Verification of analysed and forecasted winter precipitation in complex terrain

M. Schirmer[1),2)] and B. Jamieson[1)]

1) Department of Civil Engineering, University of Calgary, Calgary, AB, Canada

2) Centre for Hydrology, University of Saskatchewan, Saskatoon, SK, Canada

Correspondence to: M. Schirmer (michael.schirmer@usask.ca)

Abstract

Numerical Weather Prediction (NWP) models are rarely verified for mountainous regions during the winter season, although avalanche forecasters and other decision makers frequently rely on NWP models. Winter precipitation from two NWP models (GEM-LAM and GEM15) and from a precipitation analysis system (CaPA) was verified at approximately 100 stations in the mountains of western Canadian and northwestern US. Ultrasonic snow depth sensors and snow pillows were used to observe daily precipitation amounts. For the first time, a detailed objective validation scheme was performed highlighting many aspects of forecast quality. Overall, the models underestimated precipitation amounts, although low precipitation categories were overestimated. The finer resolution model GEM-LAM performed best in all analysed aspects of model performance, while the precipitation analysis system performed worst. An analysis of the economic value of large precipitation categories showed that only mitigation measures with low cost/loss ratios (i.e. measures that can be performed often) will benefit from these NWP models. This means that measures with large associated costs relative to anticipated losses when the measure is not performed should not or not primarily depend on forecasted precipitation.

1 Introduction

Recently, there has been a growing interest in the question of how much snow is distributed over mountainous terrain. A better knowledge thereof will improve our understanding and forecasting of natural hazards like flooding and avalanches, which affect us today. Since snow is close to its melting point, small changes in climate will influence not only natural hazards, but also drinking water resources in snow-melt dominated watersheds.

Numerical Weather Prediction (NWP) models were recently developed with increasing spatial resolutions able to capture relevant physical processes in highly complex terrain. Thus they are potentially able to be applied for flood and avalanche forecasting, for which forecasted winter precipitation is an especially relevant output variable. Furthermore, the performance of NWP models can suggest at which resolution and with which model characteristics regional climate models need to be applied to estimate winter precipitation and thus drinking water resources in a changing climate.

In a forecast demonstration project MAP D-PHASE, Rotach et al. (2009) tested the ability of a large number of high-resolution (i.e. a few kilometres grid size) NWP models to forecast floods in the Alps during summer and fall of 2007. One important outcome was that high-resolution convection-permitting models have an additional value in short-time forecasting precipitation alerts for a large variety of potential users. In a subsequent paper Weusthoff et al. (2010) investigated in detail whether high-resolution models (2.2 – 2.8 km grid size) performed better than their driving lower-resolution counterparts (6.6 – 10 km). With the same gridded verification dataset as used for MAP D-PHASE derived by radar composite. Weusthoff et al. (2010) focused on short-time forecast of accumulated 3-h precipitation using the Swiss and German COSMO and the French ALADIN/AROME models. They concluded that higher-resolution models were better or at least equal to the low-resolution models in this experiment of high complex terrain during a six month study in the summer and fall. They also observed that modelled skill varied between months and days showing that a long verification period is needed to obtain robust results.

Verifications using such complex experimental settings regularly covered only a short period. One example was the IMPROVE-2 over the Oregon Cascades during a winter storm (two days) in December 2001 (e.g. Garvert et al., 2005). It was found that a spatial resolution of 1.3 km is needed for the MM5 model to capture observed small-scale oscillations relevant for spatial precipitation differences. Garvert et al. (2005) found that precipitation observed with rain gauges was generally overpredicted especially on the leeward side of the range. Milbrandt et al. (2008, 2010) partially corrected this leeward bias with improvements in the microphysics scheme; however, a general overprediction remained. They used a Canadian Global Environmental Multiscale (GEM) model, which was also evaluated in our study. The

GEM model was also used during the Vancouver Olympic Games 2010, which led to several publications covering this short, but well documented time period (e.g. Mailhot et al., 2012).

Colle et al. (2005) applied the MM5 model at different spatial resolutions over the steep and narrow Wasatch Mountains of northern Utah during a snow storm on 12 February 2000 recorded by the IPEX IOP3 experiment. Accurate simulations required 1.33 km grid spacing. In a comparison with rain gauges they observed an underestimation of precipitation upstream of ridges.

Small-scale orographic effects on winter precipitation were studied by Mott et al. (2014), using radar data for one heavy snowfall event in March 2011. They modelled snow accumulation at the surface on a resolution of 75 m and discussed cloud microphysical as well as particle transport processes, which are not resolvable by typical NWP systems with resolutions larger than 1 km. These described effects are included in the discussion in the present paper on the limitations of comparing point measurements to NWP models in complex terrain.

Long-term verifications over four winter seasons were performed for the WRF model over complex terrain in the Colorado headwaters by Ikeda et al. (2010). For high-resolution models (2 and 6 km) they observed modelled precipitation to be 10-15% greater compared to SNOTEL rain gauges. This discrepancy was assumed to be equivalent to the estimated undercatch of rain gauges in forest clearings with typically low wind speeds. Oppositely, coarser resolution models of 18 and 36 km underpredicted precipitation amounts by 15% and 23-31%, respectively. Thus, they concluded that global and regional climate models with a typical spatial resolution (>18 km) underestimated high elevation snow fall substantially. Since their aim was to apply WRF as a regional climate model they emphasised monthly accumulated precipitation averaged over many stations rather than verifying daily (or hourly) sums at multiple station-model pairs. Therefore, performance measures for short-period accumulated precipitation, or for certain precipitation categories, were not calculated.

Daily precipitation sums were verified for the probabilistic forecast of the COSMO limited area ensemble (10 km resolution) in Switzerland both for a winter and a summer period (Fundel et al., 2010). Only a small part of the rain gauges used in this study were located in complex terrain, while the majority were located in the lowlands of northern Switzerland. Attribute diagrams showed that after calibration of the ensemble forecast the skill increased substantially. Haiden et al. (2011) presented a verification of a nowcast system INCA for one winter and one summer month in Austria. They treated precipitation as a continuous variable and used both a classical point verification method and an object-oriented approach. They concluded that after a 6-h lead time, i.e. when the nowcast was merged into the NWP model ALADIN, the model both overestimated precipitation and lost spatial agreement with observations.

The Canadian model GEM15 with a spatial resolution of 15 km was verified for winter precipitation during one month over the area of North America (Mailhot et al., 2006). A positive bias was observed for all precipitation categories, especially the lowest category. For complex terrain they mentioned a higher bias for larger precipitation categories during a verification period between February and May. During subjective verification the model was found to have a positive bias, especially on the windward side of the mountains. The same model was applied to estimate snow water equivalent (SWE) in the Canadian Rockies by Carrera et al. (2010). SWE was underestimated by the model, while monthly precipitation accumulation was overestimated for some locations. The general underestimation is opposite to studies in flat terrain and in the summer (Mailhot et al., 2006, Bélair et al., 2009). They also included the Canadian Precipitation Analysis system (CaPA) as an additional precipitation input. CaPA combines optimally model forecast, rain gauges and radar taking the 6-h forecast of GEM15 as a first guess to account for the spatial structure (Mahfouf et al., 2007). Carrera et al. (2010) concluded that the underestimation of SWE was more pronounced using CaPA than GEM15, which confirmed the hypothesised difficulties of including snow and orographic effects in a station based precipitation analysis (Mahfouf et al., 2007). CaPA was included in the present study as well. Bellaire et al. (2011, 2013) used the GEM15 model as an input for subsequent snow cover modelling. At one single station in the western Canadian Mountains the model was verified over several years and an underestimation of winter precipitation was observed.

None of the studies presented a verification of quantitative precipitation forecast (QPF) in such a detail as it is available for summer months (e.g. Bélair et al., 2009; Weusthoff et al., 2010). This detail is necessary to address the multidimensional character of a forecast, especially when several forecast systems are compared (Murphy, 1991).

The reason for this research gap may not only lie in the lower performance of NWP models in the winter and in the mountains, but also in larger measurement errors. The regularly used rain gauges are known for an undercatch bias for solid precipitation, mainly due to aerodynamic effects (e.g. Yang et al., 1998). A known problem exists with the response time, when wet snow sticks to the inside of the gauge and may be recorded hours or days later (Serreze et al., 1999). Therefore, we attempted to verify NWP in the mountains with observations from ultrasonic snow depth measurements and snow pillows, which are commonly used for forecasting avalanches and floods, as well as for a large number of snow-related research studies. Similarly, these measurement systems are not without errors and limitations are discussed in the present paper.

The aim of this present study was to explore the question of how well deterministic NWP models perform in the winter and in the mountains. A detailed quality assessment of NWP models of different spatial resolutions (2.5 km and 15 km) and a precipitation analysis system (10 km) was performed two Canadian deterministic models in the western Canadian and northwestern US American mountains. This will help decision makers to better estimate the value of NWP models by adding this long-term objective validation to their subjective experience. Additionally, this detailed quality analysis will add to the existing knowledge of how well NWP models can serve as regional climate models in the winter and in complex terrain.

2 Data and Methods

2.1 NWP models

The Canadian weather models GEM15 (Mailhot et al., 2006) and GEM-LAM (Erfani et al., 2005) with spatial resolutions of 15 km and 2.5 km, respectively, were verified against measured precipitation. In GEM15 separate schemes for shallow convection and deep convection are implemented, which are described in more detail in Bélair et al. (2009) and Mailhot et al. (2006). In addition to the same shallow convection scheme, GEM-LAM implements a cloud microphysical scheme which was used for the experimental version of GEM-LAM applied for the Vancouver 2010 Olympic Games (Mailhot et al., 2012, Jason Milbrandt, personal communication, 13 January 2015). In brief, the two-moment Milbrandt-Yau bulk microphysics scheme (Milbrandt and Yau, 2005) parameterises cloud microphysical and precipitation processes (Mailhot et al., 2012). This scheme accounts for most clouds and precipitation processes with a small contribution from the shallow convection scheme (Jason Milbrandt, personal communication, 13 January 2015). A brief description of the Milbrandt-Yau scheme can be found in Morrison et al. (2015).

Modelled data were available for the two winters 2012/13 and 2013/14. Research on such long time series was only possible with continuously downloading relevant files on a daily basis (http://weather.gc.ca/grib/index_e.html). The download was done for a project assisting the operational avalanche forecast in Canada (Bellaire et al., 2011, 2013; Bellaire and Jamieson, 2013). Continuous time series of modelled data were obtained using two initiation times per day, 6 and 18 UTC for GEM-LAM, and 0 and 12 UTC for GEM15. The first six hours were neglected to avoid model spinup issues.

Our aim was to focus on short-term forecast of precipitation considering only forecasts of up to 18 hours. This means that even though we analyzed 24-hour precipitation sums, a decision maker would have the same quality as presented by our analysis only up to 18 hours in advance (see below). This is especially meaningful for regions without weather stations, for which past hours cannot be filled with observations. In our setup, past hours were filled with output from a previous initiation time to calculate daily precipitation sums. Daily precipitation sums were analysed since (i) shorter summation periods would emphasis on rather irrelevant timing differences between model and station (see also section 3.1), (ii) decision makers are used to this summation period, (iii) SNOTEL weather stations (see section 2.2) were quality checked prior to downloading in the daily format only. The potential decrease of quality measures considering longer forecasts is discussed in section 3.1 and 3.2. To ensure a true 24 hour forecast, possible at any arbitrary time of the day, forecasts up to 30 hours were included (after excluding initial hours to avoid spinup issues).

For only one winter (2013/14) modelled data were available for the Canadian Precipitation Analysis System (CaPA) (Mahfouf et al., 2007). This system provides 6-h precipitation accumulation based on rain gauges and radar, as well as on Canada's regional model (GEM15, recently GEM10). We tested the performance of these two NWP models, limiting the data set to the last winter, and found negligible differences in presented performance measures. Thus we concluded that results were comparable between CaPA and the NWP models although the same verification period of two complete winters was not available.

Daily accumulated precipitation was analysed, i.e. the daily new snow amount (HN) in cm and new snow water equivalent amount (HNW) in mm, both calculated for a time window from 00:00 UTC to 00:00 UTC, except for the verification using SNOTEL stations (see section 2.2). This data set was available in daily resolution and a time window from 00:00 UTC to 00:00 UTC PST was used. Daily differences between snow depths defined the daily new snow amount (HN) within this study, similarly for modelled and observed amounts. Note that this is a different definition of HN than used in Fierz et al. (2009), since this procedure includes not only the settling of the new snow, but also of the underlying snow. This definition is necessary when ultrasonic snow depth sensors are used since these measurements include settling of the whole snowpack.

For forecasted HN the snow cover model SNOWPACK (Lehning et al., 2002) was used to account for settling processes in the snowpack to match measured snow depth with ultrasonic sensors (see section 2.2). SNOWPACK was forced with forecasted air temperature, relative humidity, incoming shortwave and longwave radiation and wind, using the lowest available layer in the NWP model. SNOWPACK was continuously run for a winter season. It is worth noting that drifting was disabled in SNOWPACK. Processes like saltation, sublimation and suspension were not accounted for in the model, i.e. SNOWPACK was only used to account for settling (see also section 3.4 in which the limitation of the verification data set are discussed). Investigations with snow harps showed that the snow cover model was able to match well the observed settling of single snow fall events (Steinkogler et al., 2009). The snow harps used in their study are measurement devices which combine settlement and temperature sensors. These sensors are able to track certain snow layers and measure their settling rates and temperatures. The main limitation of this model approach to account for settling in the snowpack is that parameterisations of new snow density and of the settling were developed in the Swiss Alps with different new snow densities. Comparisons of results between HN and HNW will be discussed considering the effects of new snow densities and settling in section 3.1.

Ultrasonic snow depth measurements provide no information about rain. To match modelled results to these measurements the SNOWPACK model used a modelled air temperature threshold of -0.5 ^{o}C to distinguish between rain and snow on an hourly time step.

2.2 Verification data

Figure 1 shows the location of the used weather stations. We used 95 stations with ultrasonic snow depth sensors and 101 stations with snow pillows, all at elevations above 1500 m a.s.l, from the following sources. Snow depth sensors were used to determine HN, snow pillows to determine HNW. Many stations were equipped with both snow depth and snow pillows (Fig. 1).

- SNOTEL (short for Snow Telemetry, http://www.wcc.nrcs.usda.gov/snow/)
- Ministry of Transportation and Infrastructure, BC, Canada (https://pub-apps.th.gov.bc.ca/saw-paws/weatherstation)
- Ministry of Forests, Lands and Natural Resource Operations, BC, Canada (http://bcrfc.env.gov.bc.ca/data/asp/)
- Alberta Environment and Sustainable Resource Development, AB, Canada (http://environment.alberta.ca/apps/basins/Default.aspx, individual data request)
- Glacier National Park, BC, Canada (individual data request)
- University of Northern British Columbia, BC, Canada (Déry et al., 2010)
- Own maintained weather stations.

In complex terrain large differences between modelled grid points and weather station elevations can appear based on the rather coarse terrain implementation in weather models. Figure 2 shows differences in elevation between stations and model grid points. Especially for the coarser model GEM15, the differences between the station grid point elevations were significant. Smoothing of modelled topography generally underestimated the elevation of the weather stations.

Modelled data were corrected for elevation differences following Liston and Elder (2006) for the parameters air temperature, relative humidity and precipitation. These corrections are dependent on the months of the year. For HNW only precipitation was changed. For HN the settling routine of SNOWPACK is strongly dependent on air temperature and relative humidity, which was also adjusted following Liston and Elder (2006). Test cases showed that these corrections increased the performance of the models. The effect of the elevation corrections are discussed in section 3.5. To minimise the effect of elevation corrections, the grid point closest to the station elevation was selected in a window of four (GEM15) or nine (GEM-LAM) grid points. Test cases which included only the nearest grid points showed negligible differences. This is consistent with Ikeda et al. (2010), who used different averaging and interpolation methods to compare modelled precipitation with station data with only marginal differences.

Both ultrasonic snow depth sensors and snow pillows are prone to errors. Ultrasonic snow depth sensors typically produce noisy time series (Ryan et al., 2008). However, they concluded that snow depth sensors are usually within ±1 cm of manual observations. Snow pillows are known to be erroneous when the base of the snow cover is at melting temperature, or when snow supports shear stress (Johnson and Marks, 2004). For SNOTEL stations Serreze et al. (1999) analysed total SWE at the beginning of April and concluded that 68% of the stations are within 15% of manual observations, while a bias was not found. This is an important advantage compared to rain gauges, which are known for a systematic undercatch (see Introduction). Serreze et al. (1999) concluded that this undercatch was approximately 20% for SNOTEL stations compared to snow pillow measurements in a non-time consistent and non-space consistent manner, which complicates corrections.

We addressed known difficulties with the measurement systems. The noisy snow depth 1-h data measured by ultrasonic snow depth sensors were smoothed with a 3-h moving-average filter. The analysis period was from November until March to avoid melting conditions. Preliminary data analysis showed that the correspondence of modelled and measured data strongly deteriorated at lower elevations especially for snow pillows. The reasons for this trend in elevation can be found in the measurement systems: for snow pillows this can be explained with melting conditions at the base of the snowpack, while for snow depth sensors the signal-to-noise ratio is smaller for locations with shallow snow depth. Thus, only stations above 1500 m a.s.l. were considered. For the snow pillow stations only days with air temperatures cold enough to ensure solid precipitation were considered. A daily maximum of -0.5 °C was used as a threshold, which is consistent with the threshold used in SNOWPACK to distinguish between snow and rain. After these corrections no elevation dependency was observed. Finally, measured data were quality checked by visual inspection and obvious outlier observations were removed.

The advantage of non-biased observations makes us confident that these two independent measurement systems, snow depth sensors and snow pillows, were able to provide a reliable verification dataset for winter precipitation.

2.3 Verification methods

We followed the verification methods used by Bélair et al. (2009) for the Canadian Global and Regional (i.e. GEM15) weather models. Daily accumulated precipitation was categorised using predefined thresholds which led to multicategorical contingency tables representing the empirical joint distributions of forecast and observations. These contingency tables were subsequently constructed into 2 x 2 contingency tables (Table 1), to analyse how well the models were able to forecast precipitation greater than specific thresholds (Bélair et al., 2009). The *bias* was used to detect if the models 'overforecasted' or 'underforecasted', which means the event was forecasted more or less often than observed, respectively (Wilks, 1999, p. 241):

$$bias = \frac{a+b}{a+c} . \tag{1}$$

A *bias* of 1 indicates an unbiased forecast.

As a measure quantifying the skill of a forecast the Equitable Threat Score (*ETS*) was used (Schaefer et al., 1990):

$$ETS = \frac{a-e}{a+b+c-e} , \tag{2}$$

which uses the number of hits by chance, *e*, as a reference forecast

$$e = \frac{(a+b)(a+c)}{n} , \tag{3}$$

with $n = a + b + c + d$ being the total number of observations.

This score is widely used for precipitation verification since "no"-events are regularly more frequent than "yes"-events. The *ETS* emphasises correct "yes"-events (hits), while correct negatives (*d*, see Table 1) are not considered.

Hogan et al. (2010) stated that the term 'Equitable Threat Score' is misleading, because the *ETS* is not equitable in its original definition, which requires that all random forecasts as well as constant forecasts would always receive the score 0. In spite of its misleading name this score is used frequently for precipitation verification and will be used here to compare results to other studies.

Besides quality, Murphy (1993) identified the value of a forecast, which is the incremental economic and/or other benefit realised by decision makers through the use of the forecast. We used a procedure by Richardson (2000) and Zhu et al. (2002), who linked the economic value with the 2 x 2 contingency table. Table 2 outlines this strategy: when a decision maker applies a preventive action, this will be associated with a certain cost *C*. Oppositely, if the decision maker does not apply an action and the event occurs, the decision maker suffers of a certain loss *L*, which is the sum of the protectable L_p and unprotectable loss L_u. The expenses of a forecast $E_{forecast}$ were calculated based on the empirical frequency in the contingency table:

$$E_{\text{forecast}} = [\tilde{a}(C + L_{\text{u}}) + \tilde{b}C + \tilde{c}L], \tag{4}$$

where $\tilde{a}, \tilde{b}, \tilde{c}$ are the relative frequencies of *a, b* and *c* ($\tilde{a}$ = *a/n*, $\tilde{b}$ = *b/n*, $\tilde{c}$ = *c/n*).

These expenses of a forecast were related to the expenses of decisions E_{climate} based on climatological frequency *o* only,

$$E_{\text{climate}} = \min(C + oL_{\text{u}}, oL), \tag{5}$$

and to the expenses of a perfect forecast E_{perfect}

$$E_{\text{perfect}} = o(C + L). \tag{6}$$

The relative economic value *V* was then calculated with

$$V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}}. \tag{7}$$

It can be shown that *V* is not dependent on L_{u} since it is common to each expense, and that *V* can be rewritten as a function of the cost/loss ratio C/L_{p}:

$$V = \frac{\min\left(o, \frac{C}{L_{\text{p}}}\right) - (\tilde{a} + \tilde{b})\frac{C}{L_{\text{p}}} - \tilde{c}}{\min\left(o, \frac{C}{L_{\text{p}}}\right) - o\frac{C}{L_{\text{p}}}}. \tag{8}$$

A perfect forecast would achieve *V* = 1. If the relative economic value is positive the decision maker can expect an economic benefit from the forecast, while negative values indicate an economic loss relatively to decisions based on the climatological frequency only. It is noteworthy that decisions based on the climatologic frequency will lead to either always or never applying a preventive action. Richardson (2000) stated that the point of the maximum economic value is equal with the climatological frequency and thus is not dependent on the forecasting system. At this point the expenses for both possible decisions based on the climatological frequency (i.e. always or never applying a preventive action) are the same. Thus climatology is not helpful for decision makers at this point, which results in a maximum value for the forecast system.

To show general differences between model and observation, differences in distribution of forecasted and observed precipitation categories were analysed, as well as forecasted and observed marginal totals (i.e. the sum of precipitation for each category). Spatial differences, including dependencies with elevation or with the difference between station and model elevation were additionally analysed with the multicategorical Kuipers skill score and the mean error (bias) (Wilks, 1995, p. 249 and p. 254).

3 Results and Discussion

3.1 Quality of simulated and forecasted precipitation

To obtain an overview of general differences between forecasts and observations, the frequency of predefined precipitation categories is plotted on a logarithmic scale in Figure 3. This plot as well as the following plots show results for daily accumulated snow depth (HN) measured with ultrasonic snow depth sensors (left) and snow water equivalent (HNW) measured with snow pillows (right). A total of over 26,000 days of HN and over 15,000 days of HNW were available for verification. The most obvious differences between the two measurement systems (blue bars) is the larger number of non-precipitation events (0-0.2 cm or mm per day) for the snow depth sensors. This can be explained by the different stations selected, the different number of days, rain which was only observable by snow pillows, and the fact that HN and HNW are not directly comparable. The relationship between HN and HNW is dependent on variable densities of freshly fallen snow, and variable settling rates after deposition during 24 hours. Test cases using only stations with sensors for both HN and HNW and considering only very cold days to ensure snow fall, showed that the latter argument may be the dominant since the obvious differences remained. These differences in units imply that a precise comparison between HN and HNW for the same categories is not possible. The different distributions will also influence the presented performance measures. Because of the low number of point pairs in the larger precipitation categories (60-100 and >100 cm or mm per day), no performance measures were calculated for those categories.

The NWP models showed a similar behaviour compared to observations (Fig. 3). Both NWP models, GEM-LAM (red) and GEM15 (green), tended to underestimate all precipitation categories with the prominent exception of the lowest precipitation category (0.2-5), which was consistently observed with both measurements systems. This general underestimation, as well as the overestimation of the lowest precipitation category was more pronounced with the coarser resolution model GEM15.

This general observation was confirmed with Figure 4, which shows the amount of precipitation in each category (marginal totals) instead of the frequency of events. The finer resolution model GEM-LAM was able to reproduce moderate precipitation categories.

Similarly to Fig. 3, the lowest precipitation category (0.2-5) was overestimated and higher precipitation categories underestimated. In total the model underestimated the precipitation amounts (bars). Again, GEM15 replicated this behaviour in a more pronounced way. These results were observable for both measurement systems. The total underestimation for GEM15 was 13% for HN and 16% for HNW. This is comparable to the values reported by Ikeda et al. (2010) for the WRF model in a similar spatial resolution, but not compensating for the known undercatch of the rain gauges. GEM-LAM's underestimation was only 4% and 5%, respectively. This good correspondence demonstrates that rain gauges, which have a known undercatch of 15% assuming very low wind speeds up to 2 m/s (Yang et al., 1998), are insufficient to verify the quality of NWP models.

Since the number of days differ for the precipitation analysis system CaPA the results were not plotted in Fig. 3 and 4. The results were more comparable to GEM15 than GEM-LAM. The underestimation of higher precipitation categories were even more pronounced than by GEM15. This could indicate that observations based on rain gauges in the winter and in the mountains, which are known for undercatches, impaired the precipitation analysis system compared to its first guess, the regional NWP model (GEM10). However, there are additional explanations for the decreased performance of CaPA. The rain gauges that were used are typically not located at relevant elevations and spatial interpolation techniques do not account for elevations explicitly (Carrera et al., 2010).

While in Figs. 3 and 4 precipitation categories were defined as intervals, this was changed for the following analyses, in which precipitation amounts larger than aforementioned thresholds were considered. The lowest threshold (>0.2) can be interpreted as "precipitation" vs. "no precipitation". Figure 5 shows the *bias* of GEM15 and GEM-LAM (solid lines). The *bias* relates the number of times an event was forecasted with the number of times it was observed. A ratio of 1 indicates an unbiased forecast. Only for the lowest threshold was a positive *bias* observed, which means the models were forecasting the lowest precipitation category too often. The negative biases in larger precipitation categories indicate that models forecasted higher precipitation categories less often than observed. The values for CaPA are shown only for HNW (Fig. 5b, dashed line), since this system provides only precipitation and thus not enough input parameters to run the snow cover model SNOWPACK. Consistent with the previous analyses, a larger underprediction of precipitation was observed with the *bias* analysis compared to both the NWP models: CaPA was not able to reproduce the number of observed events especially for larger precipitation categories.

The positive *bias* in the lowest category was more pronounced if calculated only for the lowest precipitation category (0.2-5) with values for HN of 1.4 and 1.7 GEM-LAM and GEM15, respectively, and for HNW 1.4 and 1.9 (not shown). For CaPA the value was 2.0.

The underestimation of larger precipitation categories is not consistent with published results. Bélair et al. (2009) reported an overestimation of all precipitation categories for GEM15. This is consistent with Mailhot et al. (2006) who mentioned an increased overestimation in the winter and in complex terrain. Similarly, Milbrandt et al. (2008, 2010) published an overestimation during their short time experiment during a winter storm in complex terrain especially for larger precipitation categories for GEM-LAM. One explanation may be regional differences. Our study showed large differences between stations which point to the necessity to include a large number of stations in such an analysis (see section 3.3). Another explanation may be found in the different duration of the verification period. In our study a long time period of two years was used. Weusthoff et al. (2010) reported varying results from month to month and pointed to the need for long verification periods. Mailhot et al. (2006) and Bélair et al. (2009) studies included periods of several months. A third explanation is the different measurement systems used. The rain gauges are prone to undercatch winter precipitation. The consistent results of two independent measurement systems in our study point to a reliable verification dataset. Furthermore, the fact that GEM15 replicated the behaviour of GEM-LAM but in a more pronounced way points to similar structural deficits in the NWP models more than to measurement errors. Also, other NWP models mentioned in the introduction were generally overestimating winter precipitation in the mountains when compared against rain gauges. To exclude false conclusions based on a known undercatch of rain gauges we suggest a verification data set with independent measurement systems, or in the case of rain gauges a thorough analysis of wind speeds at the stations used. Within the current WMO Solid Precipitation Intercomparison Experiment (SPICE, http://www.wmo.int/pages/prog/www/IMOP/intercomparisons/SPICE/SPICE.html), such independent measurements may be developed.

Our results are consistent with Bellaire at al. (2011, 2013). Their corrected results show a general underestimation (Bellaire et al., 2013), but with an overestimation of higher precipitation categories. Sascha Bellaire related this discrepancy in personal to a timing issue, since they used 3-h accumulated precipitation (personal communication, 31 July 2014). The differences in their Fig. 1b were furthermore calculated with categorization based on the model and not the observations: given the model forecasted large precipitation and the timing did not perfectly match, the probability was high that smaller precipitation amounts were observed at the same time. After switching from 3-h to daily accumulated precipitation they observed an underestimation of higher precipitation categories as well (Sascha Bellaire, personal communication, 31 July 2014). The precipitation gauge they used for several winters was placed at an especially wind protected site with wind speeds rarely above 2 m/s, which reduced the potential undercatch. Carrera et al. (2010) also reported an underestimation of SWE using GEM15. This comparison of studies points to the general picture of overestimating precipitation in the summer and underestimating in the winter and in complex terrain. It needs to be shown if this pattern is a typical characteristic for other NWP models as well, using not only rain gauges for winter verification.

While the timing of events did not play a role in Figs. 3-5, correct timing was considered in the following quality and economic value analyses. The results for the Equitable Threat Score (*ETS*) are shown in Figure 6. Larger *ETS* values stand for a larger skill of the model. For HN (Fig. 6a) *ETS* values decreased for larger precipitation thresholds, while GEM-LAM revealed better *ETS* values for all categories than GEM15. The shape of this curve is comparable to summer precipitation shown in Bélair et al. (2009) with a maximum in the lower precipitation categories.

Comparing Fig. 6a and b, higher *ETS* values were observed for HNW especially for medium precipitation categories. This cannot be explained with differences in the data set as shown by test cases for which the data were reduced to a subset of the same stations and same days. The shift of the maximum *ETS* values to larger precipitation categories may be partly explained by the different units of the measurement systems. For our dataset in average, it can be said that 30 mm of HNW is less than 30 cm HS (including settling in a 24 hour window). The relative frequency of each category ([a+c]/n) suggests that 30 mm HNW corresponded in average with 20 cm HN. This is not sufficient to explain the differences in *ETS* values. The better *ETS* values for HNW could also point to the better ability of snow pillows to observe a daily precipitation amount. On the model side in this verification setup, the higher *ETS* values may be explainable with the direct comparison of model and observations for HNW, while for HN the snow cover model SNOWPACK was needed to account for settling processes.

SNOWPACK's settling routine was thoroughly verified and improved (Steinkogler et al., 2009), but the parameterization was done in the Swiss Alps with generally higher new snow densities than in parts of the Canadian mountains. This procedure could lead to wrong settling amounts, especially for larger precipitation categories, and could thus explain the lower quality compared to HNW. We suggest interpreting the different results between HN and HNW as a potential range of model skill, which reflects the limitations of the verification data set.

Fig. 6b also shows the results obtained by CaPA. The *ETS* was smaller compared to GEM15 for most of the precipitation categories. This suggests again that the precipitation analysis system was not able to improve on the regional NWP model, which is integrated as a first guess in CaPA.

Comparing the presented values from HNW with published values for summer precipitation in mainly flat terrain (Bélair et al., 2009, Fig. 7a), the skill of the GEM15 model decreased when applied in the winter in complex terrain. The magnitude can be compared to the decrease in skill from a short-time forecast (one day) to a medium-time forecast (three days, Bélair et al., 2009, Fig 7b). The high resolution GEM-LAM in the winter and in complex terrain yielded similar results as the GEM15 model in the summer and in mainly flat terrain. It is worth noticing that these comparisons do not account for possible improvements in model development, as well as possible differences in both the verification data sets, which certainly affects skill measures.

The effect considering a true 24 hour forecast with longer forecasts of up to 30 hours was tested for GEM-LAM. This analysis was only done for a subset of stations with hourly data (i.e. all Canadian stations, see Fig. 1). This restriction was necessary to match the summation period of model and observations (01:00 UTC to 01:00 UTC), which is dictated by the initiation time of the NWP model (18:00 UTC plus 6 excluded initial forecast hours). SNOTEL stations were only available in daily format (08:00 UTC to 08:00 UTC) and could therefore not be used without including even longer forecasts.

A decrease in quality was anticipated when including longer forecast, but ETS values were not consistently worse. Higher precipitation categories showed even slightly larger ETS values (up to 0.035 larger for HNW, not shown), while lower precipitation categories showed lower ETS values of similar magnitude. This difference is small compared to the differences between GEM-LAM and GEM15 presented in Fig. 5b, which were as large as 0.15. The same observations were found for HN. We conclude that the effect of longer forecasts was much smaller than the presented differences between models of different resolution.

3.2 Economic value analysis

The economic value for three selected precipitation categories is shown in Figure 7 dependent on different cost/loss ratio (x-axis) representing all possible mitigation measures. Decision makers need to define cost/loss ratios for their specific operation and mitigation measures. The benefit of such an analysis is that all potential users are included. The disadvantage is that values for cost and especially for losses are difficult to determine. In general it can be said that measures with low cost/loss ratios will be applied rather often, since they incur low costs compared to anticipated losses. Below we also discuss an example of a typical user group, an avalanche warning service, using an estimated cost/loss ratio.

Solid lines show economic values for GEM-LAM and dashed lines for GEM15. This value addresses the question of whether the decision maker benefits or loses from a forecast in relation to decisions based on a climatological frequency only. The solid blue line in Fig. 7a shows the economic value of the lowest category for GEM-LAM when compared to measurements HN. Positive economic value can be expected for measures with cost/loss ratios between ~16% and ~67%. For measures with other cost/loss ratios the economic value was negative, which implies the decision maker will lose if he/she relies on the forecast. It would have been economically better to rely on the climatological frequency instead. Decisions based on the climatological frequency will lead to always or never applying a measure. For negative economic values it is better to use this rather simple strategy compared to decisions which are assessed each day and are based solely on forecasted precipitation amounts.

For higher precipitation categories the economic values decreased. For large precipitation categories (>30 cm, solid red line) a benefit from the forecast can only be expected for measures below a cost/loss ratios of 40%. Especially for these large forecasted precipitation events, avalanche or flood forecasters prepare or apply measures with associated costs. If these measures have large cost/loss ratios, which means they are rather expensive compared to the anticipated loss, the small or negative economic value in Fig. 7 implies that these measure should not rely on a precipitation forecast alone. Note that the point of the maximum economic value is equal to the climatological frequency, which explains the shift towards the left with higher precipitation categories.

Comparing GEM15 (dashed lines) with GEM-LAM indicates that for all precipitation categories the finer resolution model had a larger economic value. For larger precipitation categories GEM15 will only add a small benefit to a decision maker.

In Fig. 7b the same assessment is plotted when compared to snow pillow observations (HNW). The shift in maximum values for example for the lowest precipitation category reflects the different climatologic frequency (see also Fig. 3). In general, the differences between both measurement systems replicated those for the *ETS*. A lower economic value for the lowest precipitation category and higher values for larger precipitation categories can be recognised, with the same explanations as mentioned before.

The values for CaPA were comparable to GEM15 (not shown) with a slight improvement on the range of positive cost/loss ratios, but with lower maximum relative economic values especially for larger precipitation categories.

When the values of the two larger precipitation categories in Fig. 7b were compared to summer precipitation in non-complex terrain (Bélair et al., 2009), a similar conclusion can be drawn as for the *ETS* values. The performance of the GEM15 model decreased when applied in the winter and in the mountains similar to the decrease from a one-day to a three-day forecast, while the higher resolution model GEM-LAM could compensate for this decrease.

Similarly to presented test cases for ETS values, the effect of including longer forecasts (up to 30 hours) was tested for the economic value. Both for HN and HNW a similar conclusion as for the ETS values can be drawn, with in general small differences between the originally presented values in Fig. 7 and the test cases. Similarly, an increase in value for higher precipitation categories was observed and a decrease for lower precipitation categories. Differences were small (up to 0.05 for HNW, not shown), compared to the presented differences between the models in Fig. 7 (up to 0.2).

In the following we want to give an example for a typical group using a NWP model in the winter and in complex terrain, which is an avalanche warning service with the decision to close a road and to apply avalanche control (blasting). We refer to a cost/benefit evaluation presented by Rheinberger et al. (2009) for a heavily travelled road to a ski resort in Switzerland. This road is 3.2 km long and exposed to five avalanche paths. They called the scenario without avalanche sheds or other permanent structures an organizational mitigation system (OMS), for which they assessed a cost/loss ratio of ~50% (analysing their Table 6 and dividing cost by benefit for OMS at the most likely social discount rate of 1.5%). For a large precipitation category (>30 cm or mm per day) the economic value of the GEM-LAM model at this cost/loss ratio was either strongly reduced to 0.2 compared to its maximum of 0.45 for HNW (Fig. 7b), or was already negative for HN (Fig. 7a). This implies that the precipitation forecast by a NWP gives only a small or no economic benefit to such a user. Please note that this cost/loss ratio based on the calculations by Rheinberger et al. (2009) is valid for installing and running an avalanche warning service in total and not for single mitigation measures. In practice, a precipitation forecast is regularly used to prepare more expensive mitigation measures (e.g. put workers on alert and gather additional observations, before blasting and closing a road). These preparation measures have rather lower cost/loss ratios compared to actually applying mitigation measures. For these lower cost-loss ratios NWP models showed a larger economic value for the important larger precipitation categories. This indicates that an avalanche warning service will profit especially in the preparation phase from a NWP model while the actual decision to apply the measures should then be accompanied by observations.

3.3 Spatial differences

The investigated performance measures were analysed for the spatial distribution of the stations. The only obvious spatial dependency found was for the *bias* of the lowest precipitation category (0.2-5 cm or mm). As described in Fig. 5 the *bias* for this category was positive while for all other categories it was negative. The spatial distribution of the *bias* of the lowest precipitation category is shown in Figure 8a for GEM-LAM and HN. The data show positive values mainly in the US, which is covered by SNOTEL stations. The same spatial distribution is visible with HNW and in a more pronounced manner with GEM15. There are arguments for regional differences not represented in the model or for station related dependencies. The SNOTEL stations were the only data source with 24-h data. Unknown pre-processing and quality assessments before the download may have included filtering out especially low precipitation amounts and thus explain this positive *bias*. However, the fact that GEM15 replicates this spatial pattern in a more pronounced way hints also to real spatial differences not integrated in the model. Furthermore, within the US stations in Fig. 8a there was an east/west dependency with a larger overestimation of this lowest precipitation class in the west, which rather points to model than station issues.

Biases of other precipitation categories do not show a spatial pattern (not shown). The spatial dependency of the lowest precipitation category had no effect on other performance measures such as *ETS* values for single categories and the multicategorical Kuipers skill score, for which no spatial difference could be observed. Also, no dependencies with elevation were observable.

Many studies point to differences between lee and windward side of mountain ranges of different NWP models (e.g. Mailhot et al., 2006, Milbrandt et al., 2008, 2010, Liu et al., 2011). Figure 8b shows which stations over- or underestimated precipitation amounts expressed with the mean error (for GEM-LAM and HN). An obvious pattern of the station locations is not visible. The stations were subsequently grouped in four aspect categories defined by the model topography. To account for effects of different spatial resolutions this topography was also aggregated from the 2.5 km to a 12.5 km resolution. No relevant or statistically significant differences between these groups were detected. This can be explained with the more complex structure of the terrain with changing synoptic weather patterns (compared to single mountain ranges as studied in Milbrandt et al., 2008, 2010 or as in Liu et al., 2011). Using modelled updraft or downdraft characteristics of each day as a grouping indication rather than aspect may be investigated in the future to obtain terrain induced differences of model performance. Another conclusion of the variable results between stations shown in Fig. 8b is that a large number of stations are needed to prevent site specific effects on spatial scales not included in NWP models.

3.4 Limitations of the verification data set

Both observed and modelled precipitation is believed to be less accurate in the winter and in the mountains. Observations are affected by physical processes not resolvable in a NWP model of more than a kilometre resolution. These processes include saltation, suspension and sublimation of snow close to the ground and orographically induced small-scale snowfall patterns (e.g. Mott et al., 2014). The location of weather stations is generally intended to be representative to a certain area, trying to avoid previously mentioned small-scale effects. Grünewald et al. (2014) concluded that typical index sites appear not to be representative of their surroundings. However, their study regions were mainly in high-alpine and wind affected terrain, while the typical station used in our study was a SNOTEL station in a forest clearing with low to moderate wind speeds. Thus we believe that these stations were able to provide representative point observations that should be comparable to the NWP model output. Additionally, the large number of stations used in this study added to the robustness of the presented analyses. Many decision makers use snow depth sensors and snow pillows for avalanche and flood warnings. We believe information describing how well NWP models compare to those well used measurement systems to be valid and worthwhile.

3.5 Effect of elevation corrections

Model runs with elevation corrections improved all presented model performance measures compared to non-corrected test runs. These improvements were greater for the GEM15 model, since the magnitude of elevation differences were larger compared to the finer resolution model GEM-LAM. ETS values in Fig. 6 increased due to elevation corrections by up to 0.5 for GEM15 and 0.3 for GEM-LAM (not shown). For the economic value a similar increase was observed, with increases of up to 0.1 for GEM15 and 0.03 for GEM-LAM (not shown). In comparison, the presented differences in Fig. 7 between the both models are rather large with values up to 0.2. Thus, the difference caused by elevation corrections is less than the differences between both models.

An important question was if these elevation corrections improved the performance measures mainly because they compensated for a systematic error in both models, namely the underestimation of precipitation amounts. Precipitation was generally increased by the elevation corrections, since most of the grid points were lower in elevation compared to weather stations (Fig. 2). However, there are strong indications that the elevation corrections were relevant. First, the mean error (bias) of precipitation was dependent on difference in elevation between model and station before applying corrections. As expected, underestimated precipitation was observed at underestimated model grid point elevations. Elevation corrections were partly able compensate this expected dependency. Second, for GEM-LAM enough stations were available in an interval ±100 m difference to the model grid point (see Fig. 2). For this subset, our results could be reproduced without applying corrections (not shown).

4 Conclusion

In this study a long-term objective verification of winter precipitation forecasted by NWP models in mountainous terrain was presented. To assess the quality of NWP models we used two measurement systems commonly applied to measure winter precipitation, snow depth sensors and snow pillows. Thus, we could present consistent results showing a systematic underestimation by the NWP models in the winter and in the mountains. The quality and relative economic values differed between the two measurement systems, thus giving a range of possible model performance. The better correspondence of NWP with snow pillow data could point to snow pillows being more capable to observe daily precipitation amounts compared to snow depth sensors, but this needs further investigation. We suggest including several measurement systems for future verifications of NWP models of winter precipitation to address the uncertainty of the measurement systems. A large number of stations are needed to prevent site specific effects on spatial scales not included in NWP models. The analysis showed that the 2.5 km resolution model performed better than the 15 km resolution model in all analysed aspects of model performance. General characteristics such as overestimating small and underestimating large amounts were similar between both models, but more pronounced with the 15 km resolution model. This characteristic of a general underestimation is not consistent with many other related studies using rain gauges only which have a known undercatch in the winter. The precipitation analysis system designed to increase the regional NWP model's performance with observations based on rain gauges clearly failed in the winter and in the mountains. For those applications, precipitation analysis systems may be improved by including snow depth sensors and snow pillows instead of rain gauges.

We also presented an economic value discussion of the forecasted precipitation amounts. Decision makers who are able to assess the cost/loss ratio of their mitigation measures are able to define for which of their measures the forecast will deliver a benefit compared to decisions based on a climatological frequency. For larger precipitation categories we have shown that decision makers will only benefit from the forecasts if their measures can be applied rather often due to low costs compared to high anticipated losses. For measures with other cost/loss ratios it is important that decision makers include other information in their decision process, for example snow observations or weather station measurements. Finally, the better performance of the high-resolution model implies that regional climate models need to operate on a spatial resolution on a kilometre-scale to capture relevant processes in the winter and in complex terrain.

Acknowledgements

The authors would like to thank Doug Wilson from BC Ministry of Transportation and Infrastructure, Catherine Brown from Glacier National Park, BC, Stephen Déry from the University of Northern British Columbia, John Pomeroy from the University of Saskatchewan and many others for their help with providing weather station data. We are grateful to Curtis Pawliuk from Valemount Area Recreation Development Association, Alexandre Langlois and his team from the University of Sherbrooke, Kerry MacDonald from Marmot Basin Ski Resort , William Golley from Northwest Avalanche Solutions and Bradford White from Banff National Park and Mike Smith for their support with building weather stations. We are also very grateful to Erik Kulyk who helped us with assessing NWP model data and creating input for the snow cover model. For their support of this research we thank the Natural Sciences and Engineering Research Council of Canada, Canadian Avalanche Centre, TECTERRA, HeliCat Canada, Canadian Avalanche Association, Canadian Avalanche Foundation, Parks Canada, Mike Wiegele Helicopter Skiing, Canada West Ski Areas Association, Backcountry Lodges of BC Association, Association of Canadian Mountain Guides, Teck Mining Company, Canadian Ski Guide Association, Backcountry Access and the BC Ministry of Transportation and Infrastructure Avalanche and Weather Programs. For the interesting discussion we would like to thank the ASARC team at the University of Calgary, Vincent Vionnet from Centre National de Recherches Météorologiques in France, Stéphane Bélair and Jason Milbrandt from the Canadian Meteorological Centre and Sascha Bellaire from the University of Innsbruck. Many thanks to Simon Horton and Shane Haladuick for proofreading. We also would like to thank Richard Essery and one anonymous reviewer for their very valuable comments, which helped to improve this manuscript.

References

Bélair, S.; Roch, M.; Leduc, A.-M.; Vaillancourt, P. A.; Laroche, S. and Mailhot, J.: Medium-range quantitative precipitation forecasts from Canada's new 33-km deterministic global operational system, Weather and Forecasting, 24, 690-708, 2009.

Bellaire, S.; Jamieson, J. B. and Fierz, C.: Forcing the snow-cover model SNOWPACK with forecasted weather data, The Cryosphere, 5, 1115-1125, 2011.

Bellaire, S.; Jamieson, J. B. and Fierz, C.: Corrigendum to "Forcing the snow-cover model SNOWPACK with forecasted weather data" published in The Cryosphere, 5, 1115–1125, 2011, The Cryosphere, 7, 511-513, 2013.

Bellaire, S. and Jamieson, B.: Forecasting the formation of critical snow layers using a coupled snow cover and weather model, Cold Regions Science and Technology, 94, 37 - 44, 2013.

Carrera, M. L.; Bélair, S.; Fortin, V.; Bilodeau, B.; Charpentier, D. and Doré, I.: Evaluation of snowpack simulations over the Canadian Rockies with an experimental hydrometeorological modeling system, Journal of Hydrometeorology, 11, 1123-1140, 2010.

Colle, B. A.; Wolfe, J. B.; Steenburgh, W. J.; Kingsmill, D. E.; Cox, J. A. and Shafer, J. C.: High-resolution simulations and microphysical validation of an orographic precipitation event over the Wasatch Mountains during IPEX IOP3, Monthly Weather Review, 133, 2947-2971, 2005.

Déry, S. J.; Clifton, A.; MacLeod, S. and Beedle, M. J.: Blowing Snow Fluxes in the Cariboo Mountains of British Columbia, Canada, Arctic, Antarctic, and Alpine Research, 42, 188-197, 2010.

Ebert, E. E: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications, 15, 51-64, 2008.

Erfani, A.; Mailhot, J.; Gravel, S.; Desgagné, M.; King, P.; Sills, D.; McLennan, N. and Jacob, D.: The high resolution limited area version of the Global Environmental Multiscale model (GEM-LAM) and its potential operational applications, Preprints, 11th Conf. on Mesoscale Processes, Albuquerque, NM, Amer. Meteor Soc. M, 2005.

Fundel, F.; Walser, A.; Liniger, M. A.; Frei, C. and Appenzeller, C.: Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts, Monthly Weather Review, 138, 176-189, 2010.

Garvert, M. F.; Colle, B. A. and Mass, C. F: The 13-14 December 2001 IMPROVE-2 event. Part I: Synoptic and mesoscale evolution and comparison with a mesoscale model simulation, Journal of the Atmospheric Sciences, 62, 3474-3492, 2005.

Grünewald, T. and Lehning, M.: Are flat-field snow depth measurements representative? A comparison of selected index sites with areal snow depth measurements at the small catchment scale, Hydrological Processes, n/a-n/a, 2014.

Haiden, T.; Kann, A.; Wittmann, C.; Pistotnik, G.; Bica, B. and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region, Weather and Forecasting, 26, 166 - 183, 2011.

Hogan, R. J.; Ferro, C. A.; Jolliffe, I. T. and Stephenson, D. B.: Equitability revisited: Why the "equitable threat score" is not equitable, Weather and Forecasting, 25, 710-726, 2010.

Ikeda, K.; Rasmussen, R.; Liu, C.; Gochis, D.; Yates, D.; Chen, F.; Tewari, M.; Barlage, M.; Dudhia, J.; Miller, K.; Arsenault, K.; Grubišić, V.; Thompson, G. and Guttman, E.: Simulation of seasonal snowfall over Colorado , Atmospheric Research , 97, 462 - 477, 2010.

Johnson, J. B. and Marks, D.: The detection and correction of snow water equivalent pressure sensor errors, Hydrological Processes, 18, 3513-3525, 2004.

Lehning, M.; Bartelt, P.; Brown, B. and Fierz, C.: A physical SNOWPACK model for the Swiss avalanche warning. Part III: meteorological forcing, thin layer formation and evaluation, Cold Regions Science and Technology, 35, 169-184, 2002.

Liston, G. E. and Elder, K.: A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet), Journal of Hydrometeorology, 7, 217 - 234, 2006.

Liu, C.; Ikeda, K.; Thompson, G.; Rasmussen, R. and Dudhia, J.: High-resolution simulations of wintertime precipitation in the Colorado Headwaters region: Sensitivity to physics parameterizations, Monthly Weather Review, 139, 3533-3553, 2011.

Mahfouf, J.-F.; Brasnett, B. and Gagnon, S.: A Canadian precipitation analysis (CaPA) project: Description and preliminary results, Atmosphere-ocean, 45, 1-17, 2007.

Mailhot, J.; Bélair, S.; Lefaivre, L.; Bilodeau, B.; Desgagné, M.; Girard, C.; Glazer, A.; Leduc, A.-M.; Méthot, A.; Patoine, A. and others: The 15-km version of the Canadian regional forecast system, Atmosphere-Ocean, 44, 133-149, 2006.

Mailhot, J.; Milbrandt, J.; Giguère, A.; McTaggart-Cowan, R.; Erfani, A.; Denis, B.; Glazer, A. and Vallée, M.: An experimental high-resolution forecast system during the Vancouver 2010 Winter Olympic and Paralympic Games, Pure and Applied Geophysics, 1-21, 2012.

Milbrandt, J. and Yau, M.: A multimoment bulk microphysics parameterization. Part II: A proposed three-moment closure and scheme description. Journal of the Atmospheric Sciences, 62, 3065-3081, 2005.

Milbrandt, J.; Yau, M.; Mailhot, J. and Bélair, S.: Simulation of an orographic precipitation event during IMPROVE-2. Part I: Evaluation of the control run using a triple-moment bulk microphysics scheme, Monthly Weather Review, 136, 3873-3893, 2008.

Milbrandt, J.; Yau, M.; Mailhot, J.; Bélair, S. and McTaggart-Cowan, R.: Simulation of an orographic precipitation event during IMPROVE-2. Part II: Sensitivity to the number of moments in the bulk microphysics scheme, Monthly Weather Review, 138, 625-642, 2010.

Morrison, H.; Milbrandt, J. A.; Bryan, G. H.; Ikeda, K.; Tessendorf, S. A. and Thompson, G.: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part 2: Case study comparisons with observations and other schemes. Journal of the Atmospheric Sciences, 72, 287-311, 2015.

Mott, R.; Scipión, D.; Schneebeli, M.; Dawes, N.; Berne, A. and Lehning, M: Orographic effects on snow deposition patterns in mountainous terrain, Journal of Geophysical Research: Atmospheres, 119, 1419-1439, 2014.

Murphy, A.: Forecast verification: its complexity and dimensionality, Monthly Weather Review, 119, 1590-1601, 1991.

Murphy, A.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather and Forecasting, 8, 281-293, 1993.

Rheinberger, C. M.; Bründl, M. and Rhyner, J.: Dealing with the White Death: Avalanche Risk Management for Traffic Routes, Risk Analysis, 29, 76-94, 2009.

Richardson, D. S.: Skill and relative economic value of the ECMWF ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, John Wiley and Sons, Ltd, 126, 649-667, 2000.

Rotach, M. W.; Ambrosetti, P.; Appenzeller, C.; Arpagaus, M.; Fontannaz, L.; Fundel, F.; Germann, U.; Hering, A.; Liniger, M. A.; Stoll, M. and others: MAP D-PHASE: Real-time demonstration of weather forecast quality in the Alpine region, Bulletin of the American Meteorological Society, 90, 1321-1336, 2009.

Ryan, W. A.; Doesken, N. J. and Fassnacht, S. R.: Evaluation of ultrasonic snow depth sensors for US snow measurements, Journal of Atmospheric and Oceanic Technology, 25, 667-684, 2008.

Schaefer, J. T.: The critical success index as an indicator of warning skill, Weather and Forecasting, 5, 570-575, 1990.

Serreze, M. C.; Clark, M. P.; Armstrong, R. L.; McGinnis, D. A. and Pulwarty, R. S.: Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data, Water Resources Research, 35, 2145-2160, 1999.

Steinkogler, W; Fierz, C.; Lehning, M. and Obleitner, F. Systematic assessment of new snow settlement in SNOWPACK. In: J. Schweizer and A. van Herwijnen (Editors), Proceedings ISSW 2009, International Snow Science Workshop, Davos, Switzerland, 27 September – 2 October 2009, 132-135, 2009.

Weusthoff, T.; Ament, F.; Arpagaus, M. and Rotach, M. W.: Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE, Monthly Weather Review, 138, 3418-3433, 2010.

Wilks, D.: Statistical Methods in the Atmospheric Sciences, Academic Press, 467, 1995.

Yang, D.; Goodison, B. E.; Metcalfe, J. R.; Golubev, V. S.; Bates, R.; Pangburn, T. and Hanson, C. L.: Accuracy of NWS 8" standard nonrecording precipitation gauge: Results and application of WMO intercomparison Journal of Atmospheric and Oceanic Technology, 15, 54-68, 1998.

Zhu, Y.; Toth, Z.; Wobus, R.; Richardson, D. and Mylne, K.: The economic value of ensemble-based weather forecasts, Bulletin of the American Meteorological Society, 83, 73-83, 2002.

Table 1. Example of a 2 x 2 contingency table.

		Observed	
		Yes	No
Forecasted	Yes	a (hits)	b (false alarms)
	No	c (misses)	d (correct negatives)

Table 2. 2 x 2 contingency table for a cost/loss analysis. *C* stands for the costs of a user takes preventive action, while *L* stands for the loss if the event occurs and elements at risk are not protected. *L* is a sum of L_p, the loss which can be protected against and L_u, the unprotectable loss.

		Observed	
		Yes	No
Forecasted	Yes	Mitigated Loss ($C + L_u$)	Cost (C)
	No	Loss ($L = L_p + L_u$)	No costs

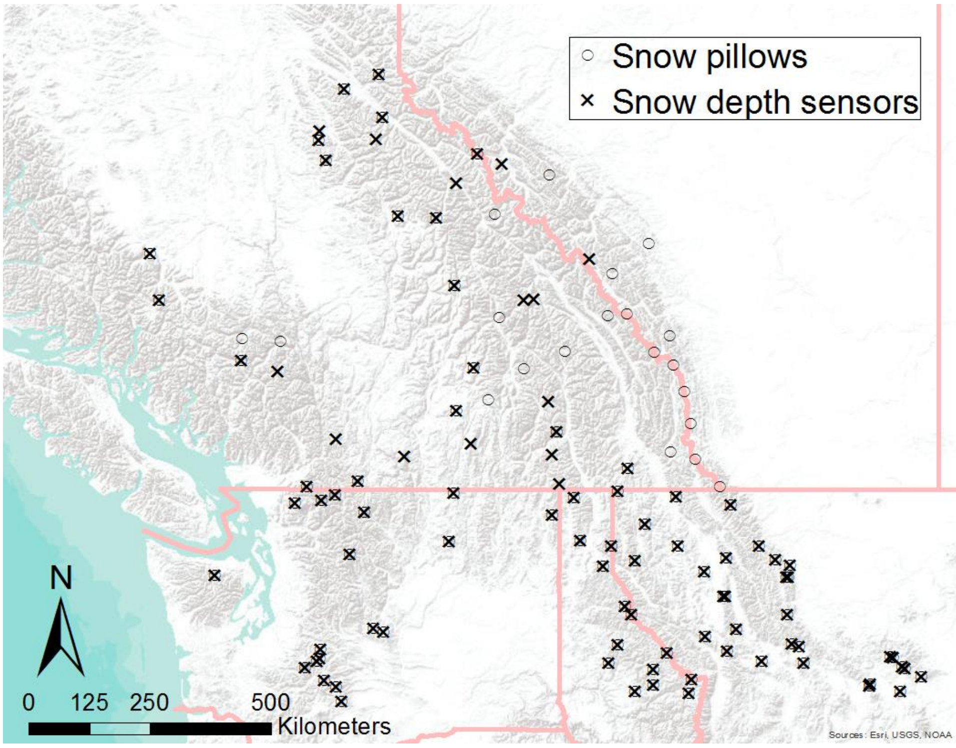

Figure 1: Locations of weather stations.

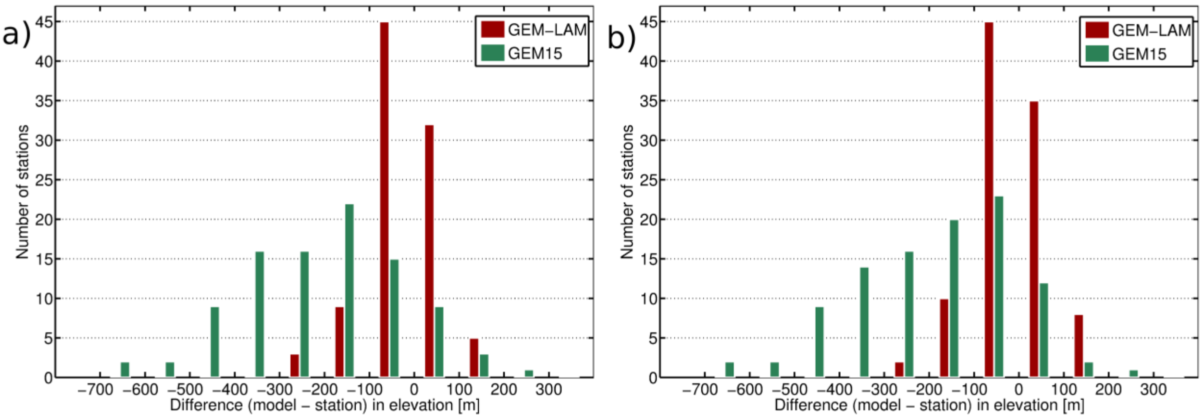

Figure 2: Differences in model and station elevation for stations with a) snow depth sensors (HN) and b) snow pillows (HNW).

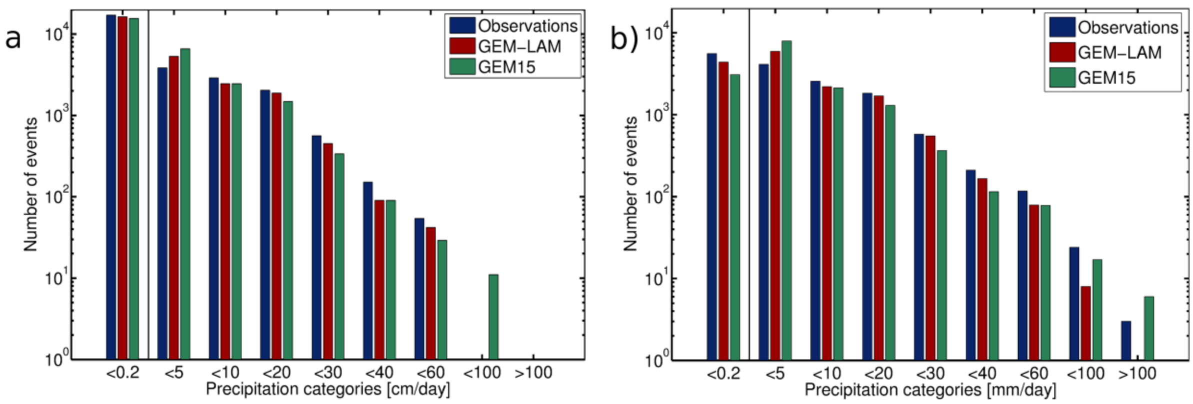

Figure 3: Frequency of daily precipitation amounts for models and observations from a) snow depth sensors (HN) and b) snow pillows (HNW). The y-axes are on a logarithmic scale. The category '<0.2' is called the non-precipitation category and '<5' is called lowest precipitation category. Categories are defined as intervals (e.g. <20 means ≥10 and <20).

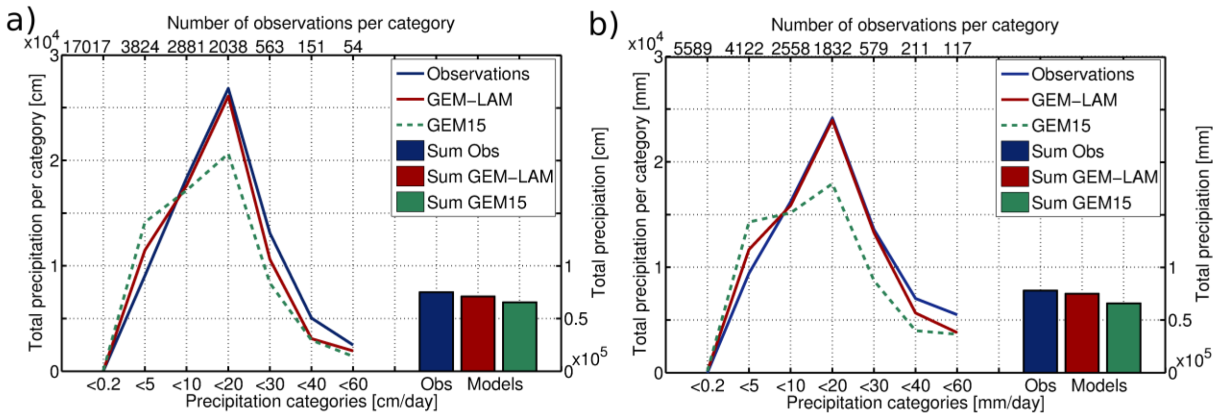

Figure 4: Sum of precipitation in each category (lines, left y-axis) and in total (bars, right y-axis) for models and observations from a) snow depth sensors (HN) and b) snow pillows (HNW). The upper x-axis shows the number of observations per category. Categories are defined as intervals (e.g. <20 means ≥10 and <20)

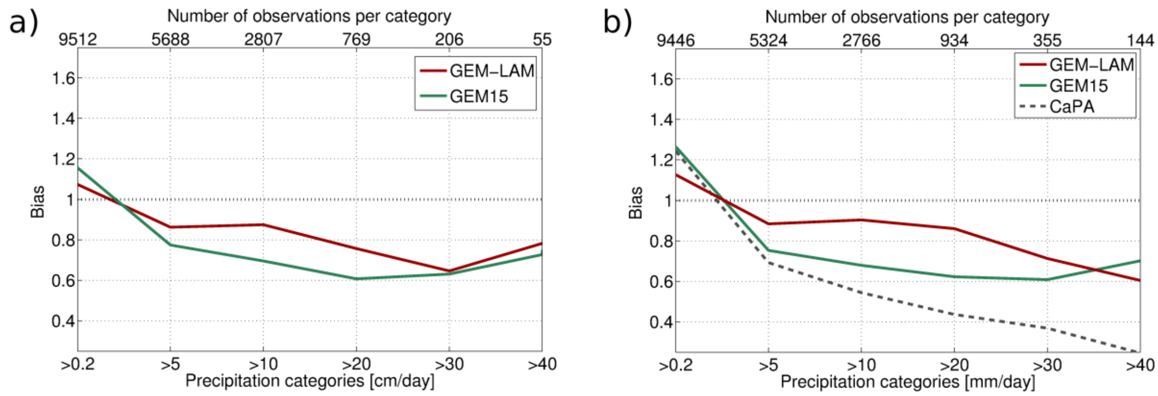

Figure 5: Modelled bias of each threshold category compared against a) snow depth sensors (HN) and b) snow pillows (HNW). The CaPA model only includes one winter of verification with approximately half of the number of observations in each category.

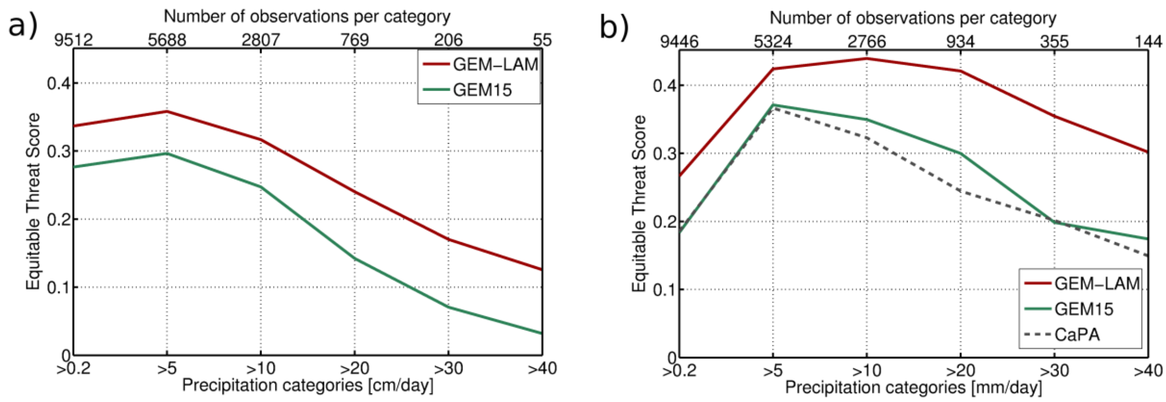

Figure 6: Equitable Threat Score (ETS) of each threshold category compared against a) snow depth sensors (HN) and b) snow pillows (HNW). The CaPA model only includes one winter of verification with approximately half of the number of observations in each category. Larger values imply better quality.

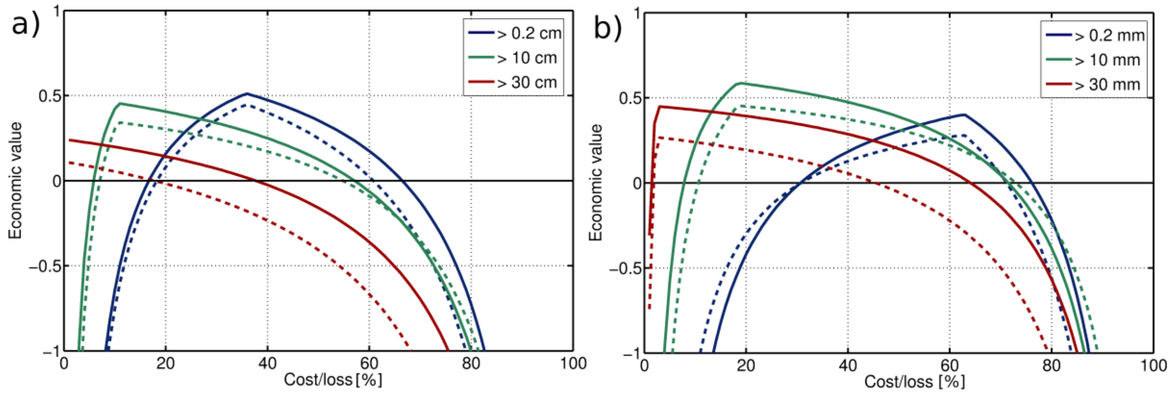

Figure 7: Economic value for three selected precipitation categories for GEM-LAM (solid lines) and GEM15 (dashed lines) compared against a) snow depth sensors (HN) and b) snow pillows (HNW).

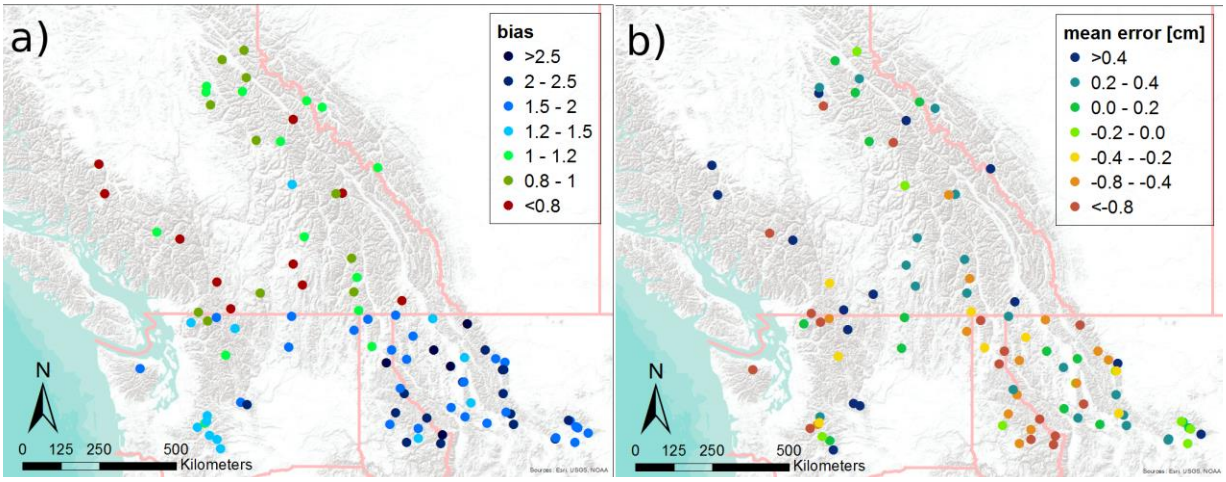

Figure 8: Spatial distribution of a) the bias of the lowest precipitation category (0.2-5 cm/day) and b) the mean error (cm) for GEM-LAM compared against snow depth sensors (HN).