https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

Evaluating methods for reconstructing large gaps in historic snow depth time series

Johannes Aschauer[1] and Christoph Marty[1]

[1]WSL Institute for Snow and Avalanche Research SLF

Correspondence: Johannes Aschauer (johannes.aschauer@slf.ch)

Abstract. Historic measurements are often temporally incomplete and may contain longer periods of missing data whereas climatological analyses require continuous measurement records. This is also valid for historic manual snow depth (HS) measurement time series, where even whole winters can be missing in a station record and suitable methods have to be found to reconstruct the missing data. Daily in-situ HS data from 126 nivo-meteorological stations in Switzerland in an altitudinal range of 230 to 2536 m above sea level is used to compare six different methods for reconstructing long gaps in manual HS time series by performing a "leave-one-winter-out" cross-validation in 21 winters at 33 evaluation stations. Synthetic gaps of one winter length are filled with bias corrected data from the best correlated neighboring station (BSC), inverse distance weighted (IDW) spatial interpolation, a weighted normal ratio (WNR) method, Elastic Net (ENET) regression, Random Forest (RF) regression and a temperature index snow model (SM). Methods that use neighboring station data are tested in two station networks with different density. The ENET, RF, SM and WNR methods are able to reconstruct missing data with a coefficient of determination (r^2) above 0.8 regardless of the two station networks used. Median RMSE in the filled winters is below 5 cm for all methods. The two annual climate indicators, average snow depth in a winter (HSavg) and maximum snow depth in a winter (HSmax), can be well reproduced by ENET, RF, SM and WNR with r^2 above 0.85 in both station networks. For the inter-station approaches, scores for the number of snow days with HS$\geq$1 cm (dHS1) are clearly weaker and except for BCS positively biased with RMSE of 18-33 days. SM reveals the best performance with r^2 of 0.93 and RMSE of 15 days for dHS1. Snow depth seems to be a relatively good-natured parameter when it comes to gap filling of HS data with neighboring stations in a climatological use case. However, when station networks get sparse and if the focus is set on dHS1, temperature index snow models can serve as a suitable alternative to classic inter-station gap filling approaches.

1 Introduction

Climatological analyses require continuous measurement series of meteorological data. Unluckily, historical measurement series are prone to contain periods of missing data. Longer data gaps can for example originate from temporally abandoning a measurement site, not properly reported measurements or archiving errors. Therefore, periods of missing data ideally need to be interpolated prior to execution of any analysis. This is also valid for manual snow depth (HS) measurement time series. For example, many instances of a whole winter of missing data are present in the manual station HS data records in Switzerland. On the other hand, long-term continuous records of HS are for example necessary to perform climatological trend analyses

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

(e.g. Matiu et al., 2021), to verify modeling studies (e.g. Olefs et al., 2020) or to calculate return levels of extreme events for constructional guidelines (e.g. Marty and Blanchet, 2012).

A number of studies have evaluated and compared methods for reconstructing missing data mostly for the two variables temperature and precipitation (e.g. Kanda et al., 2018; Woldesenbet et al., 2017; Yozgatligil et al., 2013; Kemp et al., 1983). For longer gaps, usually inter-station approaches are used where missing data of one station is imputed with the help of one or more neighboring stations (Massetti, 2014). For this purpose, most often multiple regressions, weighted averages or ratios of average values between the neighboring station and the station to be filled are used (Woldesenbet et al., 2017; Tardivo and Berti, 2012; Auer et al., 2007). More recently, also machine learning approaches have been used to estimate missing values (Kim and Pachepsky, 2010; Kashani and Dinpashoh, 2012).

Snow depth is the result of an interplay between temperature and precipitation as well as the radiation driven energy budget. Therefore, it is unclear if the methods developed for the reconstruction of other meteorological parameters are also easily applicable for snow depth time series. Additionally, for inter-station approaches there might be the problem of different relationships during accumulation and ablation phase between stations which could hinder such approaches (Bales et al., 2018). This might be especially true for stations at different elevations. Inter-station approaches are limited by the fact that a suitable set of reference stations needs to be available. Additionally, different predominant macro-scale weather patterns from one winter to the other can lead to the violation of the assumption that relationships between stations are stationary. If other meteorological parameters have been continuously measured in the period of missing HS at the target station, HS can also be derived from these parameters with snow models. For the climatological use case, temperature-index models which have been used in snow climatological impact studies (e.g. Marke et al., 2018; Notaro et al., 2011) seem to be most appropriate for this task as they only need daily precipitation and mean temperature as input variables.

Reconstruction of HS data has been done by several studies (e.g. Brown, 1996; Brown et al., 2003; Witmer, 1984; Falarz, 2002; Avanzi et al., 2020). Some of the studies focus on shorter gaps in hourly automatic measured snow data (Avanzi et al., 2020) while other studies focus on monthly means and employ very simple statistical models based on temperature only (Hughes and Robinson, 1993; Brown et al., 1995). For daily data, weighted averages of HS data from neighboring stations are employed (Matiu et al., 2021). Schöner and Koch (2016) use spatial averages and a temperature-index model to reconstruct missing daily HS data in a project of the Austrian meteorological service. However, except for Witmer (1984) who compare spatial interpolation methods for short gaps, no general comparison of different methods for reconstructing long gaps in daily HS time series exists to our knowledge. It remains unclear which methods are most appropriate for climatological analyses because the existing methods from different studies are not easily comparable and also only applicable for specific setups. For climatological analyses covering snow, most often annual or seasonal snow climate indicators are used to evaluate trends and changes in the snow cover rather than the daily values (e.g. Marty, 2008; Beniston, 2012; Buchmann et al., 2021; Marke et al., 2018; Olefs et al., 2020). These snow climate indicators are derived from daily data such as for example mean snow depth or duration of the snow cover. However, none of such studies evaluate the influence of missing data and gap filling procedures on these snow climate indicators.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

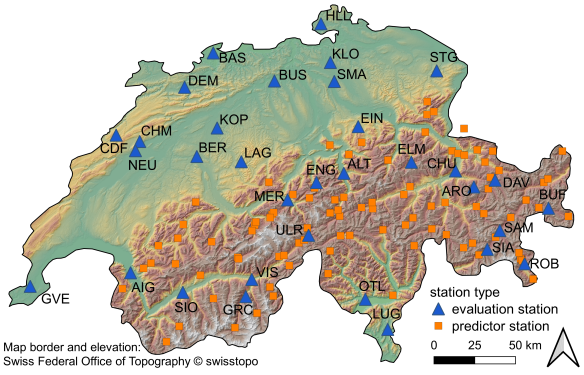

Figure 1. Location of evaluation stations (blue triangles) and predictor stations (orange squares) for the cross-validation study. The background color resembles elevation.

With this study, we perform a quantitative comparison of different methods for reconstructing typical year-long gaps in manual daily HS time series with focus on climatological analyses and the ability to reproduce important annual snow climate indicators. We compare different spatial interpolation methods as well as a simple snow model by imputing synthetic gaps in a "leave-one-winter-out" cross-validation study. The remainder of the paper is structured as follows: Used data and methods are described in Section 2, results are presented and discussed in Section 3 and concluding remarks are given in Section 4.

2 Data and methods

We use daily manual snow depth, mean temperature and sum of precipitation data from 126 nivo-meteorological stations in Switzerland. The majority (93) of the stations are primarily measuring snow related variables and not necessarily temperature and precipitation. The stations are either operated by the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) or by the WSL Institute for Snow and Avalanche Research SLF (SLF) and data is provided by these two institutions. The data covers 21 hydrological years in the period between 1999 until 2020. A hydrological year is defined as the period from September until end of August. The snow depth is measured manually between 7 and 8 a.m. local time each morning from a fixed snow stake an has the date stamp of the day of measurement. The daily sum of precipitation data is covering the period 7 a.m. of the previous day until 7 a.m. local time and has the date stamp of the previous day. Mean temperature is aggregated over the whole day and has no date shift. The change of a HS measurement of date i relative to the preceding measurement is therefore influenced by the precipitation of date $i-1$ and a combination of the temperature at the two dates i and $i-1$. For being able to test methods for reconstructing missing data in a controlled environment, a "leave-one-winter-out" cross-validation is performed. Data for one winter (Nov-Apr) is deleted (gap period) and in case a parameter training is required for the respective method, this is done on the winter data of the remaining 20 winters (training period). Locations of the stations used in the cross validation study can be seen in Fig. 1. We test the spatial interpolation methods in two different station networks in order to

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

assess sensitivity against sparser station networks. Sparser networks can be expected in areas of the world which are not as densely populated as Switzerland or in earlier times such as e.g. in the mid 20th century when much fewer stations measured snow depth in Switzerland. The dense network contains 33 evaluation stations (blue triangles in Fig. 1) as well as additional 93 neighboring predictor stations (orange squares in Fig. 1) and covers stations in an altitudinal range of 230 to 2536 m above sea level. The sparser network consists of the evaluation stations only and covers an altitudinal range of 273 to 1970 m above sea level. If two stations were situated closer than 3 km to each other, one of the two stations was excluded from the station sets. In order to test every method at the same set of stations, evaluation stations are chosen thus they have a continuous record for all three variables HS, temperature and precipitation. Therefore, gaps are only filled at the evaluation stations or both station networks. For the stations ARO, DAV and ULR we combined temperature and precipitation data measured by MeteoSwiss with HS data that was measured by the SLF at a close by partner station. Gaps shorter than three days in the HS time series have been filled by linear interpolation.

2.1 Interpolation methods

2.1.1 Selection of neighboring stations for spatial interpolation methods

Six different methods are employed to interpolate a missing winter of snow depth data at a certain station with help of neighboring stations or by using measured meteorological data at the gap station. In case neighboring stations are used as predictors for reconstructing the missing data, these stations have to be within a radius of 200 km and show an absolute elevation difference of less than 500 m. We choose these limits based on a correlation analysis of Matiu et al. (2021). For all methods which use HS data from neighboring stations, the best n correlated neighboring stations are chosen as predictor stations. If less than n stations meet the constraints defined above, the number of predictor stations is reduced accordingly. To select the best reference stations, Pearson correlations between target station and neighboring stations are computed in the training period only (see Sec. 2 for definition). The maximum number of potential predictor stations for each of the spatial interpolation methods has been determined in another cross validation study where we varied the number of maximum potential predictor stations from 3 to 25 stations. This sensitivity study is performed only on the complete station network as for the sparse network the maximum number of 25 stations would not be reached in many example cases. Results of this sensitivity study and the maximum number of potential predictor stations is discussed further in Sec. 3.1.

2.1.2 Best correlated station (BCS)

The simplest approach we test for imputing missing data is to directly use HS data from the best correlated neighboring station (BCS). Correlation is calculated in the training period and the constraints defined in Section 2.1.1 have to be fulfilled. As a simple bias correction measure, the data from the BCS is multiplied with the ratio of the mean at the target site to the mean at the BCS calculated in the training period.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

2.1.3 Inverse distance weighting (IDW)

The inverse distance weighting (IDW) method uses a weighted spatial average of neighboring stations to impute missing values at the target station, neglecting any elevation gradients. Weights are the inverse squared distance of the respective neighboring station to the target station such that

$$\hat{y} = \frac{\sum_{i=1}^{n} \frac{y_i}{d_i^2}}{\sum_{i=1}^{n} \frac{1}{d_i^2}} \tag{1}$$

where $\hat{y}$ is the estimated snow depth at the target station, n is the number of neighboring reference stations, y_i is the snow depth at neighboring station i and d_i is the distance of the neighboring station i to the target station. Imputed values are rounded to the nearest integer. IDW is besides nearest neighbor and non-weighted local averages one of the most often used methods for reconstructing climatological data (Beguería et al., 2019; Kanda et al., 2018).

2.1.4 Weighted normal ratio (WNR)

Matiu et al. (2021) use a variation of the weighted normal ratio (WNR) method for filling short and longer gaps (up to few years) in daily snow depth time series. The normal ratio method was first introduced by Paulhus and Kohler (1952) and assumes a constant ratio of the average state of two neighboring stations (Young, 1992; Yozgatligil et al., 2013). In the version of Matiu et al. (2021), missing values are filled by

$$\hat{y} = \frac{\sum_{i=1}^{n} w_i y_i \frac{\bar{y}}{\bar{y_i}}}{\sum_{i=1}^{n} w_i} \tag{2}$$

where n is the number of neighboring reference stations, y_i is the snow depth at neighboring station i, $\bar{y}$ and $\bar{y_i}$ are the mean snow depth at the target station and reference station i in the training period respectively and w_i is the weight of station i based on the vertical distance $Z - Z_i$ calculated as $w_i = e^{-\ln 2 (Z - Z_i)^2 / 250^2}$ which is a Gaussian weight function with a full width at half maximum of 500 m. Reconstructed values are rounded to the nearest cm integer. In order to have equal conditions within our method comparison, the selected neighboring stations do not need to have a correlation coefficient larger than 0.7 with the target contrary to the original version in Matiu et al. (2021).

2.1.5 Elastic Net (ENET) regression

As fourth method for reconstructing missing HS data at a target station, we use a multiple linear regression of the HS data from the best correlated neighboring stations. As the neighboring stations often are as well correlated with each other, we use Elastic Net (ENET) regularization to reduce the variance of the model (Zou and Hastie, 2005; Friedman et al., 2010). Elastic Net combines the $l1$ regularization term employed in LASSO (Tibshirani, 1996) and the $l2$ regularization term used in ridge regression (Hoerl and Kennard, 1970) and is thus able to deal with multicollinearity in the predictors. The ratio between $l1$ and $l2$ regularization and the hyperparameter α are optimized in a 5-fold cross validation on the data in the training period. Before fitting and predicting with the model, predictors and target are standard scaled to have a mean of 0 and standard deviation of

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

1 based on the data in the training period. Reconstructed values are rounded to the nearest cm integer and negative predicted values are clipped to zero.

2.1.6 Random Forest (RF) regression

As fifth method we employ Random Forest (RF) regression as a nonlinear combination of neighboring stations. A random forest is an ensemble of decision trees that are drawn from random subsets of the training data (Breiman, 2001). The prediction of the ensemble is the average of the individual trees. We use the best correlated neighboring stations as predictors that satisfy the requirements defined in Section 2.1.1. In order to capture potential different relationships between stations in the course of a snow season, we additionally pass the three seasons early Winter (Nov, Dez), mid Winter (Jan, Feb) and late Winter (Mar, Apr) as a categorical predictor to the model. Prior to fitting the model, this "seasons" predictor is one-hot encoded, whereas the other predictors of neighboring station HS data are standard-scaled as for the elastic net regression (Section 2.1.5). The random forest model has a tree number of 200 and a maximum depth of 70.

2.1.7 Snow model (SM)

As last method we make use of a simple snow model (SM). The snow model consists of a temperature-index model which is then coupled to a density model to estimate the snow depth. For estimating snow water equivalent (SWE) in the snowpack, we use the Snow-17 model which uses a temperature-index approach with a seasonally varying melt factor (Anderson, 1973). However, we do not use the density parameterization described in the former reference. Instead, we post-process the SWE time-series of the temperature-index model with a very simple density model. The density model uses an approach based on Martinec and Rango (1991) in which a time dependent density for the different layers in the snowpack is assumed:

$$\rho(t) = \rho_{max} + (\rho_0 - \rho_{max})e^{-t/\tau} \tag{3}$$

Each layer that is identified by an increase in SWE has an initial new snow density ρ_0, which is temporally increasing according to Eq. 3 at each time step t until it reaches a maximum density ρ_{max}. When SWE decreases during a day, the density model removes layers from top of the snowpack for compensating the loss in SWE. During the cross-validation, only the parameters of the density model ρ_0, ρ_{max} and τ are optimized by grid-searching a predefined reasonable parameter space during the training period for each station and synthetic gap individually to minimize the root-mean-squared error (RMSE) in the training period. No parameter optimizations are done for the melt and accumulation model and the parameters defined in Anderson (1973) are used. We considered to use a combined temperature of two days to correspond with the interval of precipitation and HS data (see Section 2. However, we found negligible differences in model performance and decided to leave the input data as is to avoid potential smoothing of temperature signals. In contrast to the inter-station methods described above, we apply the snow model over the full hydrological year in order account for snow which has already built up until November. However, scoring is only done in the winter months Nov-Apr.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

2.2 Evaluation metrics

As score metrics of the reproduced daily HS values we use the RMSE, the coefficient of determination (r^2) and the BIAS. The BIAS is calculated as the average error. RMSE and BIAS can be interpreted in the same unit as the HS measurements [cm]. As fourth metric, we use the mean arctangent absolute percentage error (MAAPE) which was introduced by Kim and Kim (2016) as a relative error term (limited to a maximum of 1.6) because of frequent close-to-zero HS values for stations at low elevation which blow up traditional relative error terms such as the mean absolute percentage error. Since we are interested in gap filling for climatological analyses, we additionally test how good the different methods are able to reproduce three snow climate indicators which are frequently used by practitioners. These snow climate indicators are i) the average snow depth in a winter (HSavg) which is widely used to test for trends in snow climatology, ii) the maximum snow depth in a winter (HSmax) which is an important indicator for e.g. prevention measures in engineering, and iii) the number of snow days with HS$\geq$1 cm (dHS1) which has vital importance for ecology and the winter tourism industry.

3 Results and Discussion

3.1 Number of potential predictor stations

The influence of the maximum number of neighboring stations is displayed in Fig. 2. Boxplots of RMSE and MAAPE scores caluclated in the reconstructed winters are shown for varying numbers of neighboring stations for the different spatial interpolation methods. The methods have been evaluated in the dense station network. IDW shows decreasing performance for both RMSE and MAAPE with increasing number of predictors. The median RMSE increases from 3.9 for one predictor station to 5.6 for 25 predictor stations. For WNR, the median MAAPE is increasing with increasing number of neighboring stations from 0.21 for one neighboring station to 0.37 for a maximum number of 25 neighboring stations. However, WNR performs best in terms of RMSE for a maximum number of 5 neighboring stations with a median RMSE of 3.1. RF and ENET generally show increasing performance with increasing number of predictor stations. For ENET, median RMSE is decreasing from 3.3 for one predictor station to 2.7 for a maximum number of 15 predictor stations. Above 15 predictor stations, a minimal increase of median RMSE to 2.8 is observable. MAAPE scores are decreasing and show a lower spread for increasing maximum number of predictor stations. However, further increase from 15 stations does not yield remarkable differences in median MAAPE and its variance. For RF, RMSE constantly decreases with increasing maximum number of predictor stations from 3.5 for one predictor station to 2.9 for a maximum number of 25 predictor stations. MAAPE scores for RF are insignificantly better for maximum station numbers of 3, 5 and 10 than for higher maximum station numbers.

Some of the methods are more sensitive to the maximum number of used neighboring stations than others. The deterministic approaches (IDW, WNR) regresses in skill for more stations because more stations introduce unnecessary noise. This is the reason why other studies that use regional averages or simple linear regressions also use only few neighboring stations for reconstructing missing data (e.g. Matiu et al., 2021; Tardivo and Berti, 2014). Regularization measures, which are both included in the ENET and RF regression, allow to choose the best predictors from a given set of predictor stations. Therefore,

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

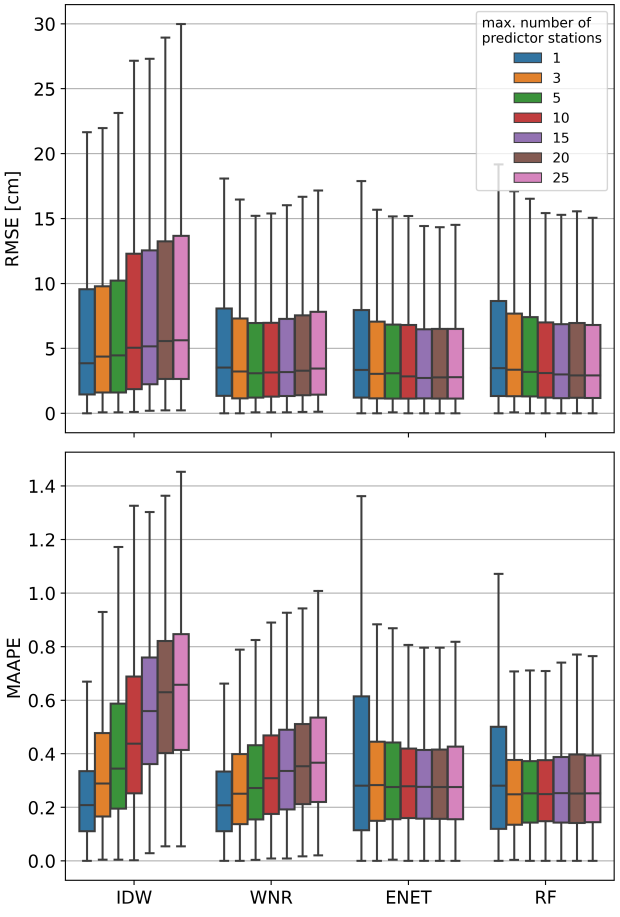

Figure 2. Boxplots of RMSE and MAAPE calculated in the individual reconstructed winters with varied maximum number of predictor stations for the spatial interpolation methods. The methods have been applied to the complete station network. For better comparison, outliers are not shown in the boxplots. Note that WNR with one predictor station is equivalent to the BCS method.

overfitting is prevented even for a larger number of predictors with these two methods. Tests on how many predictor stations are influential for the Random Forests showed that only few stations (less than ~5) share the majority of feature importance. The selected number of maximum neighboring stations for the method comparison in Section 3.2.1 and 3.2.2 is mainly based on the median RMSE and MAAPE scores presented earlier. If scores from two different maximum numbers of predictor stations are approximately equal for one method, we decided to use the lower number of stations to keep the method as simple as possible. Accordingly, we use the maximum numbers of predictor stations listed in Tab. 1 for the comparison of different methods in the following sections.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

Table 1. Selected number of neighboring stations for each method.

Method	max. # neighboring stations
BCS	1
IDW	3
Matiu	3
ENET	15
RF	10
SM[1]	0

[1]only temperature and precipitation data from the target station is used

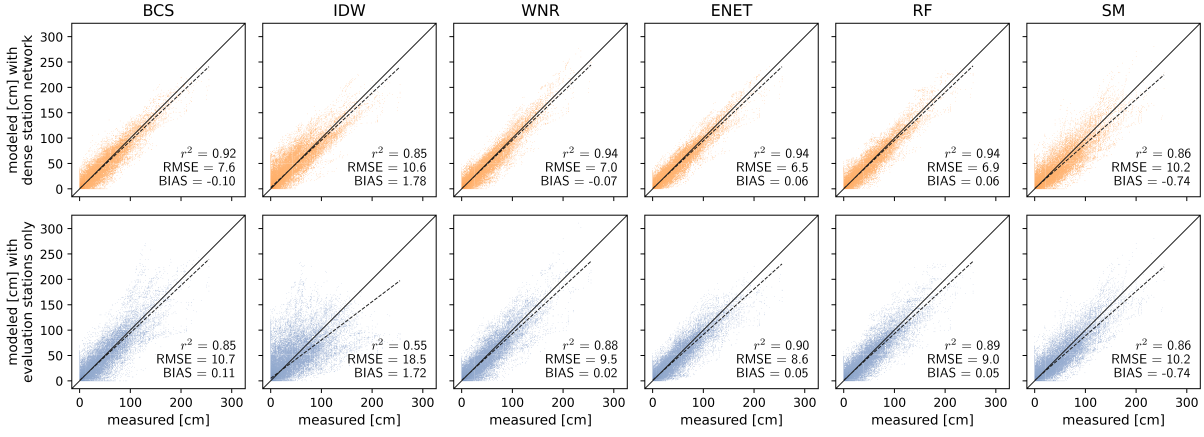

Figure 3. Reconstructed daily snow depth values plotted against the measured values for the used methods (columns). Data in the top row is calculated in the full station network, data in the bottom row is calculated using the evaluation stations only. The solid black line represents perfect predictions, the dashed line is a linear fit of predicted versus measured values. The three score metrics coefficient of determination (r^2), root-mean-squared error (RMSE), and BIAS are indicated in each panel.

3.2 Method performance

3.2.1 Daily values

Predicted daily values are plotted against measured daily values for the different methods and station densities in Fig. 3. Values are aggregated over every filled gap in the cross-validation. The three score metrics r^2, RMSE and BIAS are indicated in each panel. For both the sparse and dense station network, ENET regression yields almost always the best results for all score metrics, shortly followed by RF regression and the WNR method. In the dense station network, WNR, ENET and RF have similar score values with RMSE ranging between 6.5 and 7.0, similar r^2 of 0.94 and equally small BIAS of 0.06 for ENET

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions

and RF and BIAS of -0.07 for WNR. BCS is shortly following WNR, ENET and RF in the dense station network with r^2 of 0.92, RMSE of 7.6 and BIAS of -0.1. IDW is performing poorer than the four aforementioned methods with r^2 of 0.85, RMSE of 10.6 and a positive BIAS of 1.78. The Snow Model performs equal to IDW in the dense station network in terms of RMSE and r^2 with RMSE of 10.2 and r^2 of 0.86. SM predictions are negatively biased with BIAS of -0.74. The SM thus cannot compete with the WNR, BCS, ENET and RF methods in the dense station network. However, the SM (in contrast to IDW) can compete with the WNR and BCS methods in the sparse station network for which the RMSE increases by ~35% and ~40% compared to the dense station network, respectively. RF and ENET are less sensitive against station network density than the WNR and BCS methods but still performance decreases for decreasing station network density. RMSE in the sparser station network decreases by ~30% compared to the dense station network for RF and ENET. IDW is the most sensitive to station network density. RMSE in the sparse station network increases by ~75% and explained variance is significantly lower with r^2 of 0.55 in the sparse station network.

The RMSE scores and BIAS of daily values aggregated over all reconstructed gaps are about double as high as the the median RMSE and BIAS obtained from each gap individually (Fig. A2).

3.2.2 Annual snow climate indicators

HSavg, HSmax and dHS1 derived from the reconstructed daily data (Section 3.2.1) are plotted against the same snow climate indicators derived from the measured data in Fig. 4. Absolute errors of the same snow climate indicators derived form reconstructed data versus the HSavg derived from the measured data in the reconstructed winters are shown in Fig. 5. BCS, WNR, ENET, RF and SM yield unbiased reconstructions of HSavg for both the dense and the sparser station network with BIAS smaller 0.15 cm. For all methods, RMSE for HSavg is about 30 to 40% smaller than the RMSE derived from the aggregated daily values (see Section 3.2.1) for both the reconstructions from the dense and sparser station network. The absolute error of HSavg and HSmax increases with increasing HSavg for all methods (Fig. 5). However, the increase is much larger for BCS and IDW in the case of the sparser station network.

HSmax derived form the filled gaps shows a ~5-10% lower explained variance than HSavg. RMSE values for HSmax are larger than for HSavg but should be compared cautious because of the different scales of the two snow climate indicators. BCS, WNR, ENET, RF and the SM yield negatively biased HSmax with biases ranging from -2.3 to -7.4 cm in the dense and -1.6 to -7.4 cm in the sparse station networks, respectively. IDW shows slightly positive BIAS of 2.8 and 2.9 for the dense and sparse station networks, respectively. Median absolute errors of HSmax are increasing with increasing HSavg for all methods. For BCS and IDW, absolute errors for HSmax are increasingly sensitive to station network density for increasing HSavg.

The dHS1 is reproduced less precisely than HSavg with ~10-20% lower explained variance r^2. All methods apart from BCS and SM strongly overestimate the number of snow days with HS$\geq$1 cm of the reconstructed winters with BIAS from 14.6 to 18.4 days overestimation for the full station network and 16.0 to 23.3 days overestimation for the sparse station network. However, the BCS also slightly overestimates dHS1 with BIAS of 3.7 and 6.6 days in the dense and sparse station networks, respectively. All methods (except SM by method definition) experience an increase in BIAS of dHS1 in the sparse station

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

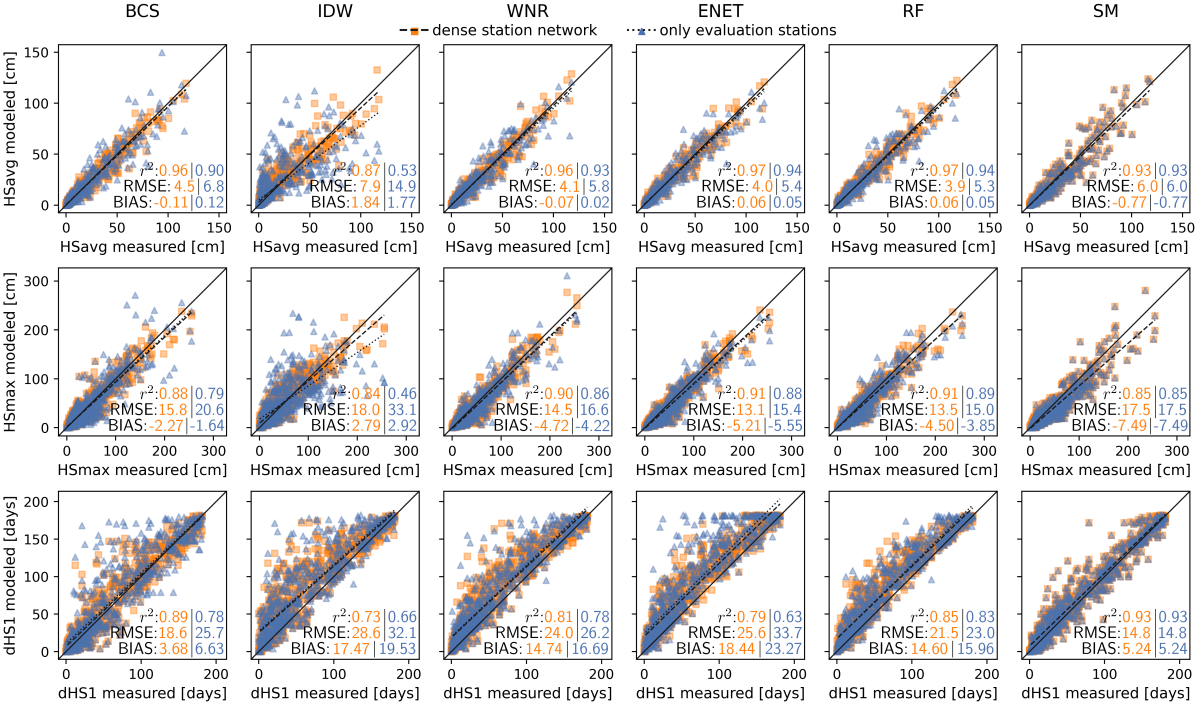

Figure 4. Modeled average snow depth in a winter (HSavg, top row), maximum snow depth in a winter (HSmax, middle row) and number of snow days with HS≥1 cm (dHS1, bottom row) of the reconstructed winters from the cross-validation trials versus the respective snow climate indicator value derived from measurements. The columns refer to the different interpolation methods. Orange squares are gaps reconstructed with the complete station network, blue triangles are gaps that have been reconstructed solely using the evaluation stations as depicted in Figure 1. The black line represents perfect predictions. The dashed and dotted lines are linear fits to the data points of the dense and sparse station networks, respectively. BIAS, RMSE and coefficient of determination (r^2) are indicated in each panel for the dense station network (orange) and the sparse station network (blue).

network compared to the dense station network. For all methods, the absolute error of dHS1 is largest in winters with HSavg below 40 cm.

3.3 Applicability and limitations

Snow depth appears to be a good-natured parameter with respect to reconstructing missing data. All methods except of IDW are able to reconstruct HS with a coefficient of determination above 0.8 regardless of the two station networks used. When deciding what method to choose, it depends on the use case (daily values or derived annual climate indicators) and the setting (station network, surrounding topography) in which one wants to reconstruct the data.

In a very dense station network such as the one in Switzerland, BCS is able to reproduce annual snow climate indicators HSavg, HSmax and dHS1 with r^2 above 0.8 and RMSE below 10 cm for the reconstructed daily HS values. This performance

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

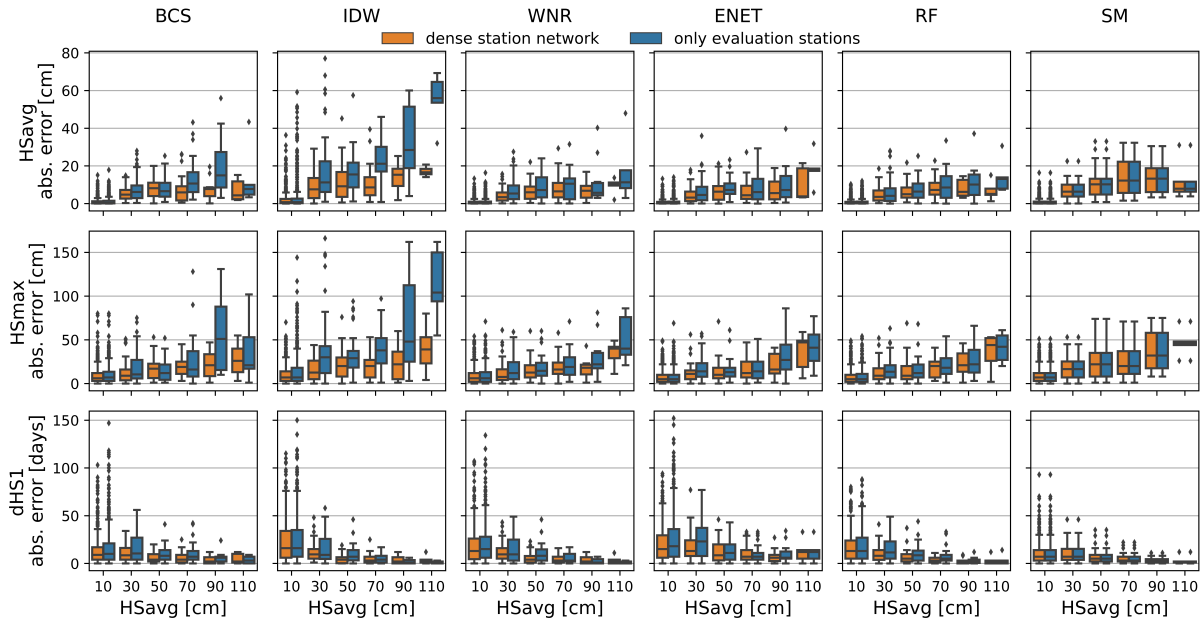

Figure 5. Boxplots of absolute errors in average snow depth in a winter (HSavg, top row), maximum snow depth in a winter (HSmax, middle row) and number of snow days with HS≥1 cm (dHS1, bottom row) calculated for 20 cm HSavg bins of the respective gap-winter and the different methods (columns) respectively. Colors of the boxplots denote the different station networks that have been used for reconstruction. Outliers in the boxplots are not shown for better comparison.

could probably be improved with more advanced bias correction of the neighboring station such as quantile-mapping (Gudmundsson et al., 2012). However, simple approaches such as BCS, IDW and to a smaller extent WNR are sensitive to the density and representativity of the station network. While this is true for every method that uses neighboring stations, more sophisticated methods such as ENET and the nonlinear RF regression are able to almost retain skill also for sparser station networks. Consequently, ENET and RF are besides the SM the most promising candidates in regions with a sparser station network.

Simple spatial averaging with IDW is not able to resemble strong gradients that are present in an alpine topography. We therefore also tested the gradient-plus-inverse-distance-squared (GIDS) method (not shown in results) introduced by Nalder and Wein (1998) which was used in a project of the Austrian Meteorologic service for imputing gaps in HS time-series (Schöner and Koch, 2016). In the sparse network GIDS performed even weaker than IDW, which is in accordance with Price et al. (2000) who observed poor results with GIDS for temperature and precipitation reconstruction in areas with strong topography.

Buchmann et al. (2021) evaluated the natural variability of annual snow climate indicators by comparing data from parallel station pairs (<3 km distance and <100 m elevation difference). They find RMSE for HSavg within a station pair to be in the same range as RMSE for reproduced HSavg with the ENET, RF, WNR and SM methods. This proves that HSavg can be reproduced reasonably well with these four methods. Even the best performing method cannot reach the quality of a parallel

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

station pair for HSmax and dHS1. RMSE of the RF method is 2 respectively 4 times larger then the median RMSE within a parallel station pair for these two snow climate indicators (Buchmann et al., 2021).

For all methods, highest median absolute errors and BIAS for dHS1 can be observed in winters with low HSavg. These winters are often characterized by an ephemeral snow cover which is building up and vanishing again in the course of the winter. Temperature index models are prone to have problems with these kinds of snow covers which could explain the weaker performance of the SM method in these consitions (Hughes and Robinson, 1993; Gray and Landine, 1988). The positive BIAS of dHS1 for the methods that use several neighboring stations may be explained as following. The probability that at least one of the neighboring stations has snow at a certain day is higher than the probability of snow at the target station. Since most of the methods combine data of the neighboring stations, this will result in statistically more days with snow. When trying to minimize BIAS in dHS1, it is therefore best to rely on only few neighboring stations. Accordingly, BCS yields predictions for dHS1 that have a lower positive BIAS. One possible approach to reproduce dHS1 more accurately than deriving it from reconstructed daily values, could be to model dHS1 directly. This could be realized by fitting a nonlinear statistical model such as random forest to the dHS1 series of the target station with dHS1 series derived from neighboring stations as predictors. However, the reduced number of data points would ideally require a longer training period of simultaneous measurements at target and neighboring stations, respectively.

An option to increase the skill of the deterministic methods BSC, IDW and WNR, is to apply stricter constraints to the neighboring stations as done in (Matiu et al., 2021) by introducing a correlation constraint to the neighboring stations (see Section 2.1.4). In the station networks applied in this study, this would lead to a failure in filling data in 15 and 20% of the filled gaps (stationyears) for the dense and sparse station network, respectively. These cases occured mostly for stations at low elevations (AIG, ALT, GVE, SIO, VIS, see Fig.1) with an ephemeral snow cover.

Due to semi-automatic quality control procedures and careful station preselection, our test data set did only contain very few missing HS values for the reference stations. However, this is rather unlikely to be encountered in a real application. Missing values in neighboring stations can be handled differently by different methods. ENET does not allow a single missing value in one of the neighboring stations in the train and gap period. On the other hand, RF and the WNR method are able to deal with missing values in the predictor stations which is a huge asset when it comes to applicability. The effect of missing values in neighboring stations on the performance has not been tested in this study. However, this is an important point to keep in mind when trying to apply any of the evaluated methods.

One potential limitation of the SM approach is, that if the snow measurements are interrupted at a certain station, possibly other variables which are needed as input for the snow model could also be missing. However, this is a rather unlikely case to encounter at least in the dataset of Switzerland. Temperature and precipitation traditionally have a higher priority for weather services than the variables associated with snow and therefore in case an issue occurred at a station, the probability of continuation of these two classic meteorologic variables is higher than for snow. After the automation of many weather stations (not for snow) in Switzerland in the 1980s, long gaps in the temperature and precipitation record are even less likely to be encountered.

A general limitation of our analysis may be the fact that the sparse station network is still dense when compared to station networks present in other regions of the world (Gubler et al., 2017). If the station network is sparser than in our example, the

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

snow model and RF should be favoured over the other approaches as these both methods show the least sensitivity to station network density in our analysis.

4 Conclusions

We compared different methods for reconstructing long gaps in daily manual HS data records as well as their ability to reconstruct the annual snow climate indicators HSavg, HSmax and dHS1. The ENET, RF, WNR and BCS method are able to reproduce daily HS values with coefficient of determination above 0.9 in the dense and above 0.8 in the sparse station network, respectively. Median RMSEs of the filled gaps are below 4 cm for all methods. The SM which does not need data from neighboring stations reveals only slightly lower coefficient of determination (0.86) for daily HS values. The two annual climate indicators HSavg and HSmax, in contrast to dHS1, can be well reproduced by BCS, ENET, RF, SM and WNR. All methods except for SM and BCS overestimate the dHS1 with BIAS of 15 to 23 days. In a sparse station network a simple snow model is best suited to resemble dHS1 most accurately with r^2 of 0.93.

The reconstruction errors of HSavg are within the natural variability of a parallel station pair. Snow depth seems to be a relatively good-natured parameter when it comes to gap filling of data with neighboring stations. However, when station networks get sparse, temperature index snow models serve as a suitable alternative to classic inter-station gap filling approaches.

Since most of the methods perform reasonably well, the choice of which method to use depends on the specific use case and setting. If a serially complete, highly correlated station is available, bias corrected data from this station is easy to calculate and, in many instances, sufficient enough to be used in a climatological use case. If no meteorological data is available at the target station and if neighboring stations regularly contain missing data as well, WNR is a suitable deterministic approach to reconstruct data from neighboring stations. Missing data in neighboring stations can also be handled by RF. If the station network is sparser than in our study and if neighboring stations are further away and weakly correlated, the snow model, ENET and RF should be favoured over the other approaches as these three methods show the least sensitivity to station network density in our analysis. If the focus of the analysis is set on dHS1, a simple snow model is best suited to reconstruct a complete missing winter. If no meteorological data is available, BCS should be the fallback solution for dHS1 in case a suitable reference station is available.

Code and data availability. Python code to perform the analysis and to use the methods on other data is available from Aschauer (2021). Due to data policies, input data to reproduce the analyses is available upon request from the authors.

Appendix A: Additional Figures

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

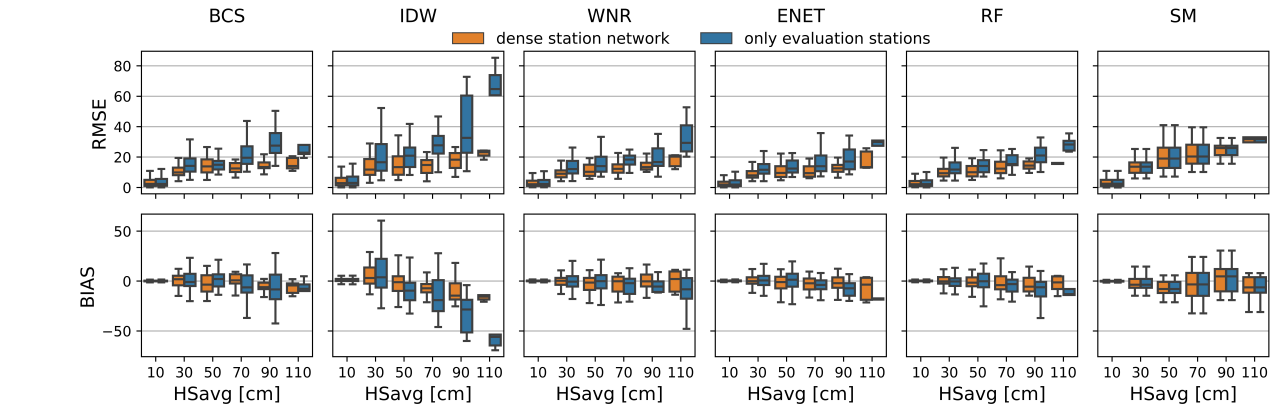

Figure A1. Boxplots of RMSE (top row) and BIAS (bottom row) calculated for 20 cm HSavg bins of the respective gap-winter and the different methods (columns) respectively. Colors of the boxplots denote the different station networks that have been used for reconstruction. Outliers in the boxplots are not shown for better comparison. The maximum number of predictors for the different methods is set as defined in Tab. 1.

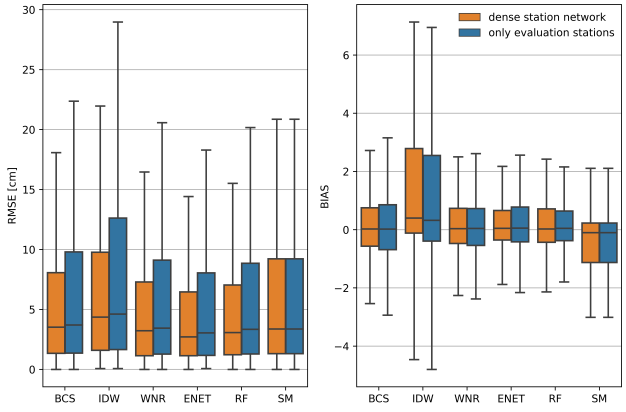

Figure A2. Boxplots for root-mean-squared error (RMSE) and BIAS of the daily values for the different methods and station densities. The maximum number of predictors for the different methods is set as defined in Tab. 1. The station network density is irrelevant for the snow model as no data from neighboring stations is used. For better comparison, outliers are not shown in the boxplots.

Author contributions. JA and CM designed the study. JA performed the analysis and drafted the manuscript. Both authors discussed the results and commented on the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

Acknowledgements. We want to thank MeteoSwiss for providing data of their meteorological stations as well as Tobias Jonas for input on the density model. Thank you to Moritz Buchmann for valuable discussions and comments on the manuscript.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

References

Anderson, E. A.: National Weather Service River Rorecast Rystem - Snow Accumulation and Ablation Model, NOAA Technical Memorandum NWS-HYDRO-17, US Depart. of Commerce, Silver Spring, MD, 1973.

Aschauer, J.: source code: Evaluating methods for reconstructing large gaps in historic snow depth time series, https://doi.org/10.5281/zenodo.4785141, 2021.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.-M., Begert, M., Müller-Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Szalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovic, Z., and Nieplova, E.: HISTALP—historical instrumental climatological surface time series of the Greater Alpine Region, International Journal of Climatology, 27, 17–46, https://doi.org/10.1002/joc.1377, 2007.

Avanzi, F., Zheng, Z., Coogan, A., Rice, R., Akella, R., and Conklin, M. H.: Gap-filling snow-depth time-series with Kalman filtering-smoothing and expectation maximization: Proof of concept using spatially dense wireless-sensor-network data, Cold Regions Science and Technology, 175, 103 066, https://doi.org/10.1016/j.coldregions.2020.103066, 2020.

Bales, R., Stacy, E., Safeeq, M., Meng, X., Meadows, M., Oroza, C., Conklin, M., Glaser, S., and Wagenbrenner, J.: Spatially distributed water-balance and meteorological data from the rain–snow transition, southern Sierra Nevada, California, Earth System Science Data, 10, 1795–1805, https://doi.org/10.5194/essd-10-1795-2018, 2018.

Beguería, S., Tomas-Burguera, M., Serrano-Notivoli, R., Peña-Angulo, D., Vicente-Serrano, S. M., and González-Hidalgo, J.-C.: Gap filling of monthly temperature data and its effect on climatic variability and trends, Journal of Climate, 32, 7797–7821, https://doi.org/10.1175/JCLI-D-19-0244.1, 2019.

Beniston, M.: Is snow in the Alps receding or disappearing?, Wiley Interdisciplinary Reviews: Climate Change, 3, 349–358, https://doi.org/10.1002/wcc.179, 2012.

Breiman, L.: Random forests, Machine learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brown, R. D.: Evaluation of methods for climatological reconstruction of snow depth and snow cover duration at Canadian meteorological stations, in: Proc. Eastern Snow Conf., 53d Annual Meeting, pp. 55–65, 1996.

Brown, R. D., Hughes, M. G., and Robinson, D. A.: Characterizing the long-term variability of snow-cover extent over the interior of North America, Annals of Glaciology, 21, 45–50, https://doi.org/10.3189/S0260305500015585, 1995.

Brown, R. D., Brasnett, B., and Robinson, D.: Gridded North American monthly snow depth and snow water equivalent for GCM evaluation, Atmosphere-Ocean, 41, 1–14, https://doi.org/10.3137/ao.410101, 2003.

Buchmann, M., Begert, M., Brönnimann, S., and Marty, C.: Evaluating the robustness of snow climate indicators using a unique set of parallel snow measurement series, International Journal of Climatology, 41, E2553–E2563, https://doi.org/10.1002/joc.6863, 2021.

Falarz, M.: Long-term variability in reconstructed and observed snow cover over the last 100 winter seasons in Cracow and Zakopane (southern Poland), Climate Research, 19, 247–256, https://www.jstor.org/stable/24866784, 2002.

Friedman, J., Hastie, T., and Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent, Journal of statistical software, 33, 1, https://doi.org/10.18637/jss.v033.i01, 2010.

Gray, D. M. and Landine, P. G.: An energy-budget snowmelt model for the Canadian Prairies, Canadian Journal of Earth Sciences, 25, 1292–1303, https://doi.org/10.1139/e88-124, 1988.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021


Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

375 Gubler, S., Hunziker, S., Begert, M., Croci-Maspoli, M., Konzelmann, T., Brönnimann, S., Schwierz, C., Oria, C., and Rosas, G.: The influence of station density on climate data homogenization, International Journal of Climatology, 37, 4670–4683, https://doi.org/10.1002/joc.5114, 2017.

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T.: Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods, Hydrology and Earth System Sciences, 16, 3383–3390,
380 https://doi.org/10.5194/hess-16-3383-2012, 2012.

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, 12, 55–67, https://doi.org/10.1080/00401706.2000.10485983, 1970.

Hughes, M. G. and Robinson, D. A.: Creating temporally complete snow cover records using a new method for modelling snow depth changes, World Data Center A, Glaciology (Snow & Ice), pp. 150–163, 1993.

385 Kanda, N., Negi, H. S., Rishi, M. S., and Shekhar, M. S.: Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of Karakoram Himalaya, Meteorological Applications, 25, 337–349, https://doi.org/10.1002/met.1699, 2018.

Kashani, M. H. and Dinpashoh, Y.: Evaluation of efficiency of different estimation methods for missing climatological data, Stochastic Environmental Research and Risk Assessment, 26, 59–71, https://doi.org/10.1007/s00477-011-0536-y, 2012.

390 Kemp, W. P., Burnell, D. G., Everson, D. O., and Thomson, A. J.: Estimating Missing Daily Maximum and Minimum Temperatures, Journal of Applied Meteorology and Climatology, 22, 1587–1593, https://doi.org/10.1175/1520-0450(1983)022<1587:EMDMAM>2.0.CO;2, 1983.

Kim, J.-W. and Pachepsky, Y. A.: Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation, Journal of hydrology, 394, 305–314, https://doi.org/10.1016/j.jhydrol.2010.09.005, 2010.

395 Kim, S. and Kim, H.: A new metric of absolute percentage error for intermittent demand forecasts, International Journal of Forecasting, 32, 669–679, https://doi.org/10.1016/j.ijforecast.2015.12.003, 2016.

Marke, T., Hanzer, F., Olefs, M., and Strasser, U.: Simulation of past changes in the Austrian snow cover 1948–2009, Journal of Hydrometeorology, 19, 1529–1545, https://doi.org/10.1175/jhm-d-17-0245.1, 2018.

Martinec, J. and Rango, A.: Indirect evaluation of snow reserves in mountain basins, IAHS Publ, 205, 111–119, 1991.

400 Marty, C.: Regime shift of snow days in Switzerland, Geophysical research letters, 35, https://doi.org/10.1029/2008gl033998, 2008.

Marty, C. and Blanchet, J.: Long-term changes in annual maximum snow depth and snowfall in Switzerland based on extreme value statistics, Climatic Change, 111, 705–721, https://doi.org/10.1007/s10584-011-0159-9, 2012.

Massetti, L.: Analysis and estimation of the effects of missing values on the calculation of monthly temperature indices, Theoretical and Applied Climatology, 117, 511–519, https://doi.org/10.1007/s00704-013-1024-8, 2014.

405 Matiu, M., Crespi, A., Bertoldi, G., Carmagnola, C. M., Marty, C., Morin, S., Schöner, W., Cat Berro, D., Chiogna, G., and De Gregorio, L.: Observed snow depth trends in the European Alps: 1971 to 2019, The Cryosphere, 15, 1343–1382, https://doi.org/10.5194/tc-15-1343-2021, 2021.

Nalder, I. A. and Wein, R. W.: Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest, Agricultural and Forest Meteorology, 92, 211–225, https://doi.org/10.1016/S0168-1923(98)00102-6, 1998.

410 Notaro, M., Lorenz, D. J., Vimont, D., Vavrus, S., Kucharik, C., and Franz, K.: 21st century Wisconsin snow projections based on an operational snow model driven by statistically downscaled climate data, International Journal of Climatology, 31, 1615–1633, https://doi.org/10.1002/joc.2179, 2011.

https://doi.org/10.5194/gi-2021-16
Preprint. Discussion started: 9 June 2021

Olefs, M., Koch, R., Schöner, W., and Marke, T.: Changes in Snow Depth, Snow Cover Duration, and Potential Snowmaking Conditions in Austria, 1961–2020—A Model Based Approach, Atmosphere, 11, 1330, https://doi.org/10.3390/atmos11121330, 2020.

Paulhus, J. L. and Kohler, M. A.: Interpolation of missing precipitation records, Monthly Weather Review, 80, 129–133, 1952.

Price, D. T., McKenney, D. W., Nalder, I. A., Hutchinson, M. F., and Kesteven, J. L.: A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data, Agricultural and Forest meteorology, 101, 81–94, https://doi.org/10.1016/s0168-1923(99)00169-0, 2000.

Schöner, W. and Koch, R.: SNOWPAT – Schnee in Österreich, Snow in Austria during the instrumental period – spatiotemporal patterns and their causes – elevancefor future snow scenarios, resreport, Zentralanstalt für Meteorologie und Geodynamik ZAMG, report for Austrian Climate Research Program, 2016.

Tardivo, G. and Berti, A.: A dynamic method for gap filling in daily temperature datasets, Journal of Applied Meteorology and Climatology, 51, 1079–1086, https://doi.org/10.1175/jamc-d-11-0117.1, 2012.

Tardivo, G. and Berti, A.: The selection of predictors in a regression-based method for gap filling in daily temperature datasets, International Journal of Climatology, 34, 1311–1317, https://doi.org/10.1002/joc.3766, 2014.

Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58, 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x, 1996.

Witmer, U.: Eine Methode zur flächendeckenden Kartierung von Schneehöhen unter Berücksichtigung von reliefbedingten Einflüssen, Phd thesis, Universität Bern, 1984.

Woldesenbet, T. A., Elagib, N. A., Ribbe, L., and Heinrich, J.: Gap filling and homogenization of climatological datasets in the headwater region of the Upper Blue Nile Basin, Ethiopia, International Journal of Climatology, 37, 2122–2140, https://doi.org/10.1002/joc.4839, 2017.

Young, K. C.: A three-way model for interpolating for monthly precipitation values, Monthly Weather Review, 120, 2561–2569, https://doi.org/10.1175/1520-0493(1992)120<2561:ATWMFI>2.0.CO;2, 1992.

Yozgatligil, C., Aslan, S., Iyigun, C., and Batmaz, I.: Comparison of missing value imputation methods in time series: the case of Turkish meteorological data, Theoretical and applied climatology, 112, 143–167, https://doi.org/10.1007/s00704-012-0723-x, 2013.

Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x, 2005.