 – What is meant by gauge configuration?

AC: This refers to the gauge model and the wind shield type (if any) that the gauge is paired with.

Action: clarified this in the text by adding "…and gauge configuration *(gauge and wind shield type)*."

RC: P5, L24 – This sentence reads a bit awkwardly and may just be missing a comma. Perhaps revise into two sentences.

Action: These sentences now read: "The algorithm removes random and systematic diurnal noise, but does not account for signal drift (an example of signal drift is a decrease in weight that occurs due to evaporation of water from the gauge bucket)."

RC: P6, paragraph 2 – How often were the temp/wind speed data missing?

AC: We can qualitatively say that the number of missing temperature and wind speed data were low, and generally the missing temperature and wind speed data coincided with missing precipitation data. Since missing data is not included in the analysis, this is largely inconsequential.

Action: none taken

RC: P6, L34 – there is an extra comma on this line

Action: fixed

RC: P6, L35 – there shouldn't be a comma after "snow".

Action: fixed

RC: P7, L5 – The term "both" is not necessary in this sentence.

Action: fixed

RC: P8, L6 – The term "both" is not necessary here either.

Action: fixed

RC: P11, L18 – season should be plural.

Action: fixed

RC: P11, L20 – I think you meant to say compounding instead of confounding.

Action: changed this to "complex".

RC: P12, L16 – There is a) missing on this line.

Action: fixed

RC: Figures 5 through 8 – One suggestion the authors might want to consider is to change the colors of the bars and use warm colors (e.g. red, orange, yellow, or shades of red) for the single Alter data and cool colors (blue, green, purple or shades of blue) for the unshielded data. This would really make the differences between the two data sets much more obvious and quickly draw the reader's attention to the points you make in the manuscript

AC: Agreed.

Action: We changed the colour of the bars so that the shielded gauges are shown in grey tones while the unshielded gauges are shown in blue tones. We also made some changes to help with discerning the bars if printed in grey scales.

Evaluation of the WMO-SPICE transfer functions for adjusting the wind bias in solid precipitation measurements

Craig D. Smith[1], Amber Ross[1,2], John Kochendorfer[3], Michael E. Earle[4], Mareile Wolff[5], Samuel Buisan[6], Yves-Alain Roulet[7], Timo Laine[8]

[1]Climate Research Division, Environment and Climate Change Canada, Saskatoon, S7N 3H5, Canada
[2]Global Institute for Water Security, University of Saskatchewan, Saskatoon, S7N 3H5, Canada
[3]Atmospheric Turbulence and Diffusion Division, ARL, National Oceanic and Atmospheric Administration, Oak Ridge, 37830, USA
[4]Meteorological Service of Canada, Environment and Climate Change Canada, Dartmouth, B2Y 2N6, Canada
[5]Norwegian Meteorological Institute, Oslo, 0313, Norway
[6]Delegación Territorial de AEMET (Spanish National Meteorological Agency) en Aragon, Zaragoza, 50007, Spain
[7]MeteoSwiss, Payerne, 1530, Switzerland
[8]Finnish Meteorological Institute, Helsinki, 00101, Finland

Correspondence to: Craig D. Smith (craig.smith2@canada.ca)

Abstract. The World Meteorological Organization (WMO) Solid Precipitation Interc Comparison Experiment (SPICE) involved extensive field intercomparisons of automated instruments for measuring snow during the 2013/2014 and 2014/2015 winter seasons. A key outcome of SPICE was the development of transfer functions for the wind bias adjustment of solid precipitation measurements using various precipitation gauge and wind shield configurations. Due to the short intercomparison period, the dataset was not sufficiently large to develop and evaluate transfer functions using independent precipitation measurements, although on average the adjustments were effective at reducing the bias in unshielded gauges from -33.4% to 1.1%. The present analysis uses data collected at eight SPICE sites over the 2015/2016 and 2016/2017 winter periods, comparing 30-minute adjusted and unadjusted measurements from Geonor T-200B3 and OTT Pluvio[2] precipitation gauges in different shield configurations to the WMO Double Fence Automated Reference (DFAR) for the verification of the transfer function. Performance is assessed in terms of relative total catch (RTC), root mean square error (RMSE), Pearson correlation (r), and Nash-Sutcliffe Efficiency (NSE). Metrics are reported for combined all precipitation types, and for snow only. The evaluation shows that the performance varies substantially by site. Adjusted RTC varies from 54% to 123%, RMSE from 0.07 mm to 0.38 mm, r from 0.28 to 0.94 and NSE from -1.88 to 0.89, depending on precipitation phase, site, and gauge configuration (gauge and wind screen type). Generally, windier sites such as Haukeliseter (Norway) and Bratt's Lake (Canada) exhibit a net under-adjustment (17% to 46%), while the less windy sites such as Sodankylä (Finland) and Caribou Creek (Canada) exhibit a net over-adjustment (2% to 23%). Although the application of transfer functions is necessary to mitigate wind bias in solid precipitation measurements, especially at windy sites and for unshielded gauges, the inconsistency in the performance metrics among sites suggests that the functions be applied with caution.

1 Introduction

The World Meteorological Organization (WMO) Solid Precipitation Interc Comparison Experiment (SPICE) was a Commission for Instruments and Methods of Observation (CIMO) initiative to assess and compare instruments and

methods for measuring solid precipitation (Nitu et al., 2012; Nitu et al., 2018). The objectives were: 1) to make recommendations for appropriate automated field reference systems; and 2) to provide guidance on the performance and operation of automated systems for measuring solid precipitation and snow on the ground. SPICE was motivated by the need for accurate and homogenized solid precipitation measurements. For example, such measurements are required for climate trend analysis in northern regions (e.g. Førland and Hanssen-Bauer, 2000; Yang and Ohata, 2001; Scaff et al., 2015). Following historical works on adjusting the systematic undercatch of solid precipitation measurements due to wind (Goodison, 1978; Sevruk et al., 1991; Goodison et al., 1998; Sevruk et al., 2009; Smith, 2009; Wolff et al., 2015; Kochendorfer et al., 2017a; Buisan et al., 2017), a methodology and a set of widely applicable transfer functions for the adjustment of high resolution (i.e. 30-min) precipitation measurements was developed. The SPICE transfer functions discussed in this study were developed for single -Alter-shielded or unshielded automated precipitation gauges (Kochendorfer et al., 2017b). Because of the symbioses between the Kochendorfer et al. (2017b) SPICE work and this evaluation, the SPICE methodology is described below in more detail and henceforth cited as K2017b.

The official SPICE intercomparison period occurred during the winters of 2013/2014 and 2014/2015. During this period, a Double Fence Automated Reference (DFAR) was operated at eight test bed sites: Bratt's Lake (XBK), Caribou Creek (CCR), the Centre for Atmospheric Research and Experiments (CARE; abbreviated to CAR), Formigal (FOR), Haukeliseter (HKL), Marshall (MAR), Sodankylä (SOD), and Weissfluhjoch (WFJ) (Fig. 1; Table 1). The DFAR was developed for use as the field reference configuration for SPICE (Nitu, 2012; Nitu et al., 2016; Nitu et al., 2018). A detailed layout and site description for each of these eight sites can be found in the WMO-SPICE site commissioning reports at: http://www.wmo.int/pages/prog/www/IMOP/intercomparisons/SPICE/SPICE.html and in the WMO-SPICE final report (IMO 131 found at http://www.wmo.int/pages/prog/www/IMOP/publications-IOM-series.html; Nitu et al., 2018).

The DFAR consisted of either a Geonor T-200B3 or an OTT Pluvio^2 automated weighing precipitation gauge with a single- Alter -shield inside the same large, octagonal double fence used for the WMO Double Fence Intercomparison Reference (DFIR). The DFIR was employed as a manual field reference configuration during previous solid precipitation intercomparisons (Yang et al., 1993; Goodison et al., 1998). The DFAR incorporated a precipitation detector as a verification tool to reduce the probability of using false precipitation reports in SPICE data analysis. The result was the development of a high confidence reference precipitation data set called the Site Event Data Set (SEDS; Reverdin, 2016). In the SEDS, a precipitation event was a 30-minute period during which athe precipitation detector (typically an optical disdrometer capable of identifying the occurrence of even light precipitation) observed precipitation for at least 18 minutes during the 30-minute period (60% of event duration) and the DFAR measured ≥ 0.25 mm. The justification for the filtering criteria used in K2017b is detailed in Kochendorfer et al. (2017a). The SEDS was then used to produce the SPICE transfer functions. By combining data from each site in Table 1, the intent was to make these multi-site transfer functions universally applicable.

1.1 The K2017b transfer functions

The transfer functions presented in K2017b were developed for both unshielded and single -Alter--shielded automated precipitation gauges by combining observations from the Geonor T-200B3 and OTT Pluvio[2] gauges (hereafter referred to as the sensors under test, or SUT), after the authors demonstrateding that the unshielded catch from both SUT types were very similar. Using the SEDS, further processing was applied to the SUT data (as justified in Kochendorfer et al., 2017a) using a minimum 30-minute threshold defined as the median of the ratio of the SUT accumulation to that of the DFAR, for the precipitation event threshold of 0.25 mm. This prevented the results from becoming biased toward the gauge used in the SEDS event selection. For wind speed, K2017b used the wind speed measurements available at each site and applied the log-profile law to produce a 30-minute average wind speed for both gauge height (U_{gh}) and the standard 10 m height (U_{10m}), either incrementing speeds to U_{10m} or decreasing speeds to U_{gh}, depending on which measured wind height speed data was deemed the best at each site. Catch efficiencies (C_E) were then calculated for each event as the ratio of the 30-minute SUT precipitation accumulation to that from the DFAR over the same period. From K2017b, two functional forms were used to fit the data:

$$C_E = e^{-a(U)(1-\tan^{-1}(b(T_{air}))+c)},$$ (1)

and

$$C_E = (a)e^{-b(U)} + c,$$ (2)

where U is wind speed (specifically either U_{gh} or U_{10m}) in m s^{-1}, T_{air} is air temperature in degrees C, and a, b, and c are coefficients to fit the data to the model. Eq. 1 and Eq. 2 listed here are referred to as Eq. 3 and Eq. 4 in K2017b. To reduce the impact of fewer events at higher wind speeds and the potential impacts of blowing snow, the SEDS data was filtered further to remove events with wind speeds higher than 7.2 m s^{-1} (9 m s^{-1}) at gauge height (10 m).

The key difference between Eq. 1 and Eq. 2 is the inclusion of temperature dependency in Eq. 1. This allows for a continuous (3-dimensional) transfer function at all temperatures without having explicit knowledge of the precipitation phase. This curve is shown in Fig. 2 (red) for both the single -Alter--shielded (solid) and unshielded (dashed) gauges and a temperature of -5 °C. Equation 2, however, requires an assessment of precipitation phase (liquid, solid, or mixed), with each phase having unique coefficients. K2017b used temperature to discriminate phase for Eq. 2 and assumed the following: solid precipitation occurs at $T_{air} < $ -2 °C; liquid precipitation occurs at $T_{air} > $ 2 °C; and mixed precipitation occurs at -2 °C $\leq T_{air} \leq$ 2 °C. Equation 2 is plotted in Fig. 2 (blue) using the coefficients for snow for the single -Alter--shielded (solid) and unshielded (dashed) gauges. For both equations and gauge configurations, unique coefficients were derived for each of the two wind speed measurement heights. The SUT precipitation were adjusted

3

using both Eq. 1 and Eq. 2 by employing the appropriate coefficients, depending on phase/temperature, wind measurement height, and shield configuration. For Eq. 2, precipitation classified as rain ($T_{air} > 2$ °C) assumed a C_E equal to 1. The coefficients used in Eq. 1 and Eq. 2 are detailed in K2017b and are not repeated here.

1.2 K2017b results

Following the end of the SPICE project, the performance of the SPICE transfer functions was evaluated as described in K2017b, using the same SEDS data used to develop those transfer functions. As discussed above, the data from all eight sites in Table 1, including data from multiple gauges of the same configuration at each of the sites, were combined to fit the transfer function models. This model, developed by pooling the data from multiple sites, was then applied to each individual gauge at each site. By applying a model developed from data collected at multiple sites to individual gauges at each individual site, the authors maintained some independence between the model development and the evaluation. The evaluation of the universal transfer function equations was completed for each individual site, arguably making the evaluation at least marginally independent from the pooled, multi-site measurements used to develop the transfer functions. The K2017b results were based on four metrics: the root mean square error (RMSE), mean bias, Pearson correlation (r), and percentage of events (PE) between the DFAR and SUT that agreed within a specified threshold (typically 0.1 mm). It should also be noted that the overall performance metrics summarized in K2017b included all precipitation phases, whether adjusted by the transfer functions (i.e. solid and mixed) or not (i.e. liquid).

K2017b showed that the SPICE transfer functions reduced the overall bias in the unshielded precipitation gauges (both Geonor T-200B3 and OTT Pluvio^2, all sites combined) from -33.4% to 1.1%, but the results varied by site, with CAR and WFJ showing over-adjustment and HKL, FOR, and XBK showing under-adjustment. For the most part, the r, RMSE, and PE were slightly improved after adjustment, but these also varied by site. K2017b also showed that, in general, the mountainous sites experienced larger errors after adjustment, with one mountainous site (WFJ) being over-adjusted and the other two (HKL and FOR) being under-adjusted.

1.3 Motivation for the extended evaluation

The impetus for this extended evaluation was twofold:

1) The methodology used during SPICE for developing and evaluating the transfer functions used only a subset of the observed data (the SEDS), and although this was a robust methodology for developing transfer functions, it did not provide a comprehensive evaluation of the adjustments under circumstances more typical of users collecting precipitation data in the field where the data is less filtered to remove smaller amounts more natural circumstances. This evaluation will use winter precipitation data as they would be collected in an operational or monitoring application, produce time series of adjusted winter precipitation, and evaluate the performance of the adjustment relative to the reference.

4

2) The dataset used for the evaluation of transfer functions in K2017b was not completely independent of that used to develop the functions. The present evaluation will use two additional years of data collected following the end of the SPICE intercomparison period, and hence will provide an independent means of assessment.

5

This evaluation will examine the performance of the SPICE transfer functions for precipitation measurements from each of the eight intercomparison sites shown in Fig. 1 for the 2015/2016 and 2016/2017 winter seasons. Each of these sites continuesd to operate a DFAR following the SPICE intercomparison period, a critical component for assessing transfer function performance. The assessment will be conducted for both unshielded and single- Alter shielded gauges using wind speed heights of 10 m and gauge height, where available. Extending the K2017b evaluation, this assessment will not only look at the performance using data over the entire winter season (regardless of precipitation phase), it will also isolate the performance to adjust for snow measurements.

2 Methods

Precipitation and ancillary meteorology data at 1-minute resolution for the 2015/2016 and 2016/2017 winter periods were obtained from the eight SPICE sites listed in Table 1. The data were quality controlled using the same techniques employed in SPICE (Nitu et al., 2018), which involved automated range and jump checks and supervised removal of remaining outliers. Where available, service logs were provided by site hosts to assist in data quality control and the identification of outliers due to servicing (e.g. rapid drops or increases in precipitation gauge bucket weights) or maintenance (e.g. instrument malfunctions or other human interventions that may impact the data). The Geonor T-200B3 gauges employed three transducers and the output from all three were averaged to produce a single time series. Next, the high-resolution precipitation data were subjected to the same Gaussian filter as the SPICE data used in K2017b to dampen high frequency noise; however, it was decided not to develop an event database such as the SEDS, but rather to use an alternate process to develop a consistent 30-minute precipitation time series to more closely reflect real-world precipitation datasets. This alternate process involves the use of a modified "Brute-Force" filter initially described in Pan et al. (2016), henceforth called the Neutral Aggregation Filter (NAF). NAF is described in more detail in Smith et al. (2019).

2.1 Neutral Aggregation Filter

The NAF algorithm removes noise in cumulative precipitation time series by iteratively balancing positive and negative noise and accumulating positive changes exceeding the noise by a user-defined threshold (Δ^*, e.g. 0.05 or 0.2 mm, depending on the gauge precision) (Smith et al., 2019) such that the total accumulated positive increases in bucket weight after filtering are forced to equal the total end-of-season bucket weight. The algorithm removes random and systematic diurnal noise, but does not account for signal drift (an , for example of signal drift is a decrease in weight that occurs due to evaporation of water from within the gauge bucket). Signal drift can result in estimation errors, which can be mitigated using an iterative manual process, with the NAF output as a first guess. This process is called NAF Supervised (NAF-S) and lets the user select the beginning and the end points of segments within the time

series where evaporation is occurring. The process then removes these segments so that they have no impact on the time series. Because there is some user subjectivity in selecting the beginning and end points of impacted segments, this process is completed by a single user employing pre-determined and consistent criteria. Although beyond the scope of the present work, testing NAF and NAF-S on both simulated and observed precipitation time series over an entire winter season, including both noise and evaporation drift, showed the technique to be effective with low error as compared to the control. The end product of the NAF-S filter is a clean, time consistent 1-minute accumulating time series with preserved data gaps (i.e. no gap filling) for each gauge configuration for each season.

2.2 Amalgamation and adjustment

To produce accumulation periods consistent with the K2017b validation, the 1-minute NAF-S precipitation accumulation time series were resampled to 30-minute accumulation amounts via differentiating the bucket weights between the start and end of the 30-minute period. The continuity of the time series was maintained, despite data gaps, through the assumption that the gauge continues to accumulate precipitation despite logger or power outages. In the event of missing data, precipitation accumulation during outages was calculated based on the differential bucket weight between the start and end of the outage and recorded as an accumulation at the end of the outage. Although this preserves the total accumulation during the outage, information related to the timing of the events during the outage are not preserved, and the accumulation data need to be flagged. Protocols for adjusting the data for undercatch are noted below.

The 1-minute wind speeds (U_{10m} and U_{gh} where available) and air temperatures (generally measured at 1.5 m) were averaged over the same 30-minute periods as the accumulated precipitation amounts. Site specific details on the ancillary measurements can be found in the SPICE site commissioning reports (referenced in Section 1). If more than 10 minutes of wind or temperature data were missing in any 30-minute period, the data were flagged as missing and were not used in the adjustment procedure.

The resultant time series for each test gauge at each site were adjusted separately using both Eq. 1 and Eq. 2. Each 30-minute accumulation was adjusted individually if the following conditions were met: 1) both the start and end bucket weights for the 30-minute period were not missing, such that the differential could be determined; and 2) no more than 10 of the 30-minute values of either wind speed or temperature were missing. Periods that did not meet these criteria were preserved in the time series, but were flagged as being unadjusted and were not included in the validation.

For adjustments using Eq. 1, the pre-determination of the precipitation phase was not necessary, as the transfer function is continuous with temperature and not directly dependent on precipitation phase. For the purpose of adjusting precipitation using Eq. 2, phase is determined by air temperature using the phase regimes outlined in Section 1.1 with rain assuming a catch efficiency of 1 (and is therefore not adjusted). The relative total amounts (Snow/Mixed/Rain) as measured by the DFAR at each site are shown in Table 1. The same maximum wind speed thresholds for adjustments were employed here as in K2017b, which were 7.2 m s^{-1} and 9.0 m s^{-1} at gauge height (generally 2 m above ground) and at 10 m height respectively. Wind speeds above these thresholds were set at the threshold value to

6

avoid over adjustment and the increased uncertainty in the transfer functions above the wind speed threshold. Figure 2 suggests that Eq. 2 can exceed a catch efficiency of 1 at low wind speeds. There is no obvious physical explanation for the portion of the Eq. 2 catch efficiency function (originally published by K2017b) that is > 1.0 and this is related to the empirical fit of the catch efficiency curve to the original SPICE data. For this current assessment, calculated catch efficiencies > 1 were infrequent and occurrences were automatically set to 1.

The resulting data (https://doi.org/10.1594/PANGAEA.907379; Smith et al., 2019b) includes a sub-set of adjusted and unadjusted 30-minute precipitation amounts for each SUT gauge configuration at each site (adjusted using Eq. 1 and Eq. 2 and using either U_{gh} or U_{10m} or both where available), the 30-minute DFAR data, and the accumulated gap preserved time series for each with flags identifying the periods that were not adjusted.

The performance of adjustments were assessed using the relative total catch (RTC; defined as the total catch of the gauge under test as compared to the DFAR, expressed as a percentage), the root mean square error, Pearson correlation, , and the Nash-Sutcliffe Efficiency (NSE). The assessment considered overall performance for all precipitation types combined, as well as for snow, alone. In the latter case, the assessment of snow adjustments using Eq. 1 employed the same temperature threshold for snow as K2017b (T_{air} < -2 °C). Intercomparison results are reported for both unshielded and single -Alter-shielded sensors under test, which can be either a Geonor T-200B3 or OTT Pluvio[2]. Where multiple gauges of the same configuration are present at a site, these gauges are assessed individually and as a combined dataset.

For evaluation purposes, and where possible, the following circumstances will be assessed for ~~both~~ all precipitation types (includes rain, snow, and mixed) and snow only: a) adjustments using Eq. 1 vs. Eq. 2, b) adjustments using gauge height vs. 10 m wind speeds, and c) adjusting single -Alter-shielded vs. unshielded gauges. Based on site-by-site evaluations, some insight will be provided as to the performance of transfer functions in different environments, and under different climate characteristic conditions.

2.2 Performance metrics

Four statistical metrics are used to assess the performance of the transfer functions for adjusting precipitation. These are relative total catch (RTC), root mean square error (RMSE), Pearson correlation (r) and Nash-Sutcliffe Efficiency (NSE). RTC is the ratio of accumulated catch of the SUT versus the accumulated catch of the reference over the same period and using the same filter or thresholds (i.e. for snow). RTC is expressed as a percentage of the reference precipitation and reflects the capability of the transfer functions to correctly adjust seasonal and long term total precipitation. RMSE is used to estimate the uncertainty of unadjusted and adjusted precipitation measurements relative to the reference. However, the variability in the magnitude of the RMSE amongst the sites should be interpreted with caution, as a small RMSE at a site with low precipitation rates may be more significant than a higher RMSE at a site that has higher precipitation rates. For that reason, the RMSE will mainly be used to assess the relative performance of Eq. 1 and Eq. 2 at each site, as well as the relative performance of the transfer functions for the single Alter-shielded and unshielded gauges. 3.3 Pearson correlation. The Pearson correlation assesses the strength of linear relationships between the SUT and the references before and after adjustments using Eq. 1 and Eq. 2. The Nash-Sutcliffe Efficiency

(NSE) coefficient is an alternative goodness-of-fit indicator for the agreement between the reference and the unadjusted and adjusted precipitation measurements. The measure has a higher sensitivity to bias and outliers than the Pearson Correlation coefficient and this can have both advantages and disadvantages in this assessment. It is included here to complement the other metrics used in this analysis. The NSE coefficient can vary from $-\infty$ to 1, where 1 indicates a perfect fit between the reference and the SUT.

3 Results

3.1 Time series

The impact and the performance of transfer functions for adjusting precipitation can be examined by comparing the accumulation time series for unadjusted and adjusted data to the reference. Figures 3 and 4 show unadjusted, adjusted (Eq. 1 only), and reference (DFAR) time series of precipitation accumulation for the unshielded and single -Alter-- shielded gauges at all sites for the 2015/2016 and 2016/2017 winter seasons, respectively. Where more than one gauge with the same shield configuration was present at a site, and where more than one wind speed height was available, results for only one gauge and wind speed height were selected for illustrative purposes.

Figures 3 and 4 show the relative impacts of wind on undercatch for each of the eight sites and the relative effectiveness of the transfer function adjustments on each SUT configuration (shielded and unshielded) for the two winter seasons separately. Note that the season lengths vary by site and season (depending on both the actual length of the winter season and on data availability) and the scale of the vertical axis changes with site and season to show the relative scale of the bias and the adjustment. Gaps in the series represent missing data, with total accumulations during the gap obtained from the bucket weight change (in both the DFAR and the SUT); the gap accumulations were preserved but not adjusted.

The precipitation amounts vary by site and season, but the general trends in SUT undercatch as compared to the DFAR are consistent. Accordingly, the impact of the adjustment also appears to be consistent. Without considering precipitation phase partitioning, unadjusted precipitation (solid lines) relative to the DFAR was always lowest at the windy sites of XBK and HKL. Referring to unadjusted precipitation, unshielded gauges (blue lines) always catch less precipitation than the single -Alter--shielded gauges (red lines) at all sites, and during both winters. The transfer function (only temperature dependent Eq. 1 adjustments shown here) appears to be less effective at the windier sites (wind speeds during precipitation events are shown in Table 2), with a substantial undercatch remaining after adjustment. Further, precipitation is over-adjusted at some of the less windy sites (CAR, CCR, and SOD).

3.2 Relative total catch

The relative total catch (RTC) is an important performance metric for climate monitoring and analysis and reflects the capability of the transfer functions to correctly adjust seasonal and long term total precipitation. The RTC metrics for the single -Alter--shielded gauges are shown in Table 2 and for the unshielded gauges in Table 3 for Eqs. 1 and 2, combining the data from the two intercomparison seasons (CCR being the exception, since data is only available for

8

2016/2017). RTC is shown for ~~both~~ all precipitation phases and for snow, and for both wind measurement heights (U_{10m} and U_{gh}), where available. If there are multiple gauges per site, the RTC is reported separately for each gauge and for the combined gauge dataset. Figure 5 summarizes the snow RTC in the tables for the combined results for both winter seasons. When more than one gauge exists, the U_{gh} was used for the adjustment for all sites except FOR,

5   which only reported U_{10m}.

The RTC values in Tables 2 and 3 indicate that the unadjusted catch for snow is lower than the unadjusted catch for all precipitation types at most of the sites. The magnitude of the difference depends on the relative amounts of solid and liquid precipitation received during the season, as well as the wind speeds during snow. The biggest difference in the unadjusted RTC values between all precipitation types and snow occurred at sites where more rain occurred during

10   the intercomparison season, as the gauge catch for rainfall is naturally higher (Yang et al., 1998; Smith, 2008) and biases the total catch. At XBK, CAR, FOR, and HKL, removing rain and mixed precipitation from the statistics had a large impact on the RTC, both pre- and post-adjustment, and provides a more realistic metric for assessing how well the transfer functions are performing for the adjustment of snow measurements.

Although the sample size is smaller (fewer unshielded than single -Alter- shielded gauges), the unadjusted catch of

15   the unshielded gauges (Table 3) was lower than the unadjusted catch of the single -Alter- shielded gauges (Table 2). From the single -Alter- shielded RTC in Table 2, focusing on snow, the differences between the Eq. 1 and Eq. 2 results were small, varying within 1 to 2 %. This can also be seen in the combined results in Fig. 5. The difference between Eq.1 and Eq. 2 was greater for the unshielded gauges (Table 3) with Eq. 1 performing better than Eq. 2 for snow at XBK (+12 %) and HKL (+10% using U_{gh}). Equation 2 tended to under-adjust the unshielded gauges more than Eq. 1

20   at XBK and HKL.

At sites with both wind speed heights available for use in the adjustment (XBK, CAR, HKL, and MAR), the data shown in Table 2 for single -Alter- shielded gauges suggest that using U_{gh} reduces the extent of over-adjustment (CAR, MAR) or under-adjustment (XBK and HKL) relative to U_{10m} (i.e. adjustments closer to 100%). This holds for the unshielded gauge adjustments in Table 3, with the exception of MAR, which shows a large under-adjustment using

25   U_{gh} as compared to U_{10m}. This may suggest that wind speeds are biased low at MAR, which is consistent with comments made by Kochendorfer et al. 2017a stating that the ground height wind measurement were~~may be~~ shadowed in some directions. Although the sample size was small, there is reason to suggest from an RTC perspective that U_{gh} outperforms U_{10m} when used in Eq.1 and 2 for adjusting snow measurements. For this reason, where available, the U_{gh} rather than U_{10m} wind adjustments are shown in Fig. 5 and subsequent figures.

30   The differences in RTC for adjusted measurements from single -Alter- shielded gauges versus those from unshielded gauges are mixed. At the windier HKL and XBK sites, Fig. 5 suggests that the adjusted RTC for the unshielded gauge is just as high as or higher than for the single -Alter- shielded gauge (Eq. 2 at XBK being the exception). The over-adjustments at SOD and CCR are exaggerated for the unshielded gauge, but the unshielded adjustment is closer to 100% at CAR and WFJ. MAR is an outlier in Fig. 5, possibly due to the potential issue with U_{gh}, but the U_{10m}

9

adjustment of the unshielded gauge (Table 3, snow) has an RTC closer to 100% (105% and 106% for Eq. 1 and Eq. 2 respectively) than the single -Alter--shielded gauge (117%).

Including the RTC for both wind speed heights but excluding the combined gauge statistics, the adjustment for snow using Eq. 1 increases the mean catch efficiency for the single -Alter--shielded gauge from 61% to 88% and for the unshielded gauge from 48% to 92%.

3.2 RMSE

~~The RMSE was used to estimate the uncertainty of unadjusted and adjusted precipitation measurements relative to the reference. However, the variability in the magnitude of the RMSE amongst the sites should be interpreted with caution, as a small RMSE at a site with low precipitation rates may be more significant than a higher RMSE at a site that has higher precipitation rates. For that reason, the RMSE will mainly be used to assess the relative performance of Eq. 1 and Eq. 2 at each site, as well as the relative performance of the transfer functions for the single-Alter shielded and unshielded gauges.~~

Table 4 (single -Alter--shielded gauges) and Table 5 (unshielded gauges) show the RMSE for each available SUT at each site, as well as the RMSE when multiple SUTs are combined. As with RTC, the metric is provided for both transfer functions and using both wind speed heights, where available.

Comparing the RMSE values between Table 4 and Table 5, the RMSE values for both the adjusted and unadjusted unshielded gauges are higher than their single -Alter--shielded counterparts. However, the RMSE differences between all precipitation phases and those for snow are inconsistent, with RMSE occasionally lower for snow than for all phases, and vice versa. For the single -Alter--shielded and unshielded gauges, respectively, 46% and 36% of the RMSE values are either lower or the same for snow as compared to all precipitation phases. The differences in adjusted RMSE with wind measurement height are also small. For single -Alter--shielded gauges, CAR has a lower RMSE for U_{gh}, MAR has a higher RMSE for U_{gh}, and HKL and XBK show similar RMSE values for U_{gh} and U_{10m}. For unshielded gauges, the RMSE for U_{gh} is lower than that for U_{10m} at XBK and CAR, and higher at HKL and MAR.

When considering each gauge at each site, the anticipated decrease in the RMSE following adjustment (whether for all precipitation types or snow only) is not universal. This is illustrated in Fig. 6, which shows the RMSE results for combined SUT snow datasets before and after adjustment using U_{10m} (where possible). For single -Alter--shielded gauges, the RMSE increases with adjustment at CAR, MAR, SOD and WFJ (although the increase is small at all sites but WFJ). The decrease in RMSE is small at XBK and CCR. The differences between RMSE results using Eq. 1 and Eq. 2 for single -Alter--shielded gauges are insubstantial (< 0.005 mm). These differences are larger for the unshielded gauges as shown in Fig. 6 (as high as 0.05 mm).

3.3 Pearson correlation

~~The Pearson correlation (r) assesses the strength of linear relationships between the SUT and the references before and after adjustments using Eq. 1 and Eq. 2.~~ Similar to previous metrics, the single -Alter--shielded and unshielded r-

values are shown separately in Tables 6 and 7 respectively, and are plotted for snow in Fig. 7. In theory, the adjustment using a transfer function should strengthen the linear relationship (increase r) between the adjusted and the reference measurements by removing the non-linearity associated with wind bias.

For single -Alter--shielded gauges, unadjusted r-values for all precipitation types range from 0.83 at HKL to 0.96 at CCR. For unshielded gauges, the unadjusted r-values for all precipitation types are only slightly lower than their shielded counterparts. The unadjusted r-values for single -Alter--shielded gauges for snow are generally lower than for all precipitation types, especially at the windy sites of HKL and XBK. The unshielded, unadjusted values for snow follow similar trends as all precipitation types.

The results show that adjusting measurements for all precipitation types with either transfer function has little impact on the r-values, with greater variability in r values observed for the unshielded gauges. The impact of adjustments on snow measurements are shown in Fig. 7. Generally, the r-values for the single -Alter--shielded gauges are improved with adjustment, but the change is small (< 0.07). The largest improvements are observed for the HKL and XBK measurement datasets. With the unshielded adjustment, Unshielded gauges at most sites also show an improvement in r-value with adjustment, with the most significant increases observed for the HKL and MAR datasets.

For both all precipitation phases and snow only (Fig. 7), the differences between r-values following the application of Eq. 1 and Eq. 2 are negligible for the single -Alter--shielded gauges. For the unshielded gauges, Eq. 2 results in higher r values than Eq. 1, but the differences are very small (< 0.03).

For sites with both wind speed measurement heights, the correlations appear to be independent from the measurement height. The only exception is for the unshielded adjustment of snow measurements at MAR, where correlations based on transfer function application using U_{gh} are significantly less than those using U_{10m}. This likely results from shadowing effects on the U_{gh} data.

3.4 Nash-Sutcliffe Efficiency

The Nash-Sutcliffe Efficiency (NSE) coefficient is an alternative goodness-of-fit indicator for the agreement between the reference and the unadjusted and adjusted precipitation measurements. The measure has a higher sensitivity to bias and outliers than the Pearson Correlation coefficient and this can have both advantages and disadvantages in this assessment. It is included here to complement the other metrics used in this analysis. The NSE coefficient can vary from -∞ to 1, where 1 indicates a perfect fit between the reference and the SUT. The NSE for the single -Alter--shielded gauges and the unshielded gauges are shown in Tables 8 and 9 respectively and the values for snow are shown in Fig. 8.

Unadjusted NSE for single -Alter--shielded gauges ranges from 0.55 to 0.90 for all precipitation and from 0.08 to 0.85 for snow. For the unshielded gauges, the NSE ranges from 0.36 to 0.83 for all precipitation and -0.55 to 0.75 for snow. Similiar to the Pearson Correlation analysis, post-adjusted NSE for the single -Alter--shielded gauges changes only a small amount with larger but inconsistent changes in the post-adjusted NSE for unshielded gauges. For unshielded

gauges, some sites show a marked improvement in NSE after adjustment (HKL, MAR, and WFJ) while some sites show substantial decreases in NSE after adjustment (XBK and CAR). The unshielded NSE coefficients for XBK, both adjusted and unadjusted, are the only values less than 0 (and are clipped in Fig. 8).

From Fig. 8, the impact of the adjustment for single -Alter-shielded gauges using either Eq. 1 or Eq. 2 is relatively small. However, where the post-adjusted r-values for unshielded gauges are often similar to those for the single -Alter-shielded gauges, most sites clearly show a lower post-adjusted unshielded NSE as compared to their single -Alter counterparts. As for the relative performance of Eq. 1 vs. Eq. 2, NSE shows little difference for the single -Alter--shielded gauges with perhaps a slight advantage to Eq. 2 for adjusting the unshielded gauges. Finally, for sites with both wind speed heights, the differences in the NSE metric are more pronounced than for the Pearson Correlation, but results remain mixed and varied by site.

4 Discussion

The current application of the universal transfer functions developed in SPICE to two winter seasons of precipitation data at eight locations produced variable results depending on site location. The discussion will focus on snow to avoid the complexnfounding influence of precipitation phases, in varying proportions, on the assessment results. Based on the relative total catch results, the transfer functions tend to under-adjust snow for single -Alter--shielded gauges at the windy sites of XBK and HKL with mean 10 m wind speed during snow of 6.1 and 5.3 m s^{-1}, respectively (Table 2). The results from FOR were similar, despite relatively lower mean wind speeds of 4 m s^{-1} at 10 m. The SOD, CCR, WFJ, and MAR sites were characterized by the lowest wind speeds, with mean values at 10 m smaller than 3.2 m s^{-1}. For single -Alter--shielded gauges at these sites, the RTC following adjustment varied from 105% at SOD and CAR to 113% at MAR and WFJ. The above trends are similar to the bias results of K2017b for all precipitation phases, but more pronounced due to the focus on snow.

The adjusted RTC results showed greater variability for unshielded gauges, performing better at some sites and poorer at others relative to the performance for the single -Alter--shielded gauges. Unexpectedly, the increase in RTC after adjustment of the unshielded gauges at the windy sites (XBK and HKL) was quite effective. However, upon examination of the RMSE results, the errors associated with adjusting the unshielded gauges were generally higher compared to those associated with the single -Alter--shielded gauges. The NSE points to this as well. Since the unadjusted RTC of unshielded gauges are lower (especially at windy sites), the adjustment for unshielded measurements must be larger, so signal noise and other random measurement errors after adjustment are magnified. This also applies to very windy sites, where the required magnitude of the adjustment is necessarily larger. This propagation of errors through adjustments is discussed in greater detail by Kochendorfer et al. (2018).

The evaluation of transfer function performance is complicated by observations for which the DFAR detects a measurable amount of precipitation, but the SUT does not. A gauge measurement of zero cannot be adjusted with a transfer function. This impacts the performance metrics and cannot be ignored in the context of real-world applications of transfer functions (e.g. adjusting precipitation measurements for use in forecast validation). The limited utility of transfer functions in this regard is due to the configuration's capability to catch snow and not because of the transfer

12

function. To assess the relative impact of these types of events on the evaluations, descriptive statistics were calculated for events at XBK, HKL, and SOD.

For XBK, there were 498 30-minute snow events during which the DFAR measured an accumulation value greater than zero. The single-Alter-shielded gauge did not report precipitation during 285 of those events, which accounted for 14% of the total DFAR accumulation over both measurement seasons, and had a mean wind speed of 6 m s^{-1} (7.5 m s^{-1}) at gauge height (10 m). The number of events during which the unshielded gauge did not report precipitation was even higher: 376 events in total, accounting for 24% of the total DFAR precipitation, characterized by lower mean wind speeds (5.4 m s^{-1} at gauge height and 6.8 m s^{-1} at 10 m). For HKL, 860 of 1881 snow events were not reported by the single-Alter-shielded gauge (8% of the total DFAR accumulation), and 966 of the 1881 events were not reported by the unshielded gauge (10% of the total DFAR accumulation). One can speculate that the influence of missed reports was more significant at XBK due to the drier nature of the site and of the falling snow, which made the snow more susceptible to deflection around the gauge inlet. At SOD, where wind speed has considerably less impact on gauge catch, 413 of 1656 events reported by DFAR were not reported by the single-Alter-shielded gauge (about 6% of the total DFAR accumulation). The number of occurrences for the unshielded gauge at SOD was nearly identical to the single-Alter-shielded gauge. At windy sites, when precipitation goes undetected by a non-reference gauge, there is a negative impact on the transfer function performance metrics, but it also means that the effectiveness of the transfer function is reduced when it is applied to operational observations using those same gauge configurations. Since more shielding (e.g. double-Alter) generally means a higher catch (Watson et al., 2008; Smith, 2009; Rasmussen et al., 2012; Kochendorfer et al., 2017b), more shielding would also reduce the number of unmeasured events. Perhaps another option to increase detection of small events in cold and windy locations would be the use of optical disdrometers paired with the conventional accumulating gauges.

Even though some adjusted RTC values are closer to 100% for unshielded gauges (such as at HKL, CAR and WFJ), the RMSE was generally lower and the NSE consistently higher for single-Alter-shielded gauges, and combined with a lower frequency of missed measurements, supports the use of more shielding for solid precipitation measurements.

In general, the application of transfer functions resulted in under-adjustment at the windier sites and over-adjustment at the less windy sites; however, there is no clear relationship between the mean wind speed at a site and transfer function performance. The general performance of the transfer functions for single-Alter-shielded gauge measurements, from the perspective of RTC, likely also depends on other factors, such as crystal characteristics (Thériault et al., 2012) or aerodynamic peculiarities at the intercomparison sites affecting the representative wind speed measurements (as discussed in K2017b). Based on the present results, we found that the transfer function performance varied by site and the windy sites were under-adjusted while the less windy sites were over-adjusted, but the magnitude of the adjustment error and the specific causes of error were difficult to determine. Although beyond the scope of this work, an alternative to universal transfer functions may be to develop site-specific transfer functions. Applicability to other sites with similar conditions could be assessed using a site classification process based on climate parameters and principle components analysis, such as that shown in Pierre et al. (2019).

The differences in performance between Eq. 1 and Eq. 2 for adjusting snow measurements from single -Alter–shielded gauges were small: RTC generally varied by less than 2%, and RMSE, r and NSE were nearly identical. This was likely an artifact of the way that phase was determined in the methodology; even though C_E is a function of air temperature in Eq. 1, the data for snow are a subset of the precipitation data based on the same phase discrimination used for Eq. 2. However, the metrics for all precipitation types were also similar, which is consistent with the results in K2017b, and suggests that transfer function selection is essentially a matter of user preference. In that respect, the "simpler" Eq. 2 is less simple in that user is required to determine the phase based on temperature thresholds, while Eq. 1 requires no phase discrimination. The decision would appear to be more complicated for unshielded gauges, likely because of the increased uncertainty in both the measurement and the adjustment. Generally, the differences are still quite small but Eq. 2 shows a slight advantage with RMSE and r with a more obvious advantage in NSE. As with the single -Alter–shielded gauges, the decision likely should be based on personal preference. It would be interesting, however, to explore refining the coefficients for Eq. 2 using optical disdrometers or present weather sensors to identify dominant precipitation phase and employing such instruments when performing adjustments. It may also be worthwhile to assess the performance of Eq. 2 while using hydrometeor temperature approximation, as described in Harder & Pomeroy (2013), for phase discrimination.

Only four of eight test sites measured both U_{10m} and U_{gh}, so it is difficult to draw conclusions regarding the influence of wind speed measurement height on the performance of the transfer functions. For those four sites, the RTC using U_{gh} was closer to 100% for many of the adjustments, but RMSE and NSE values varied and r-values were nearly identical. The U_{gh} is a direct measurement of the wind speed at gauge height, and does not rely on the potentially problematic assumption that wind speed at 10 m is representative of wind speed at gauge height. This assumption relies on the estimation of surface roughness, which changes with vegetation cover, snowfall, and drifting. It also neglects the impact of increasing snow depth on the relationship between the gauge height wind speed and the 10 m height wind speed. However, depending on the distance between the SUT and the U_{gh} measurement, the U_{gh} measurement may also not be representative of the wind speed at the SUT due to interference from instruments, wind shields, and other obstructions between the wind sensor and the gauge.

As noted in K2017b, discrepancies in the various wind speed measurements (whether instrument, height, or exposure related) make it difficult to ascertain any advantage or disadvantage of using one wind speed height over the other. It is recommended to use the best wind speed data available at a given site for transfer function adjustment, but to be cognizant of the issues related to spatial representation of wind speed at the site. Additional uncertainty related to wind speed may be attributed to the variability within the 30-min mean period. Although this wasn't included in the current analysis, previous work by Wolff et al. (2015) and Nitu et al. (2018) at HKL showed that the impact of high frequency variability in the wind speed over 30-min periods on transfer functions was negligible.

5 Conclusions

The evaluation of the performance of WMO SPICE transfer functions using an independent, post-SPICE dataset showed that the performance varies by site and shield configuration and is considerably reduced when only assessing

14

their performance for snow. Generally, the application of the transfer functions to measurements from sites with higher wind speeds resulted in an under-adjustment, while producing an over-adjustment for measurements from less windy sites. This trend was not universal, which indicates that the performance is also linked to local climatic conditions affecting snowfall characteristics. On average, the transfer functions resulted in an increase in the RTC of snow measurements from single -Alter- shielded gauges (unshielded gauges) from 61% (48%) to 88% (92%), but also produced an under-adjustment as low as 54% and an over-adjustment as high as 123%. Although the RTC values imply improved transfer function performance when adjusting unshielded gauges relative to single -Alter- shielded gauges, the higher RMSE and lower correlation and NSE for unshielded adjustments suggest otherwise. Further, the unshielded gauges were shown to completely miss a larger proportion of events and accumulated precipitation relative to the DFAR than the shielded gauges, raising the critical point that precipitation that is not recorded by the gauge configuration cannot be adjusted. The differences in performance observed for Eq. 1 and Eq. 2 were small enough that the choice of transfer function should largely depend on the availability of observed precipitation phase data as well as user preference. With only four sites collecting wind speed data at both 10 m and gauge height, it was difficult to determine if the wind speed measurement height significantly affected transfer function performance. RTC was generally closer to 100% when gauge height winds were used for the adjustment, but the RMSE, NSE and correlation results were mixed. Regardless, and perhaps more importantly, users must also carefully consider potential issues with obstructions and spatial representativeness when selecting a wind speed measurement.

Ultimately, eight DFAR intercomparison sites were insufficient to address the variability in performance of the SPICE transfer functions and more intercomparison sites with a DFAR are needed in various cold region climate regimes for more thorough assessments, a key recommendation from the WMO-SPICE project (Nitu et al., 2018) For the most part, and especially at locations that experience relatively high wind speeds during snowfall events, the application of the adjustment improved the usability of the observations. This study also suggests a high degree of uncertainty in applying these adjustments in networks that geographically span many different climate regimes, and additional work is required to assess and minimize that uncertainty.

Data availability

The quality controlled 30-minute data set used in this publication is available at https://doi.org/10.1594/PANGAEA.907379, will be made available via an online repository prior to the publication of this manuscript.

Author contribution

C.S. is the lead author and completed the bulk of this analysis. A.R. was responsible for data management including coding, quality control and processing. J.K. provided advice and expertise on the use and assessment of the transfer functions. M.E. provided advice on analysis and manuscript development as well as contributing to the design and implementation of the WMO-SPICE data quality control procedures used on these data. J.K., M.E., M.W., S.B., Y.R., T.L., and C.S. were core participants of the WMO-SPICE data team and were instrumental in the collection and provision of these data.

Competing interests

The authors declare that they have no conflict of interest.

References

Buisán, S. T., Earle, M. E., Collado, J. L., Kochendorfer, J., Alastrué, J., Wolff, M., Smith, C. D., and López-
Moreno, J. I.: Assessment of snowfall accumulation underestimation by tipping bucket gauges in the Spanish
operational network, Atmos. Meas. Tech., 10, 1079-1091, https://doi.org/10.5194/amt-10-1079-2017, 2017.

Goodison, B. E.: Accuracy of Canadian snow gauge measurements, J. Appl. Meteorol., 17, 1542–1548, 1978.

Goodison, B., Louie, P., and Yang, D.: The WMO solid precipitation measurement intercomparison, World
Meteorological Organization Publications WMO/TD No. 872, 1998.

Harder, P., Pomeroy, J.: Estimating precipitation phase using a psychrometric energy balance method, Hydrol.
Process. 27, 1901-1914, doi:10.1002/hyp.9799, 2013.

Kochendorfer, J., Rasmussen, R., Wolff, M., Baker, B., Hall, M. E., Meyers, T., Landolt, S., Jachcik, A., Isaksen,
K., Brækkan, R., and Leeper, R.: The quantification and correction of wind induced precipitation measurement
errors, Hydrol. Earth Syst. Sci., 21, 1973–1989, https://doi.org/10.5194/hess-21-1973-2017, 2017a.

Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., Earle, M. E., Reverdin, A., Wong, K.,
Smith, C. D., Yang, D., Roulet, Y.-A., Buisan, S., Laine, T., Lee, G., Aceituno, J. L. C., Alastrué, J., Isaksen, K.,
Meyers, T., Brækkan, R., Landolt, S., Jachcik, A., and Poikonen, A.: Analysis of single-Alter-shielded and
unshielded measurements of mixed and solid precipitation from WMO-SPICE. Hydrology and Earth System
Sciences, 21(7), 3525-3542. https://doi.org/10.5194/hess-21-3525-2017, 2017b.

Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., Earle, M. E., Reverdin, A., Wong, K., Smith, C. D., Yang, D., Roulet, Y.-A., Meyers, T., Buisan, S., Isaksen, K., Brækkan, R., Landolt, S., and Jachcik, A.: Testing and development of transfer functions for weighing precipitation gauges in WMO-SPICE, Hydrol. Earth Syst. Sci., 22, 1437-1452, https://doi.org/10.5194/hess-22-1437-2018, 2018.

5   Nitu, R., Y.-A. Roulet, M. Wolff, M. Earle, A. Reverdin, C. Smith, J. Kochendorfer, S. Morin, R. Rasmussen, K. Wong, J. Alastrué, L. Arnold, B. Baker, S. Buisán, J.L. Collado, M. Colli, B. Collins, A. Gaydos, H.-R. Hannula, J. Hoover, P. Joe, A. Kontu, T. Laine, L. Lanza, E. Lanzinger, GW Lee, Y. Lejeune, L. Leppänen, E. Mekis, J.-M. Panel, A. Poikonen, S. Ryu, F. Sabatini, J. Theriault, D. Yang, C. Genthon, F. van den Heuvel, N. Hirasawa, H. Konishi, K. Nishimura and A. Senese: WMO Solid Precipitation Intercomparison Experiment (SPICE) (2012-2015), World
10 Meteorological Organization Instruments and Observing Methods Report No. 131, Geneva, 2018.

Nitu, R., Rasmussen, R., Roulet, Y.: WMO SPICE: Intercomparison of instruments and methods for the measurement of solid precipitation and snow on the ground, overall results and recomendations, WMO Technical Conference on meteorological and environmental instruments and methods of observations, pages 1–7, Madrid, Spain, https://www.wmo.int/pages/prog/www/IMOP/publications/IOM-
15 125_TECO_2016/Session_3/K3A_Nitu_SPICE.pdf, last accessed July 2018, 2016.

Nitu, R., Rasmussen, R., Baker, B., Lanzinger, E., Joe, P., Yang, D., Smith, C., Roulet, Y., Goodison, B., Liang, H., Sabatini, F., Kochendorfer, J., Wolff, M., Hendrikx, J., Vuerich, E., Lanza, L., Aulamo, O., and V. Vuglinsky: WMO intercomparison of instruments and methods for the measurement of solid precipitation and snow on the ground: organization of the experiment, WMO Technical Conference on meteorological and environmental
20 instruments and methods of observations, pages 16–18, Brussels, Belgium, 2012.

Pan, X., Yang, D., Li, Y., Barr, A., Helgason, W., Hayashi, M., Marsh, P., Pomeroy, J., and Janowicz, R. J.: Bias corrections of precipitation measurements across experimental sites in different ecoclimatic regions of western Canada, The Cryosphere, 10, 2347-2360, https://doi.org/10.5194/tc-10-2347-2016, 2016.

Pierre, A., Jutras, S., Smith, C., Kochendorfer, J., Fortin, V., Anctil, F.: Evaluation of catch efficiency transfer
25 functions for unshielded and single-Alter-shielded solid precipitation measurements, Journal of Atmospheric and Oceanic Technology, 36, 865-881, https://doi.org/10.1175/JTECH-D-18-0112.1, 2019.

Rasmussen, R.M., Baker, B., Kochendorfer, J., Myers, T., Landolt, S., Fisher, A., Black, J., Theriault, J., Kucera, P., Gochis, D.J., Smith, C., Nitu, R., Hall, M., Cristanelli, S., Gutmann, E.: How Well Are We Measuring Snow? The
30 NOAA/FAA/NCAR Winter Precipitation Test Bed, Bulletin of the American Meteorological Society, 93, 811-829, 2012.

Reverdin, A.: Description of the Quality Control and Event Selection Procedures used within the WMO-SPICE project, Madrid, Spain, September, 2016.

17

Scaff, L., Yang, D., Li, Y., and Mekis, E.: Inconsistency in precipitation measurements across the Alaska–Yukon border, The Cryosphere, 9, 2417-2428, https://doi.org/10.5194/tc-9-2417-2015, 2015.

Sevruk, B., Ondras, M., and Chvila, B.: The WMO precipitation measurement intercomparisons, Atmos. Res., 92, 376–380, 2009.

5

Sevruk, B., Hertig, J.-A., and Spiess, R.: The effect of a precipitation gauge orifice rim on the wind field deformation as investigated in a wind tunnel, Atmos. Environ. A-Gen., 25, 1173–1179, 1991.

Smith, C.D.: Correcting the wind bias in snowfall measurements made with a Geonor T-200B precipitation gauge
10   and Alter wind shield, CMOS Bulletin SCMO, 36(5), 162-167, 2008.

Smith, C. D.: The relationship between snowfall catch efficiency and wind speed for the Geonor T-200B precipitation gauge utilizing various wind shield configurations, Canmore, Alberta, 20–23 April 2009, 115–121, 2009.

15

Smith, C. D., Yang, D., Ross, A., and Barr, A.: The Environment and Climate Change Canada solid precipitation intercomparison data from Bratt's Lake and Caribou Creek, Saskatchewan, Earth Syst. Sci. Data, 11, 1337–1347, https://doi.org/10.5194/essd-11-1337-2019, 2019.

20   Smith, C.D, Ross, A., Kochendorfer, J., Earle, M.E., Wolff, M., Buisan, S., Roulet, Y., Laine, T.: The Post-SPICE (2015/2016 and 2016/2017) Winter Precipitation Intercomparison Data. PANGAEA, https://doi.org/10.1594/PANGAEA.907379, 2019b

25   Thériault, J. M., Rasmussen, R., Ikeda,K., Landolt, S.: Dependence of snow gauge collection efficiency on snowflake characteristics. J. Appl. Meteor. Climatol., 51, 745–762, https://doi.org/10.1175/JAMC-D-11-0116.1, 2012.

Watson, S., Smith, C., Lassi, M., Misfeldt, J: An evaluation of the effectiveness of the double-
30   Alter wind shield for increasing the catch efficiency of the Geonor T-200B precipitation gauge, Canadian Meteorological and Oceanographic Society Bulletin, 36(5), 168-175, 2008.
Wolff, M. A., Isaksen, K., Petersen-Øverleir, A., Ødemark, K., Reitan, T., and Brækkan, R.: Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study, Hydrol. Earth Syst. Sci., 19, 951–967, https://doi.org/10.5194/hess-19-951-2015, 2015.

35

Wolff, M. A., Isaksen, K., Petersen-Øverleir, A., Ødemark, K., Reitan, T., and Brækkan, R.: Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study, Hydrol. Earth Syst. Sci., 19, 951–967, https://doi.org/10.5194/hess-19-951-2015, 2015.

5   Yang, D., Metcalfe, J.R., Goodison, B.E., Mekis, E: True SNOWfall. An evaluation of the double fence intercomparison reference gauge, in 50th Eastern SNOW Conference and 61st Western SNOW Conference, Quebec City, 105-111, 1993.

Yang, D., Goodison, B.E, Metcalfe,J.R., Golubev, V.S., Bates,R., Pangburn,T., Hanson,C.L.: Accuracy of NWS 8" standard nonrecording precipitation gauge: Results and application of WMO intercomparison. J. Atmos. Oceanic
10   Technol., 15, 54–68, https://doi.org/10.1175/1520-0426, 1998.

Yang, D., Ohata, T.: A bias-corrected Siberian regional precipitation climatology, Journal of Hydrometeorology, 2(2), 122-139, https://doi.org/10.1175/1525-7541(2001)002<0122:ABCSRP>2.0.CO;2, 2001.

15

Table 1: Site name (and country), abbreviation (Abbr.), latitude, elevation above sea level, ~~and~~ mean daily air temperature, and relative total amounts of precipitation by phase (Snow/Mixed/Rain).~~.~~

Site (Country)	Abbr.	Lat.	Elev.	Mean T_{air}[*]	Relative Total Precip Amounts by Phase[**] Snow/Mixed/Rain (%)
Bratt's Lake (Canada)	XBK	50.20°	585 m	-2.1°C	43/16/40
CARE (Canada)	CAR	44.23°	251 m	-0.9 °C	28/33/39
Caribou Creek (Canada)	CCR	53.94°	519 m	-4.5 °C	85/15/1
Formigal (Spain)	FOR	42.76°	1800 m	-2.1 °C	28/49/23
Haukeliseter (Norway)	HKL	59.81°	991 m	-2.0 °C	44/35/21
Marshall (USA)	MAR	39.59°	1742 m	4.0 °C	32/45/23
Sodankylä (Finland)	SOD	67.37°	179 m	-0.4 °C	49/42/8
Weissfluhjoch (Switzerland)	WFJ	46.83°	2537 m	-4.4 °C	78/21/1

[*]daily average mean air temperature over the year from site climatology (Nitu et al., 2018)

5   [**]As measured by the DFAR for the October through April periods of 2015/2016 and 2016/2017 and calculated using temperature thresholds listed in Section 1.1.

Table 2: Relative total catch (RTC) for the single -Alter-shielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017. Results are provided for available wind speed measurement heights (10 m and gh for guage heightU_{10m} and U_{gh}) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge heightU_{gh} data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases				Snow only			
				Unadjusted RTC (%)	Eq. 1 RTC (%)	Eq. 2 RTC (%)	Mean Wind m s^{-1}	Unadjusted RTC (%)	Eq. 1 RTC (%)	Eq. 2 RTC (%)	Mean Wind m s^{-1}
XBK	G	G	10 m	63	79	77	6.4	40	69	67	6.1
		G	*gh*	63	*80*	*79*	4.9	42	*71*	*70*	4.8
CAR	P	G	10 m	86	107	105	4.2	75	112	112	4.2
		P	10 m	85	105	104	4.3	71	107	106	4.3
		C	10 m	86	106	105	4.2	73	109	109	4.2
		G	*gh*	86	*104*	*104*	3.2	75	*108*	*109*	3.3
		P	*gh*	85	*102*	*102*	3.2	71	*102*	*103*	3.3
		C	*gh*	86	*103*	*103*	3.2	73	*105*	*106*	3.3
CCR*	G	G	gh	82	110	109	2.3	83	110	110	2.2
FOR	P	P	10 m	72	82	82	3.1	46	61	61	4.0
HKL	G	G	10 m	57	74	72	5.7	43	62	61	5.3
		G	10 m	58	76	75	5.7	44	65	64	5.2
		P	10 m	55	71	69	5.8	38	55	54	5.4
		C	10 m	57	74	72	5.7	42	61	60	5.3
		G	*gh*	57	*80*	*78*	5.6	43	*70*	*69*	5.2
		G	*gh*	58	*82*	*80*	5.6	44	*73*	*72*	5.1
		P	*gh*	55	*77*	*75*	5.7	38	*62*	*61*	5.3
		C	*gh*	57	*80*	*78*	5.6	42	*69*	*67*	5.2
MAR	G	G	10 m	83	107	105	3.8	84	117	117	3.2
		G	*gh*	85	*101*	*101*	2.2	86	*113*	*113*	2.0
SOD	P	P	gh	92	104	104	1.5	91	106	105	1.3
WFJ	P	P	gh	82	111	111	2.5	79	113	113	2.8

*2016/2017 winter only

5

10

15

21

5

Table 3: Relative total catch (RTC) for the unshielded gauges (G for Geonor, P for Pluvio^2) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017. Results are provided for available wind speed measurement heights (10 m and gh for guage height U_{10m} and U_{gh}) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge height U_{gh} data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases				Snow only			
				Unadjusted RTC (%)	Eq. 1 RTC (%)	Eq. 2 RTC (%)	Mean Wind m s^{-1}	Unadjusted RTC (%)	Eq. 1 RTC (%)	Eq. 2 RTC (%)	Mean Wind m s^{-1}
XBK	G	G	10 m	53	76	72	6.4	23	70	58	6.1
		G	*gh*	53	*79*	74	4.9	23	*69*	*58*	4.8
CAR	P	G	10 m	-	-	-	4.2	-	-	-	4.2
		P	10 m	75	106	104	4.3	50	115	109	4.3
		C	10 m	-	-	-	4.2	-	-	-	4.2
		G	gh	-	-	-	3.2	-	-	-	3.3
		P	*gh*	75	*102*	*101*	3.2	48	*98*	*96*	3.3
		C	gh	-	-	-	3.2	-	-	-	3.3
CCR*	G	G	gh	72	125	121	2.3	72	123	122	2.2
FOR	P	P	10 m	-	-	-	3.1	-	-	-	4.0
HKL	G	G	10 m	47	71	68	5.7	32	63	61	5.3
		G	10 m	-	-	-	5.7	-	-	-	5.2
		P	10 m	-	-	-	5.8	-	-	-	5.4
		C	10 m	-	-	-	5.7	-	-	-	5.3
		G	*gh*	46	*85*	*77*	5.6	32	*83*	*73*	5.2
		G	gh	-	-	-	5.6	-	-	-	5.1
		P	gh	-	-	-	5.7	-	-	-	5.3
		C	gh	-	-	-	5.6	-	-	-	5.2
MAR	G	G	10 m	60	89	86	3.8	58	105	106	3.2
		G	*gh*	60	*78*	*78*	2.2	50	*79*	*80*	2.0
SOD	P	P	gh	83	104	105	1.5	83	111	111	1.3
WFJ	P	P	gh	59	100	97	2.5	53	101	97	2.8

*2016/2017 winter only

22

Table 4: Root mean square error (RMSE) for the single -Alter- shielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017 for available wind speed measurement heights (10 m and gh for guage height~~U$_{10m}$ and U$_{gh}$~~) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge height~~U$_{gh}$~~ data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases			Snow only		
				Unadjusted RMSE (mm)	Eq. 1 RMSE (mm)	Eq. 2 RMSE (mm)	Unadjusted RMSE (mm)	Eq. 1 RMSE (mm)	Eq. 2 RMSE (mm)
XBK	G	G	10 m	0.141	0.131	0.133	0.138	0.126	0.125
		G	*gh*	0.141	*0.130*	*0.131*	0.138	*0.126*	*0.125*
CAR	P	G	10 m	0.150	0.172	0.164	0.197	0.216	0.211
		P	10 m	0.149	0.184	0.172	0.149	0.166	0.161
		C	10 m	0.150	0.178	0.168	0.176	0.193	0.189
		G	*gh*	0.150	*0.157*	*0.157*	0.197	*0.203*	*0.203*
		P	*gh*	0.149	*0.162*	*0.160*	0.149	*0.143*	*0.144*
		C	*gh*	0.150	*0.160*	*0.158*	0.176	*0.177*	*0.177*
CCR*	G	G	gh	0.103	0.100	0.097	0.095	0.095	0.096
FOR	P	P	10 m	0.339	0.268	0.277	0.483	0.379	0.379
HKL	G	G	10 m	0.293	0.244	0.248	0.348	0.289	0.289
		G	10 m	0.287	0.258	0.259	0.344	0.319	0.317
		P	10 m	0.301	0.265	0.270	0.358	0.307	0.306
		C	10 m	0.293	0.256	0.259	0.349	0.305	0.305
		G	*gh*	0.293	*0.245*	*0.251*	0.348	*0.288*	*0.290*
		G	*gh*	0.287	*0.266*	*0.267*	0.344	*0.328*	*0.329*
		P	*gh*	0.301	*0.270*	*0.276*	0.358	*0.305*	*0.308*
		C	*gh*	0.293	*0.260*	*0.265*	0.350	*0.308*	*0.309*
MAR	G	G	10 m	0.285	0.235	0.242	0.132	0.125	0.128
		G	*gh*	0.273	*0.243*	*0.251*	0.144	*0.151*	*0.153*
SOD	P	P	gh	0.070	0.072	0.073	0.072	0.076	0.075
WFJ	P	P	gh	0.169	0.216	0.213	0.180	0.231	0.228

*2016/2017 winter only

Table 5: Root mean square error (RMSE) for the unshielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017 for available wind speed measurement heights (10 m and gh for guage heightU_{10m} and U_{gh}) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge heightU_{gh} data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases			Snow only		
				Unadjusted RMSE (mm)	Eq. 1 RMSE (mm)	Eq. 2 RMSE (mm)	Unadjusted RMSE (mm)	Eq. 1 RMSE (mm)	Eq. 2 RMSE (mm)
XBK	G	G	10 m	0.172	0.205	0.185	0.180	0.245	0.202
		G	*gh*	0.172	*0.204*	*0.183*	0.182	*0.236*	*0.198*
CAR	P	G	10 m	-	-	-	-	-	-
		P	10 m	0.202	0.325	0.270	0.241	0.352	0.302
		C	10 m	-	-	-	-	-	-
		G	gh	-	-	-	-	-	-
		P	*gh*	0.202	*0.277*	*0.249*	0.241	*0.301*	*0.271*
		C	gh	-	-	-	-	-	-
CCR*	G	G	gh	0.131	0.165	0.147	0.122	0.148	0.143
FOR	P	P	10 m	-	-	-	-	-	-
HKL	G	G	10 m	0.335	0.252	0.270	0.382	0.283	0.287
		G	10 m	-	-	-	-	-	-
		P	10 m	-	-	-	-	-	-
		C	10 m	-	-	-	-	-	-
		G	*gh*	0.335	*0.293*	*0.293*	0.382	*0.344*	*0.320*
		G	gh	-	-	-	-	-	-
		P	gh	-	-	-	-	-	-
		C	gh	-	-	-	-	-	-
MAR	G	G	10 m	0.447	0.347	0.373	0.216	0.135	0.142
		G	*gh*	0.411	*0.345*	*0.357*	0.284	*0.240*	*0.242*
SOD	P	P	gh	0.084	0.080	0.083	0.084	0.087	0.087
WFJ	P	P	gh	0.310	0.274	0.242	0.336	0.296	0.259

*2016/2017 winter only

Table 6: Pearson correlation (r) for the single -Alter- shielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017 for available wind speed measurement heights (10 m and gh for guage heightU_{10m} and U_{gh}) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge heightU_{gh} data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases			Snow only		
				Unadjusted r	Eq. 1 r	Eq. 2 r	Unadjusted r	Eq. 1 r	Eq. 2 r
XBK	G	G	10 m	0.89	0.89	0.89	0.67	0.71	0.71
		G	*gh*	0.89	*0.89*	*0.89*	0.67	*0.70*	*0.71*
CAR	P	G	10 m	0.96	0.95	0.95	0.83	0.86	0.86
		P	10 m	0.96	0.95	0.96	0.94	0.94	0.94
		C	10 m	0.96	0.95	0.96	0.88	0.90	0.90
		G	*gh*	0.96	*0.96*	*0.95*	0.83	*0.86*	*0.86*
		P	*gh*	0.96	*0.96*	*0.96*	0.94	*0.95*	*0.95*
		C	*gh*	0.96	*0.96*	*0.96*	0.88	*0.90*	*0.90*
CCR*	G	G	gh	0.94	0.95	0.94	0.94	0.94	0.95
FOR	P	P	10 m	0.92	0.94	0.94	0.86	0.89	0.89
HKL	G	G	10 m	0.83	0.87	0.87	0.71	0.78	0.78
		G	10 m	0.83	0.85	0.85	0.70	0.72	0.72
		P	10 m	0.84	0.87	0.86	0.73	0.78	0.78
		C	10 m	0.83	0.86	0.86	0.71	0.76	0.76
		G	*gh*	0.83	*0.87*	*0.86*	0.71	*0.78*	*0.77*
		G	*gh*	0.83	*0.85*	*0.85*	0.70	*0.72*	*0.72*
		P	*gh*	0.84	*0.87*	*0.86*	0.72	*0.78*	*0.78*
		C	*gh*	0.83	*0.86*	*0.86*	0.71	*0.76*	*0.75*
MAR	G	G	10 m	0.87	0.91	0.90	0.92	0.94	0.94
		G	*gh*	0.86	*0.89*	*0.88*	0.90	*0.91*	*0.91*
SOD	P	P	gh	0.92	0.92	0.92	0.90	0.90	0.90
WFJ	P	P	gh	0.95	0.94	0.94	0.95	0.94	0.94

*2016/2017 winter only

5

10

15

Table 7: Pearson correlation (r) for the unshielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017 for available wind speed measurement heights (10 m and gh for guage height~~U_{10m} and U_{gh}~~) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge height~~U_{gh}~~ data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases			Snow only		
				Unadjusted r	Eq. 1 r	Eq. 2 r	Unadjusted r	Eq. 1 r	Eq. 2 r
XBK	G	G	10 m	0.84	0.75	0.79	0.30	0.28	0.32
		G	*gh*	0.84	*0.75*	*0.79*	0.30	*0.30*	*0.33*
CAR	P	G	10 m	-	-	-	-	-	-
		P	10 m	0.94	0.88	0.91	0.86	0.83	0.85
		C	10 m	-	-	-	-	-	-
		G	*gh*	-	-	-	-	-	-
		P	*gh*	0.94	*0.90*	*0.92*	0.86	*0.84*	*0.86*
		C	*gh*	-	-	-	-	-	-
CCR*	G	G	gh	0.92	0.93	0.94	0.91	0.92	0.93
FOR	P	P	10 m	-	-	-	-	-	-
HKL	G	G	10 m	0.79	0.87	0.84	0.69	0.80	0.80
		G	10 m	-	-	-	-	-	-
		P	10 m	-	-	-	-	-	-
		C	10 m	-	-	-	-	-	-
		G	*gh*	0.79	*0.84*	*0.82*	0.69	*0.78*	*0.77*
		G	gh	-	-	-	-	-	-
		P	gh	-	-	-	-	-	-
		C	gh	-	-	-	-	-	-
MAR	G	G	10 m	0.70	0.79	0.76	0.85	0.93	0.92
		G	*gh*	0.72	*0.79*	*0.77*	0.65	*0.71*	*0.70*
SOD	P	P	gh	0.89	0.90	0.90	0.88	0.88	0.88
WFJ	P	P	gh	0.88	0.89	0.90	0.87	0.87	0.89

*2016/2017 winter only

Table 8: Nash-Sutcliffe Efficiency (NSE) index for the single -Alter- shielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017 for available wind speed measurement heights (10 m and gh for guage heightU_{10m} and U_{gh}) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge heightU_{gh} data are shown in red and italicized.

Site	DFAR	Gauge	Wind Height	All precipitation phases			Snow only		
				Unadjusted NSE	Eq. 1 NSE	Eq. 2 NSE	Unadjusted NSE	Eq. 1 NSE	Eq. 2 NSE
XBK	G	G	10 m	0.76	0.81	0.80	0.08	0.24	0.25
		G	*gh*	0.74	*0.79*	*0.78*	0.09	*0.25*	*0.26*
CAR	P	G	10 m	0.90	0.87	0.88	0.65	0.58	0.60
		P	10 m	0.91	0.85	0.87	0.80	0.76	0.77
		C	10 m	0.90	0.86	0.88	0.72	0.67	0.68
		G	*gh*	0.90	*0.89*	*0.89*	0.65	*0.63*	*0.63*
		P	*gh*	0.91	*0.89*	*0.89*	0.80	*0.82*	*0.82*
		C	*gh*	0.90	*0.89*	*0.89*	0.72	*0.72*	*0.72*
CCR*	G	G	gh	0.85	0.86	0.87	0.85	0.85	0.85
FOR	P	P	10 m	0.78	0.86	0.85	0.42	0.64	0.64
HKL	G	G	10 m	0.56	0.69	0.68	0.26	0.49	0.49
		G	10 m	0.57	0.65	0.65	0.27	0.37	0.38
		P	10 m	0.55	0.65	0.64	0.23	0.44	0.44
		C	10 m	0.56	0.67	0.66	0.26	0.43	0.43
		G	*gh*	0.56	*0.69*	*0.67*	0.26	*0.49*	*0.49*
		G	*gh*	0.57	*0.63*	*0.63*	0.27	*0.33*	*0.33*
		P	*gh*	0.55	*0.64*	*0.62*	0.23	*0.44*	*0.43*
		C	*gh*	0.56	*0.65*	*0.64*	0.26	*0.42*	*0.42*
MAR	G	G	10 m	0.73	0.82	0.81	0.80	0.82	0.81
		G	*gh*	0.73	*0.79*	*0.77*	0.78	*0.76*	*0.75*
SOD	P	P	gh	0.84	0.83	0.83	0.80	0.78	0.79
WFJ	P	P	gh	0.87	0.80	0.80	0.85	0.76	0.76

*2016/2017 winter only

10

15

27

Table 9: Nash-Sutcliffe Efficiency (NSE) index for the unshielded gauges (G for Geonor, P for Pluvio[2]) as compared to the DFAR reference at each site for the combined winters of 2015/216 and 2016/2017 for available wind speed measurement heights (10 m and gh for guage height~~U_{10m} and U_{gh}~~) and for Eq. 1 and Eq. 2. Metrics are separated by precipitation phase (all and snow only). Where both wind speed heights are available, the gauge height~~U_{gh}~~ data are shown in red and italicized.

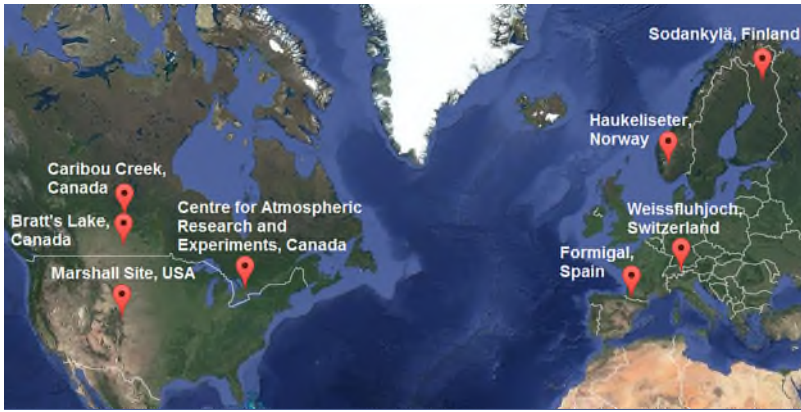| Site	DFAR	Gauge	Wind Height	All precipitation phases			Snow only		
				Unadjusted NSE	Eq. 1 NSE	Eq. 2 NSE	Unadjusted NSE	Eq. 1 NSE	Eq. 2 NSE
XBK	G	G	10 m	0.62	0.41	0.54	-0.55	-1.88	-0.97
		G	*gh*	0.58	*0.38*	*0.51*	0.30	*-1.66*	*-0.87*
CAR	P	G	10 m	-	-	-	-	-	-
		P	10 m	0.83	0.55	0.69	0.49	-0.83	0.20
		C	10 m	-	-	-	-	-	-
		G	gh	-	-	-	-	-	-
		P	*gh*	0.83	*0.67*	*0.74*	0.49	*0.21*	*0.36*
		C	gh	-	-	-	-	-	-
CCR*	G	G	gh	0.76	0.61	0.69	0.75	0.63	0.65
FOR	P	P	10 m	-	-	-	-	-	-
HKL	G	G	10 m	0.42	0.67	0.62	0.11	0.51	0.50
		G	10 m	-	-	-	-	-	-
		P	10 m	-	-	-	-	-	-
		C	10 m	-	-	-	-	-	-
		G	*gh*	0.42	*0.55*	*0.55*	0.11	*0.28*	*0.37*
		G	gh	-	-	-	-	-	-
		P	gh	-	-	-	-	-	-
		C	gh	-	-	-	-	-	-
MAR	G	G	10 m	0.36	0.62	0.56	0.45	0.79	0.76
		G	*gh*	0.41	*0.59*	*0.56*	0.15	*0.39*	*0.38*
SOD	P	P	gh	0.78	0.79	0.78	0.74	0.72	0.72
WFJ	P	P	gh	0.58	0.67	0.74	0.48	0.60	0.70

*2016/2017 winter only

28

Figure 1: The SPICE intercomparison sites used in the development and evaluation of the SPICE transfer functions (base map obtained from Google earth; Data SIO, U.S. Navy, GEBCO ©2018 Google; Image Landsat/Copernicus).
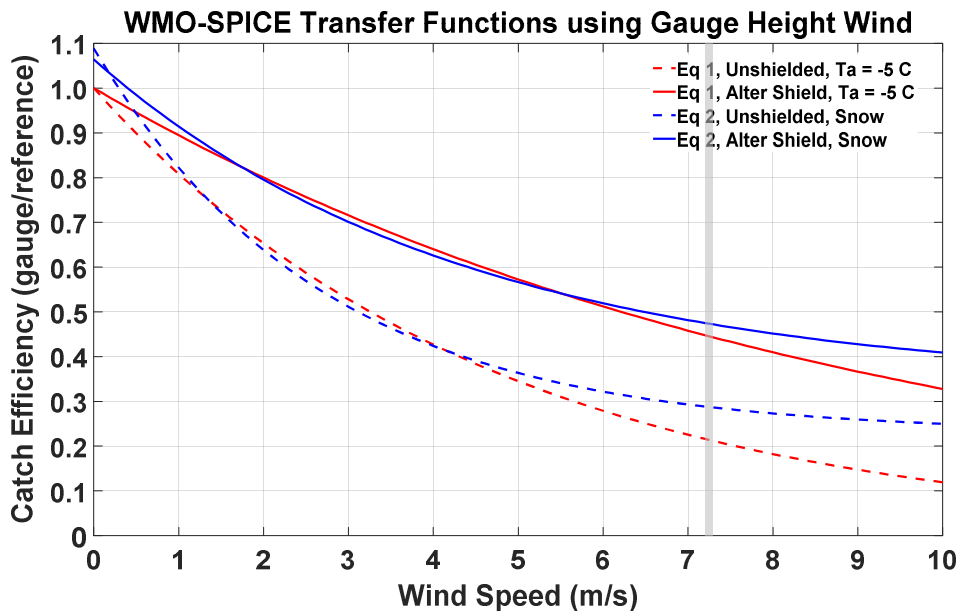

Figure 2. WMO-SPICE transfer functions from K2017b, Eq. 1 (red), Eq. 2 (blue) for unshielded (dashed), and single -Alter- shielded (solid) gauges. Eq. 1 is plotted using an air temperature of -5° C and Eq. 2 is plotted for snow. Both transfer functions are plotted for U_{gh} with the maximum wind speed threshold shown at 7.2 m s^{-1}.

Figure 3: Time series of unadjusted (solid) and adjusted (dashed) precipitation time series for single -Alter- shielded (red) and unshielded (blue) gauges as compared to the DFAR (black) during the 2015/2016 winter season at the eight SPICE sites.
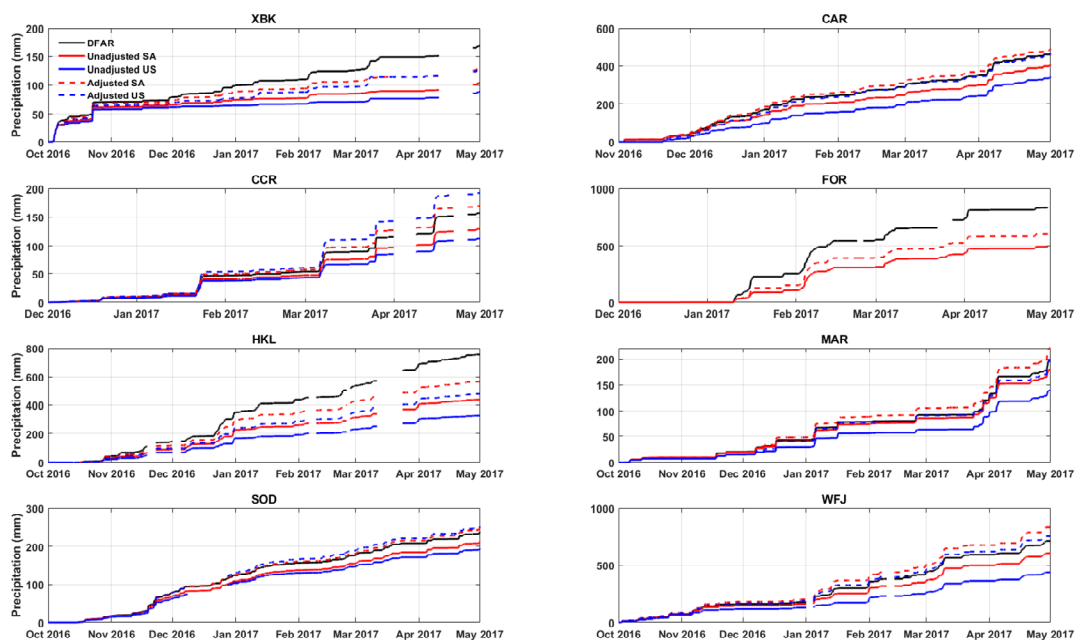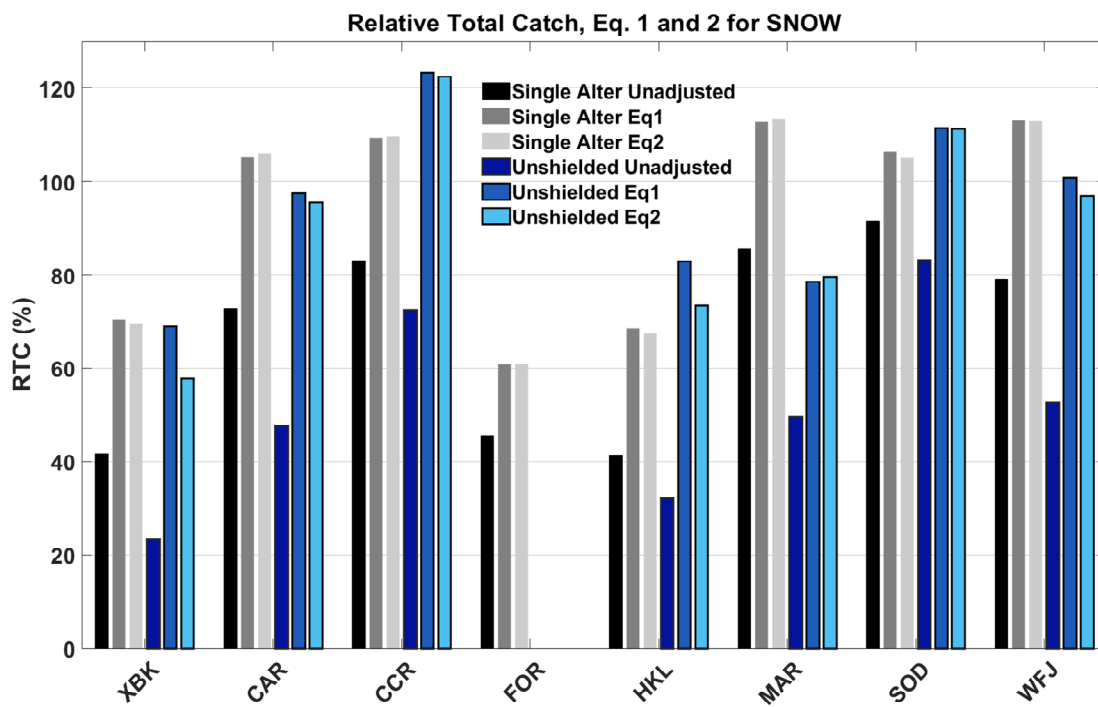
30

Figure 4: Time series of unadjusted (solid) and adjusted (dashed) precipitation time series for single -Alter- shielded (red) and unshielded (blue) gauges as compared to the DFAR (black) during the 2016/2017 winter season at the eight SPICE sites.

31

Figure 5: Relative total catch of snow (as compared to the DFAR) for ~~single~~ single -Alter and unshielded gauges for both Equations 1 and 2 at each of the eight SPICE sites, combining 2015/2016 and 2016/2017 winter seasons.
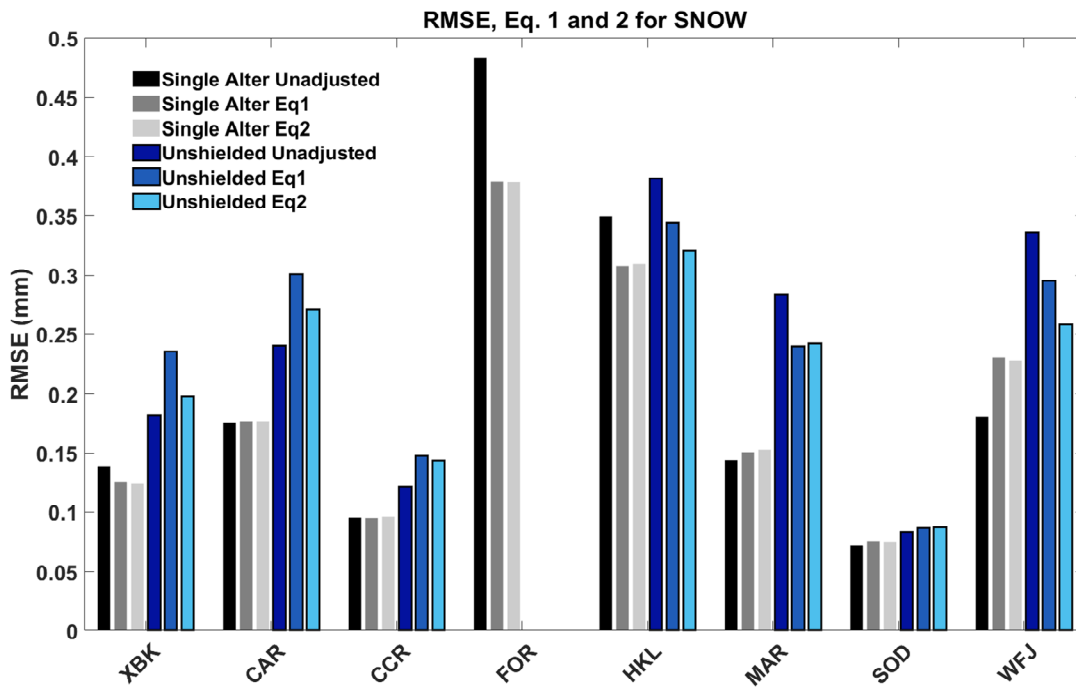
Figure 6: RMSE for snow for single -Alter and unshielded gauges for unadjusted measurements and for adjustments using Equations 1 and 2 at each of the eight SPICE sites, combining 2015/2016 and 2016/2017 winter seasons.
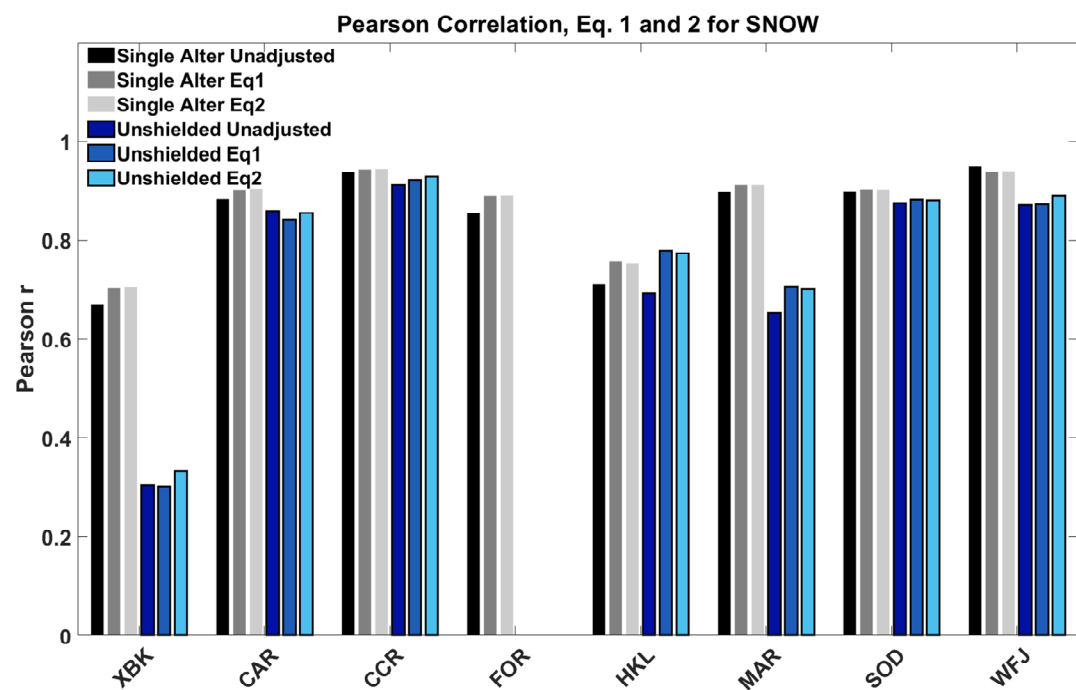
Figure 7: Pearson r for ~~single~~ single -Alter and unshielded gauges for unadjusted measurements and for adjustments using Equations 1 and 2 at each of the eight SPICE sites, combining 2015/2016 and 2016/2017 winter seasons.
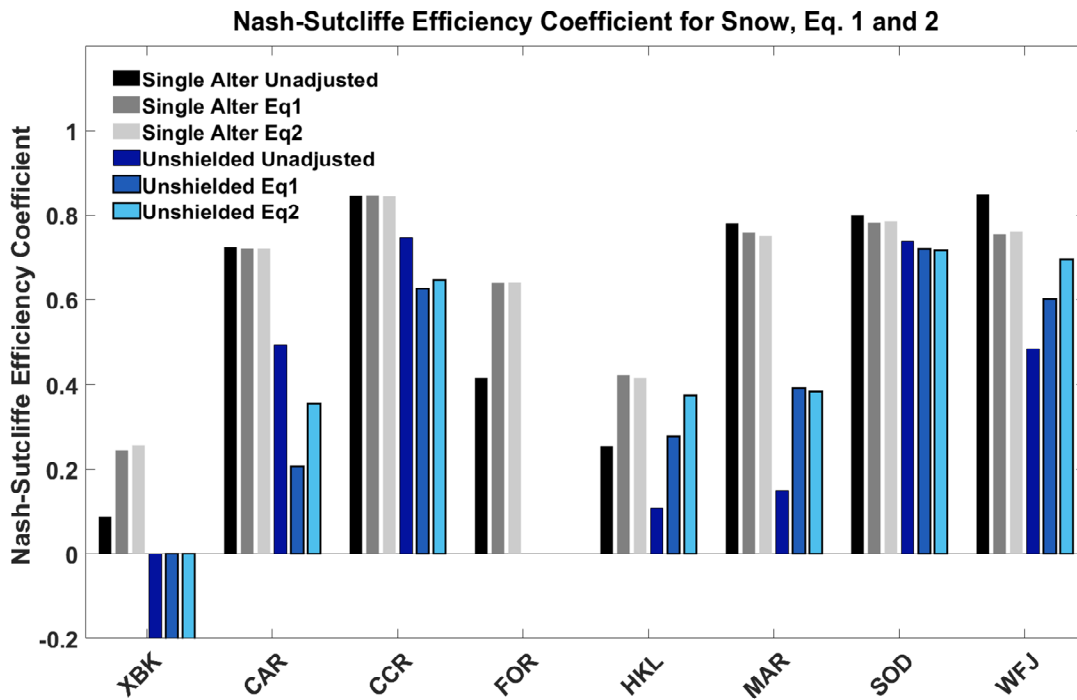
Figure 8: Nash-Sutcliffe Efficiency Coefficient for single ~~single~~ Alter and unshielded gauges for unadjusted measurements and for adjustments using Equations 1 and 2 at each of the eight SPICE sites, combining 2015/2016 and 2016/2017 winter seasons.

5