

# Robust climate scenarios for sites with sparse observations: a two-step bias correction approach

Jan Rajczak,<sup>a\*</sup> Sven Kotlarski,<sup>a</sup> Nadine Salzmänn<sup>b</sup> and Christoph Schär<sup>a</sup>

<sup>a</sup> Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

<sup>b</sup> Department of Geosciences, University of Fribourg, Switzerland

**ABSTRACT:** Observed and projected climatic changes demand for robust assessments of climate impacts on various environmental and anthropogenic systems. Empirical-statistical downscaling (ESD) methods coupled to output from climate model projections are promising tools to assess impacts at regional to local scale. ESD methods correct for common model deficiencies in accuracy (e.g. model biases) and scale (e.g. grid vs point scale). However, most ESD methods require long observational time series at the target sites, and this often restricts robust impact assessments to a small number of sites. This paper presents a method to generate robust climate model based scenarios for target sites with short and (or) sparse observational data coverage. The approach is based on the well-established quantile mapping method and incorporates two major steps: (1) climate model bias correction to the most representative station with long-term measurements and (2) spatial transfer of bias-corrected model data to represent target site characteristics. Both steps are carried out using the quantile mapping technique. The resulting output can serve as end user-tailored input for climate impact models. The method allows for multivariate and multi-model ensemble scenarios and additionally enables to approximately reconstruct data for non-measured periods. The method's applicability is validated using (1) long-term weather stations across the topographically and climatologically complex territory of Switzerland and (2) sparse data sets from Swiss permafrost research sites located in challenging conditions at high altitudes. It is shown that the two-step approach performs well and offers attractive quality, even for extreme target locations. Uncertainties, however, remain and primarily depend on (1) data availability and (2) the considered variable. The two-step approach itself involves large uncertainties when applied to short reference data sets or spatially heterogeneous variables (e.g. precipitation, wind speed). For temperature, results are promising even when using very short calibration periods.

**KEY WORDS** climate impact assessment; climate change; regional climate models; projections; bias correction; statistical downscaling

## 1. Introduction

Climate change and climate variability foster a growing demand for sophisticated and robust climate impact assessments that serve various research fields and decision makers (IPCC: Collins *et al.*, 2013; Hartmann *et al.*, 2013; Kirtman *et al.*, 2013). Such assessments are usually undertaken using impact models that are driven by raw or statistically post-processed output of climate model simulations (e.g. Salzmänn *et al.*, 2007; Bosshard *et al.*, 2011, 2013; Scherler *et al.*, 2013). Projections of future climatic conditions primarily rely on output from global and regional climate model (GCM, RCM) ensembles (e.g. PRUDENCE: Christensen and Christensen, 2007, CMIP3: Meehl *et al.*, 2007, ENSEMBLES: van der Linden and Mitchell, 2009, NARCCAP: Mearns *et al.*, 2009, CMIP5: Taylor *et al.*, 2011, EURO-CORDEX: Kotlarski *et al.*, 2014) driven by greenhouse gas emission scenarios. Even though climate

models offer an attractive tool for regional to continental assessments of climate change, the grid resolution of current RCMs in the range of 10 to 50 km is still too coarse to be directly applied in impact studies operating at regional to local scale. This is particularly true for regions of complex topography, where climatic variations over a climate model gridbox can be large. In addition, climate model simulations are subject to considerable biases and uncertainties (e.g. Flato *et al.*, 2013) originating from internal variability, model and scenario uncertainties (Hawkins and Sutton, 2009).

In order to bridge the gap between models and both the desired scales (grid versus local scale) and accuracy (model biases), empirical-statistical downscaling (ESD) methods are a promising tool to translate climate model output to the local scale and in turn provide end user-tailored climate scenarios (Fowler *et al.*, 2007; Maraun *et al.*, 2010). The principle of ESD methods is to reveal a statistical relationship between a climate model simulation and observations for historical time periods. These relations are then summarized into transfer functions that are used to translate RCM (or GCM) projections

\* Correspondence to: J. Rajczak, Institute for Atmospheric and Climate Science, ETH Zurich, Universitaetstrasse 16, 8092 Zurich, Switzerland.  
E-mail: jan.rajczak@env.ethz.ch

to the local scale and to correct for certain aspects of model biases.

Different ESD methods were systematically compared in recent studies (Themeßl *et al.*, 2011; Teutschbein and Seibert, 2012; Räisänen and Rätty, 2013; Rätty *et al.*, 2014). Many of these studies identify quantile mapping (in the following QM) as the best performing method in cross-validation frameworks.

The basic principle of QM is to correct for errors in the entire distributional behaviour of a model simulation (Wood *et al.*, 2004; Boé *et al.*, 2007; Déqué, 2007). In essence, the  $n^{\text{th}}$  quantile of the model distribution is translated into the  $n^{\text{th}}$  quantile of the observed distribution, based on simulated and observed data in a reference period. The distributional adjustment carried out during QM implicitly corrects for both errors related to the different scales and systematic model biases.

Gudmundsson *et al.* (2012) compared different implementations of QM for daily precipitation data and found that a nonparametric empirical approach outperforms different implementations relying on theoretical distributional assumptions. Gutjahr and Heinemann (2013) even showed that an empirical formulation can outperform theoretical ones in terms of extreme precipitation (far tails of the frequency distribution).

In the framework of climate scenarios, QM exhibits numerous advantages. It enables to generate multivariate and transient scenarios from climate model output. As QM corrects for distributional errors, it is able to project changes in mean and variance. It is also able to account for changes in temporal variability and weather types, provided that the considered climate model projects such changes.

However, a number of open questions remain. For instance, values that exceed (undershoot) the range of values of the calibration period are not calibrated. Furthermore, it is not a priori clear whether the spatial and temporal covariance structure between different local-scale measurements (e.g. a station network) is conserved after application of QM (see also Maraun, 2013; Keller *et al.*, 2014; Wong *et al.*, 2014). These problems, along with an inflation of variance, occur particularly for spatially heterogeneous variables such as precipitation and when applying coarse model output in regions of complex topography where subgrid-scale variability differs substantially from grid-scale variability (von Storch, 1999; Maraun, 2013). Maraun (2013) also showed that trends can be distorted. It is also not fully understood whether multivariate scenarios are physically consistent as individual variables are usually treated separately. However, the work of Wilcke *et al.* (2013) indicates that the inter-variable relation of an RCM simulation is conserved by QM in terms of correlations between pairs of variables. Moreover, impact-relevant indices such as multi-day precipitation statistics might not be represented as precisely as single-day statistics after QM (Addor and Seibert, 2014).

Independent of the applied method, all ESD approaches rely on a stationary statistical relationship between observations and climate model output. This prerequisite

requires a calibration using long-term data records, which constrains the proper application of ESD approaches to a limited number of sites for which sufficiently long time series are available (e.g. a period of 30 years). Especially in a climate scenario context, where observations are linked to a free-running GCM or a GCM-driven RCM experiment, it is inevitable to consider a long period for calibration. This is because observations and models are afflicted by inter-annual and decadal variability that may yield anomalously warm and cold (or wet and dry, etc.) periods. Considering a GCM-driven climate simulation, the modelled temporal variability (day-to-day, year-to-year) is independent from observed variability and only long-term time-integrating statistics can be expected to match. In such a setting, a calibration using only few years of data will result in a non-representative transfer function. In a GCM-RCM coupled climate scenario framework, a robust applicability of ESD approaches is therefore limited to stations (data sets) with long-term measurements where non-synchronous phases of inter-annual to decadal variability cancel each other over time.

Stations that come into consideration when targeting long data sets are usually maintained and operationally run by national weather services. In the meantime, observations from various research sites situated along sensitive environments become increasingly available. However, a majority of these research-site observations are not suitable for a direct application of ESD approaches, due to short data coverage and proneness to missing values. Also, these measurements are typically not available at a quality comparable with those from national weather services, for instance because of difficult access in the case of high-altitude sites. Yet, these are often the locations that are targeted by climate impact assessments. Providing robust climate scenarios for such sites poses a special challenge and up to now includes particularly large uncertainties. It is exactly here where our approach tries to add important value. We present a methodology for the generation of climate scenarios specifically designed for sites with short and sparse data coverage. The proposed method includes (1) the systematic correction of model biases and (2) an adequate representation of local-scale conditions. It is based on the well-established QM method and is designed for an application in alpine environments but versatile in its applicability for other environments. The method also enables an approximate reconstruction of meteorological conditions during unobserved periods.

The paper is structured as follows. First, we introduce the data used and explain the employed methods and proposed two-step procedure. Then, we present a validation of the method and follow up with an exemplary climate scenario application in a high-alpine environment. The paper is finalized with a conclusion.

## 2. Data

Observed and modelled data at daily resolution are used. Eight variables with daily resolution are taken

into account, which are typically of major interest for impacts research, namely: mean (*tas*), minimum (*tasmin*; Figure S1, Supporting Information) and maximum (*tasmax*; Figure S2) 2 m temperature, precipitation sum (*pr*), mean relative humidity (*hurs*; Figure S3), mean wind speed (*wss*; Figure S4), maximum wind gusts (*wsgsmax*; Figure S5) and mean global radiation (*rsds*; Figure S6).

### 2.1. Observations

We use observational data from three sources, of which two are available at the site scale and one on a regular grid. Long-term (1981–2013), daily multivariate observations from a network of 44 stations operated by the Swiss national weather service MeteoSwiss (MCH) are used as reference for a systematic downscaling and bias correction of climate model output and for the evaluation of the presented procedure.

Additionally, multivariate observations from 12 permafrost research sites are used (further on referred to as *TEMPS* sites according to the framing research project *The Evolution of Mountain Permafrost in Switzerland*). These sites represent target sites for the generated climate scenarios in a prototype study. The locations of the MCH stations and the *TEMPS* sites are shown in Figure 1. In general, measurements from the *TEMPS* sites cover only short time periods of a few years, usually covering some fraction of the 1995–2013 period. Moreover, precipitation measurements are not available for these sites and meteorological time series typically contain missing values.

We also employ gridded analyses at daily resolution with a spatial resolution of 2 km for the variables temperature (Frei, 2014) and precipitation (RhiresD; MeteoSwiss, 2014), which are typically based on 80 and 450 station observations, respectively, over Switzerland. The precipitation analysis is used as a proxy for the missing precipitation measurements at the *TEMPS* sites for the period 1981–2010 using values from the 2 km gridbox within which a site is located.

It is important to mention that, particularly for precipitation and in complex topography, the use of gridded analyses implies significant limitations. The statistical procedures underlying RhiresD and the jump from grid to station scale imply major uncertainties. In effect, daily precipitation intensities are smoothed towards smaller (larger) values at locations and on days with actually large (small) event intensities. The precipitation analysis has an effective resolution of 15–20 km. The highly variable Alpine precipitation patterns cannot fully be represented, and localized heavy events (e.g. thunderstorms) might be missed completely, as not captured by the underlying station network. Also, temporal trends in the gridded analyses can be distorted, as the set of stations used to construct the analysis is not stationary in time.

### 2.2. Regional climate model data

In a case study (Section 5.), we employ an ensemble of 14 regional climate model (RCM) simulations from the

EU-ENSEMBLES project (van der Linden and Mitchell, 2009). All models provide data at a horizontal resolution of 0.22° (25 km, see Figure 1), are driven by the IPCC SRES A1B greenhouse gas emission scenario (Nakicenovic and Swart, 2000) and consist of a combination of nine RCMs nested into seven different GCMs (see Table 1). The GCM-RCM matrix samples a large range of model uncertainties and thus offers a comprehensive range of possible future estimates.

## 3. Methods

### 3.1. Standard QM

We apply a nonparametric empirical implementation of the well-established QM technique, with a time-dependent correction function calibrated for each day of the year (*DOY*) using a 91-day moving window centred over the considered day. Our implementation follows previous works by Themeßl *et al.* (2011, 2012) and Wilcke *et al.* (2013), where a raw climate model time series  $X$  at time ( $t$ ) and location ( $s$ ) is corrected to a time series  $Y$  according to

$$Y_{t,s} = ecdf_{DOY,s}^{obs,cal^{-1}} \left[ ecdf_{DOY,s}^{mod,cal} (X_{t,s}) \right] \quad (1)$$

Where *ecdf* is the empirical cumulative distribution function, *obs* denotes observations and *mod* raw model output both in a calibration period (*cal*, e.g. 1981–2010).

$X$ , *mod*, respectively, can also be a time series from another station, as is the case in the second step of the proposed approach (S2, Figure 4, Section 3.2.1. – spatial transfer).

One important drawback of QM is that it calibrates only for the range of observed and simulated values within the calibration period. In the current study, values that lie outside the calibrated range (new extremes) are corrected according to the largest (99<sup>th</sup>), respectively lowest (1<sup>st</sup>), quantile that has been calibrated for (Themeßl *et al.*, 2012).

Under warming conditions, this raises particular concerns regarding temperature, as future values increasingly exceed the range of observed values in historical time periods. To project such changes as robust as possible, we choose a moving window width of 91 days to sample data for a correction function, arguing that, in contrast to smaller window widths (e.g. 31-day), new extremes occur less often and sampling uncertainty is reduced. However, the loss of information regarding the annual cycle that increasingly dominates the specification of an *ecdf* for larger window widths should not be disregarded. For instance, using a time-independent implementation (365-day window) for temperature will yield a correction function in which lower (upper) quantiles are defined by days in winter (summer) and would not be representative for the different systematic biases seen throughout the cycle of the year.

Here, we promote the implementation of a 91-day window, motivated by results shown in Figure 2. The results are based on all 44 MCH stations (Figure 1) and the simulation of maximum daily temperature of the HadRM3Q16

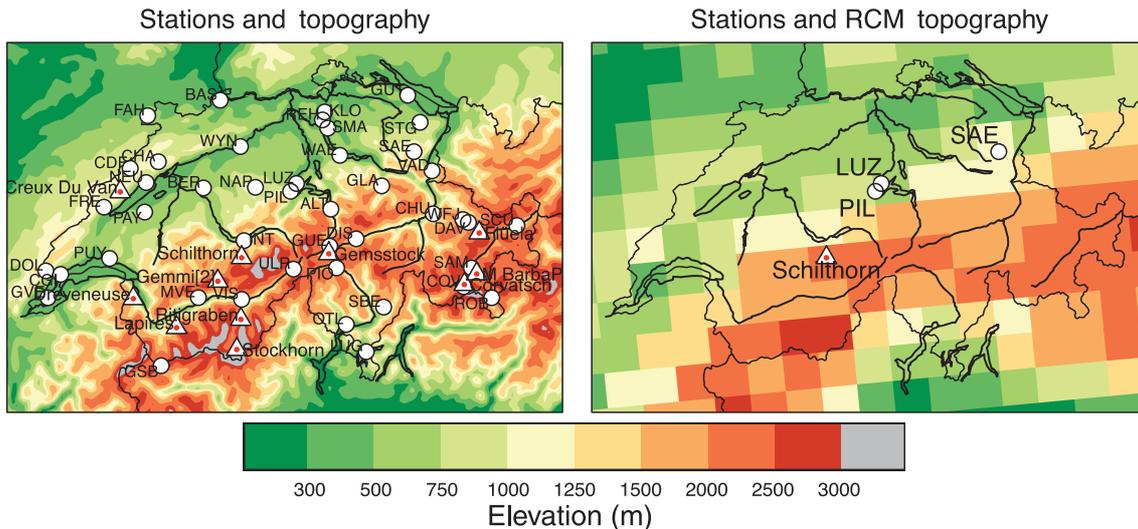


Figure 1. Location of the used MCH weather stations (44, circles in the left panel) and location of exemplary mountain permafrost research sites (12, triangles in the left panel). The left panel depicts high-resolution topography and the right panel the respective topography as presented in the ETH-HadCM3Q0 RCM run by ETH Zurich (Table 1) at a horizontal resolution of 25 km. The right panel shows only the observational sites that are of particular relevance for the paper.

Table 1. Regional climate models used.

Abbreviation	Institute	RCM	GCM	Emission scenario
SMHI-BCM	SMHI	RCA	BCM	IPCC SRES A1B
ETHZ-HadCM3Q0	ETH Zurich	CLM	HadCM3Q0	
HC-HadCM3Q0	Met Office (Hadley Centre)	HadRM3Q0		
C4I-HadCM3Q16	C4I	RCA3	HadCM3Q16	
HC-HadCM3Q16	Met Office (Hadley Centre)	HadRM3Q16		
SMHI-HadCM3Q3	SMHI	RCA	HadCM3Q3	
HC-HadCM3Q3	Met Office (Hadley Centre)	HadRM3Q3		
MPI-ECHAM5	MPI	REMO	ECHAM5	
DMI-ECHAM5	DMI	HIRHAM		
KNMI-ECHAM5	KNMI	RACMO		
SMHI-ECHAM5	SMHI	RCA		
ICTP-ECHAM5	ICTP	REGCM3		
CNRM-ARPEGE	CNRM	ALADIN	ARPEGE	
DMI-ARPEGE	DMI	HIRHAM		

RCM. It depicts at annual and seasonal (winter, DJF; summer, JJA) scale the projected values of the median correction factor (quantile) used (upper row, a, b and c; in the following QMED), and the fraction of correction terms used from the far tail of the present day frequency distribution (>95<sup>th</sup> percentile; bottom row, d, e and f; in the following F95). For both indices, distinct changes take place during the 21<sup>st</sup> century, indicated by an increase of QMED towards larger quantiles and a proportionally larger fraction of correction factors drawn from upper percentiles for all implementations (31-day, 91-day and 365-day moving window widths) and all seasons (year, summer, winter). It is obvious that a 91-day window shows smaller changes than the 31-day window, particularly for F95. We argue that the 91-day implementation shows less sensitivity to yield correction factors drawn from high percentiles of the calibration period. Moreover, Figure 2 indicates the necessity of using a time-dependent implementation, as correction factors in summer (winter) are sampled

from higher (lower) quantiles exclusively when using a time-independent implementation (365-day). The 91-day window seems to be an optimal compromise, especially for variables that feature an annual cycle.

The results are supported by a validation shown by Rätty *et al.* (2014) who also recommend a calibration focusing on seasonal (91-day), instead of monthly (~31-day), transfer functions.

### 3.2. The necessity of climatological-scale calibration

Short and sparse observations reduce the robustness and limit the applicability of ESD techniques coupled to climate model projections (Section 1.). We illustrate this by a simple example presented along Figure 3. It depicts time series of the differences in annual mean temperature between the summit station Pilatus and (1) another proximate mountain station (Säntis, black, located about 88.4 km to the East of Pilatus) and (2) the GCM-RCM chain output (grey) from the corresponding

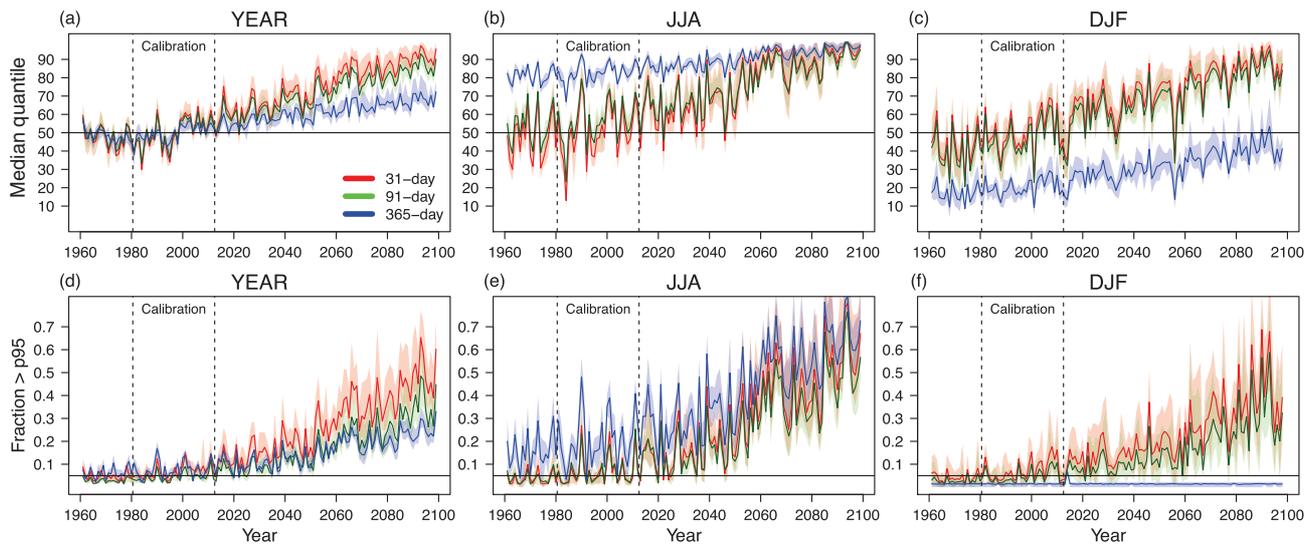


Figure 2. Temporal stability/stationarity of correction functions for 31-day (red), 91-day (green) and 365-day (blue) moving windows used for calibration at day of year basis in period 1981–2012. Correction is undertaken for the variable daily temperature maximum as simulated by the HadRM3Q16 RCM (Table 1). The upper panel presents the median quantile used for correction in a moving window of 30 years. The lower panel presents the fraction of days that are corrected using correction factors drawn from quantiles larger than the 95<sup>th</sup> percentile with respect to the calibration period (1981–2012). Shading denotes the entire range of outcomes from the 44 stations and lines the mean across all stations.

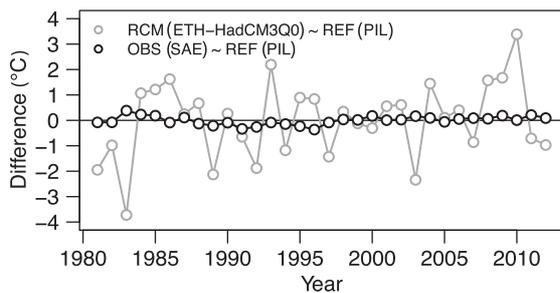


Figure 3. Illustration of annual mean differences between MCH station Pilatus and MCH station Saentis (black) and the ETH-HadCM3Q0 Regional Climate Model simulation for the gridbox located over the station Pilatus (grey). Temperature differences are expressed as the yearly differences subtracted by the mean difference in period 1981–2012.

model gridbox in which Pilatus is located. The small variability of the temperature difference between Säntis and Pilatus indicates a stationary interrelation and, most of all, a matching year-to-year correspondence. Obviously, there is no such stationary relationship and no year-to-year correspondence between the (free-running) GCM-RCM output and the observations at Pilatus. Therefore, one would ideally use as many years as possible to adequately describe and interpret systematic biases between the two series. That is, ESD calibration would have to be carried out on climatological time scales (>20 years). In contrast, there is a clear relation between the neighbouring stations Säntis and Pilatus and only a few years would apparently be enough to characterize systematic differences between both series.

### 3.2.1. Two-step bias correction approach

To overcome the elaborated limitations, we propose a technique to generate robust climate scenarios at target sites

with only short or sparse observational data coverage. The technique essentially consists of a two-step QM procedure (i.e. QM is applied twice) summarized in Figure 4 and described in the following.

*First step (S1): climate model bias correction and downscaling:* Climate model projections are bias corrected to match observations at operational weather stations with long-term observational records (e.g. MCH stations using the calibration period 1981–2010). Bias correction is done via QM (Section 3.1., Equation (1)) and implicitly incorporates the downscaling from the grid to the local scale. The calibration of a QM transfer function is undertaken using long-term observational measurements aiming to remove systematic biases.

*Second step (S2): spatial transfer:* The bias-corrected climate scenarios are spatially transferred from a most representative station (an MRS, i.e. an MCH station for which bias-corrected scenarios are available after S1) to the specific target site (i.e. a TEMPS site). The spatial transfer is also undertaken using the QM technique (Section 3.1., Equation (1) with the target station representing *obs* and the MRS *mod*). The QM correction function used to spatially transfer the bias-corrected time series of an MRS to the target site is based on pairwise daily observations at both locations during the overlapping measurement period. The observations do not have to be contiguous. S2 assures that the generated series for the target site has characteristics representative for local-scale conditions at this site.

The proposed two-step procedure is implemented in a multivariate framework, in which each variable is treated separately. The MRS from which data are spatially

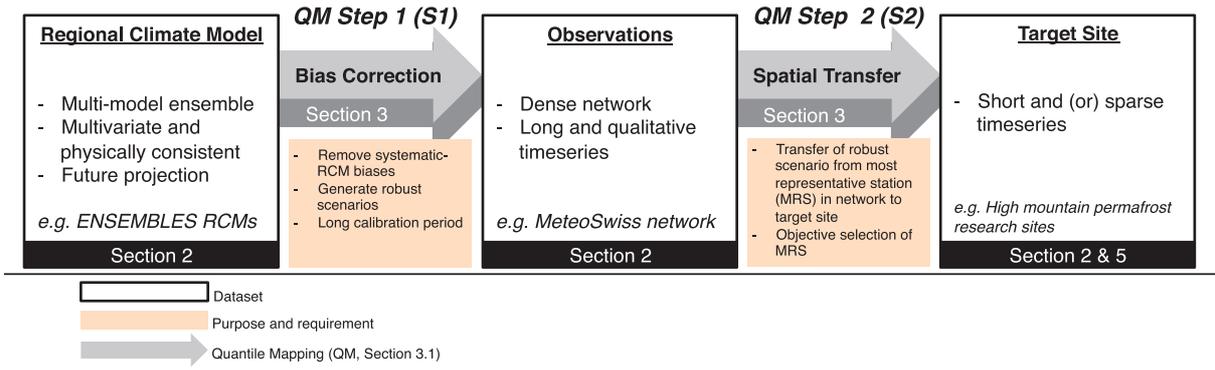


Figure 4. Schematic of the two-step bias correction approach.

transferred to the target site is objectively selected from a network of surrounding stations.

Concerning S2, it should be sufficient to use only few years for calibration between the MRS and a target site, as the interrelation and character of the difference between two similar and proximate locations should be stationary in time (see also Figures 3 and 7 and Section 4.1.). To assure this stationarity assumption also under changing climatic conditions and potential changes in flow regimes (e.g. altered frequencies of weather types), the MRS should be in appropriate proximity to the target site or at least within the same climatic region.

### 3.2.2. Objective selection of a most representative station (MRS)

In our approach, the selection of an MRS for a specific TEMPS target site is based on a combined correlation analysis using observed values at daily resolution for the two key variables mean temperature and precipitation. The time series of both variables at a target site are systematically compared to synchronous measurements from stations within a surrounding long-term observational network (here: the MCH stations). The similarities with the available set of MCH stations are quantified for each variable separately and are finally combined for the selection of an MRS. The station network should ideally surround the target site in close proximity (horizontal distance) and topographic height (vertical distance), and ideally sample characteristic climatic conditions. As a result, the network should possess stations with reasonably large correlations to avoid a misleading selection.

Correlation coefficients form the backbone of our proposed objective selection procedure for various reasons. Correlation is a measure of similarity that is independent of biases in mean or variance but explicitly addresses similarities in the temporal structure, namely the daily correspondence of the respective time series. Further, QM as envisaged bias correction approach implicitly corrects for errors in mean and variance, but not for errors in the daily correspondence. Therefore, errors in the daily correspondence between an MRS and a target site should be as small as possible.

We consider the Pearson correlation coefficient for temperature. The Spearman rank correlation coefficient is

applied for precipitation, as this variable is not normally distributed and dry days occur. For a given TEMPS site, the analysis reveals for each MCH station ( $s$ ) and for each variable ( $v$ ) a correlation coefficient  $r_{v,s}$ . For each variable, a set of correlation coefficients from all considered MCH stations is summarized in a vector  $\mathbf{r}_v$ . In the next step, each element  $r_{v,s}$  is normalized according to

$$n_{v,s} = (r_{v,s} - \min(\mathbf{r}_v)) / \max(r_v - \min(\mathbf{r}_v)) \quad (2)$$

Thus, the station ( $s$ ) showing the largest (smallest) correlation of variable ( $v$ ) is assigned the value 1 (0). The station with the largest correlation ( $n_{v,s} = 1$ ) is equivalent to being the most representative station for variable  $v$ . If a set of considered stations possesses numerous large values, normalization will inflate the spread within a set of  $r_v$ . It also accentuates effects that, to some extent, can be attributed to the correspondence at the daily timescale rather than the annual cycle, which may contribute to larger absolute values for variables with an annual cycle.

The MRS is then finally selected based on the highest rank within a combined measure, where normalized correlation coefficients for the two considered variables temperature and precipitation are added up to a final score ( $F$ ) for each station ( $s$ ), according to

$$F_s = [(w_{pr} \times n_{pr,s}) + (w_{tas} \times n_{tas,s})] / (w_{pr} + w_{tas}) \quad (3)$$

where  $w_{pr}$  ( $w_{tas}$ ) is a weighting term for the respective variable. Here, we apply the same weights for both variables. Depending on the importance of a variable for a specific end user application, one could alternatively apply different weights, or include merely a single variable, or even include additional variables. The station that yields the highest value  $F$  is referred to as the MRS for a given (TEMPS) target site. In this context, it is compulsive to define one single MRS station for all considered variables. Even though some stations may indicate better agreement for individual variables, it is not recommended to choose different stations for different variables, because this may lead to physical inconsistencies.

The objective selection can also be undertaken using another reasonable selection method. Actually, data interpolation (e.g. Alexander *et al.*, 2006; Donat *et al.*, 2013)

and assimilation approaches in numerical weather prediction operate with similar procedures that account for the spatial correlation structure.

We here stick to the presented selection procedure, as it is easy to implement and produces meaningful results. Moreover, if correlation is large, this indicates a reasonable temporal correspondence of quantiles and hence offers an attractive foundation to spatially transfer data from one location to another using the QM approach.

### 3.2.3. Approximate reconstruction of meteorological conditions

The proposed method additionally offers the possibility to approximately reconstruct observational data for non-measured periods. This enables, for instance, a more robust calibration of impact models.

Instead of transferring bias-corrected climate model data, observations from the MRS for a specific target site can be used within the spatial transfer approach (S2, Figure 4) using QM. The reconstruction period is thus constrained by the availability of data at the MRS. Further, the accuracy of the reconstructed data depends on the considered variable and temporal resolution. In this respect, it is to be noted that the evaluation presented in Section 4.2. implicitly validates the reconstruction performance.

### 3.3. Validation framework

For evaluating the second step (S2) of the two-step QM approach, i.e. the spatial transfer from an MRS to the target site, three different validation strategies are considered and applied to daily data of exemplary station pairs in the period 1981–2012. We do not validate the first step (S1, Figure 4) of the approach, on account of the existing literature that demonstrates the ability and skill of QM in correcting climate model biases at the local scale (Section 1.). A brief description of the used validation techniques, followed by a short explanation of the employed skill measures, is given in the following.

**Leave-one-out approach (LOOA):** Individual predictions for every single year out of the set of 32 years are validated separately with calibration based on the remaining 31 years. This results in 32 independent evaluations from which the average skill is derived.

**Split sample approach (SSA):** The period 1981–2012 is split into quarters, with the first three quarters (24 years, 1981–2004) used for calibration and the last quarter (8 years, 2005–2012) used for validation. As Figure 8 shows, this approach also evaluates whether the method is able to account for long-term trends; the observed trend in temperature induces a warmer validation than calibration period.

**Limited data approach (LDA):** Similar to the SSA, the series is split into the periods 1981–2004 and 2005–2012, with the first period used for calibration and the second for validation. Within the period 1981–2004, different calibration period lengths are tested, starting with 1 and steadily increasing to 23 years with an increment of 1. By random combination of years, 24 different predictions

(calibrations) for each length are validated against observations for the period 2005–2012. For instance, if the length is 3, one realization could consist of the years 1982, 1999, and 2003. The purpose is to simulate both a lack of data availability by using different sample sizes and to quantify uncertainties by using different combinations of calibration years.

#### 3.3.1. Skill scores

The validation focuses on a set of different skill scores. For all scores,  $X_{pred}$  (or  $X_{obs}$ ) is defined as a daily time series of predicted (observed) values in a validation period. Bias describes the simple offset between a predicted and observed time series defined as

$$Bias = \bar{X}_{pred} - \bar{X}_{obs} \quad (4a)$$

where  $\bar{X}$  is the mean value over the validation period. For precipitation, we apply a fractional bias

$$Bias = \frac{\bar{X}_{pred}}{\bar{X}_{obs}} \quad (4b)$$

Correlation coefficients ( $r$ ) are employed to assess the temporal correspondence between predicted and observed values at daily resolution.  $r$  is defined as

$$r = \frac{cov(X_{pred}, X_{obs})}{\sigma(X_{pred}) \sigma(X_{obs})} \quad (5)$$

where  $\sigma$  is the standard deviation and  $cov$  the covariance. Depending on the considered variable, we apply either the Pearson correlation or the Spearman rank correlation coefficients.

The Perkins score ( $PS$ , Perkins *et al.*, 2007) quantifies the distributional agreement between two data sets based on their distributional overlap in terms of common bins between predicted and observed probability density functions (histograms). It is defined as

$$PS = \sum_{i=1}^n \min(h_{pred}(n), h_{obs}(n)) \quad (6)$$

where  $h_{pred}$  or  $h_{obs}$  is the normalized frequency of predicted (observed) values within a bin ( $n$ ). Perfect distributional agreement will yield a value of 1; no overlap of the distributions will result in a value of 0.  $PS$  is sensitive to the choice of the bin size. We apply a bin size of 1 K for temperature and 1 mm/d for precipitation (1% for *hurs*, 1 W/m<sup>2</sup> for *rsds*, 0.1 m/s for wind speed).

The mean absolute error ( $MAE$ ) addresses the mean absolute daily offset between prediction and observation in a validation period.

$$MAE = \frac{1}{n} \sum_{i=1}^n |X_{pred} - X_{obs}| \quad (7)$$

While *bias* (Equations (4a) and (4b)) and  $PS$  (Equation (6)) focus on time-integrating characteristics and  $r$  (Equation (5)) and  $MAE$  (Equation (7)) address

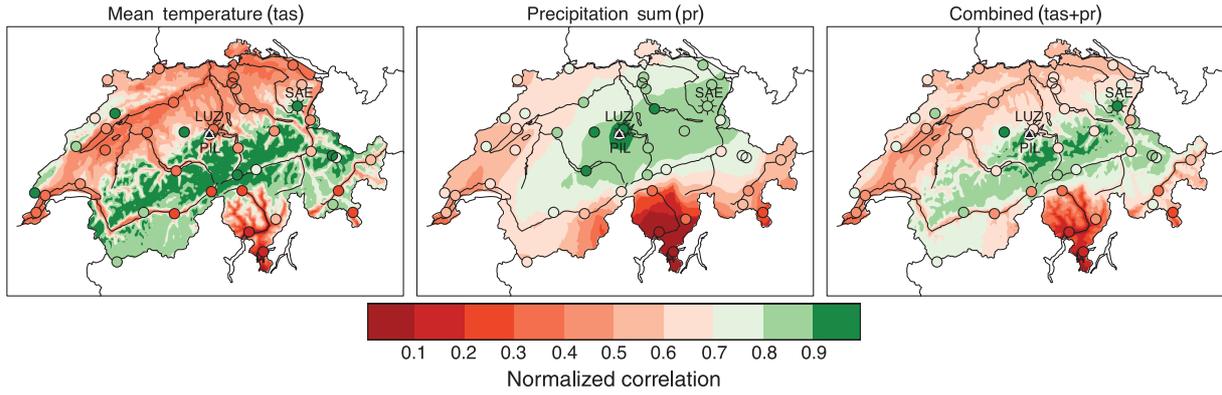


Figure 5. Visualization of the objective selection of the most representative station (MRS) for the example station PIL (Pilatus) using 43 stations and the 2 km observational grid by MCH in period 1981–2012. Left panel shows normalized Pearson correlation for daily mean 2-m temperature, middle panel normalized Spearman rank correlation for the daily precipitation sum and the right panel the combination of normalized temperature and precipitation correlations (see also Equation (3)). Dark green colouring denotes values larger than 0.9 and hence usability for generating data at the site PIL.

similarities in the daily correspondence and may show markedly weaker skill than time-integrating measures.

Furthermore, impact-relevant validation scores are included for mean temperature and precipitation. For precipitation, we define the fractional bias (Equation (4b)) in the wet day frequency ( $fre_{pred}/fre_{obs}$ ) and the 95<sup>th</sup> percentile of all-day precipitation ( $q95_{pred}/q95_{obs}$ ). For temperature, we use the fractional bias (Equation (4b)) of the freezing degree-day sum ( $FDD_{pred}/FDD_{obs}$ ), where  $FDD$  is the sum of negative daily mean temperatures.

#### 4. Evaluation of approach

We evaluate the two-step approach in a multivariate framework and focus on the second step exclusively (Section 3.2. and Fig. 4, S2).

For the evaluation of S2, we chose the station Pilatus (PIL, 2106 m) of the MCH network as an exemplary target site and use observations at the MRS instead of bias-corrected model data in order to validate the spatial transfer. That is, our validation strategy evaluates the ability of reconstructing data at target sites (Section 3.2.3.).

Choosing PIL as a target site is motivated by its centred position within the measurement network of MCH and its challenging location at high altitude in the alpine foothills. Indeed, many sites that aim for robust scenarios also possess similarly challenging locations.

We first illustrate the selection of an MRS for PIL and continue by validating the spatial transfer (S2) from the found MRS Säntis (SAE) to the target site PIL. In addition, results for the transfer from the closest station Lucerne (LUZ) to PIL are shown.

##### 4.1. Selection of a most representative station

The objective selection (Section 3.2.2.) is carried out for PIL based on daily temperature and precipitation observations from 43 surrounding MCH stations in the period 1981–2012. Additionally, we use daily-gridded observations of temperature and precipitation provided at a 2 km

grid for the entire domain of Switzerland for the period 1981–2010 (Frei, 2014). The gridded data are used to illustrate the spatial specification of representative areas and to highlight all facets of the Swiss climate.

Figure 5 depicts the spatial distribution of normalized correlation values for daily mean temperature (left panel), daily precipitation sum (middle panel) and a combination of both according to Equation (3) (right panel). The latter illustrates the criterion on which the objective selection is based to find representative conditions. In addition, Figure 6 offers a more detailed analysis with normalized and absolute correlation values as a function of the horizontal (left column) and vertical distance (right column) to the target site.

For daily mean temperature (Fig. 5 (left) and Fig. 6a and b), the results show a distinct representativeness for the target site at stations and in areas (grid cells) that are at similar elevation or located at similar topographical settings (mountain ridges situated in the northern Alpine region). The closest station LUZ, which is located in the lower foothills of PIL, shows weaker skill than a majority of stations from the analysed network, which are more distant but show a smaller elevation difference. This is primarily attributable to weather types that regularly invert or lower elevation-dependent temperature gradients (e.g. inversions). The annual cycle clearly dominates a large fraction of the daily temperature correlation (large absolute correlation values at right x-axis labelling in Fig. 6a and b).

The results for daily precipitation (Fig. 5 (middle) and Fig. 6c and d) indicate a much stronger relation between representativeness and proximity to the target site. Moreover, the location in a similar climatic region has a major influence on the degree of similarity. In Switzerland, the specification of similar climatic regions is largely governed by topography, the distribution of mountain ridges and the orientation towards atmospheric flow. In the example presented for PIL, the closest station (LUZ) is most correlated in terms of precipitation. In contrast to temperature, precipitation similarities are not governed by elevation. Areas in close proximity and in similar orientation towards

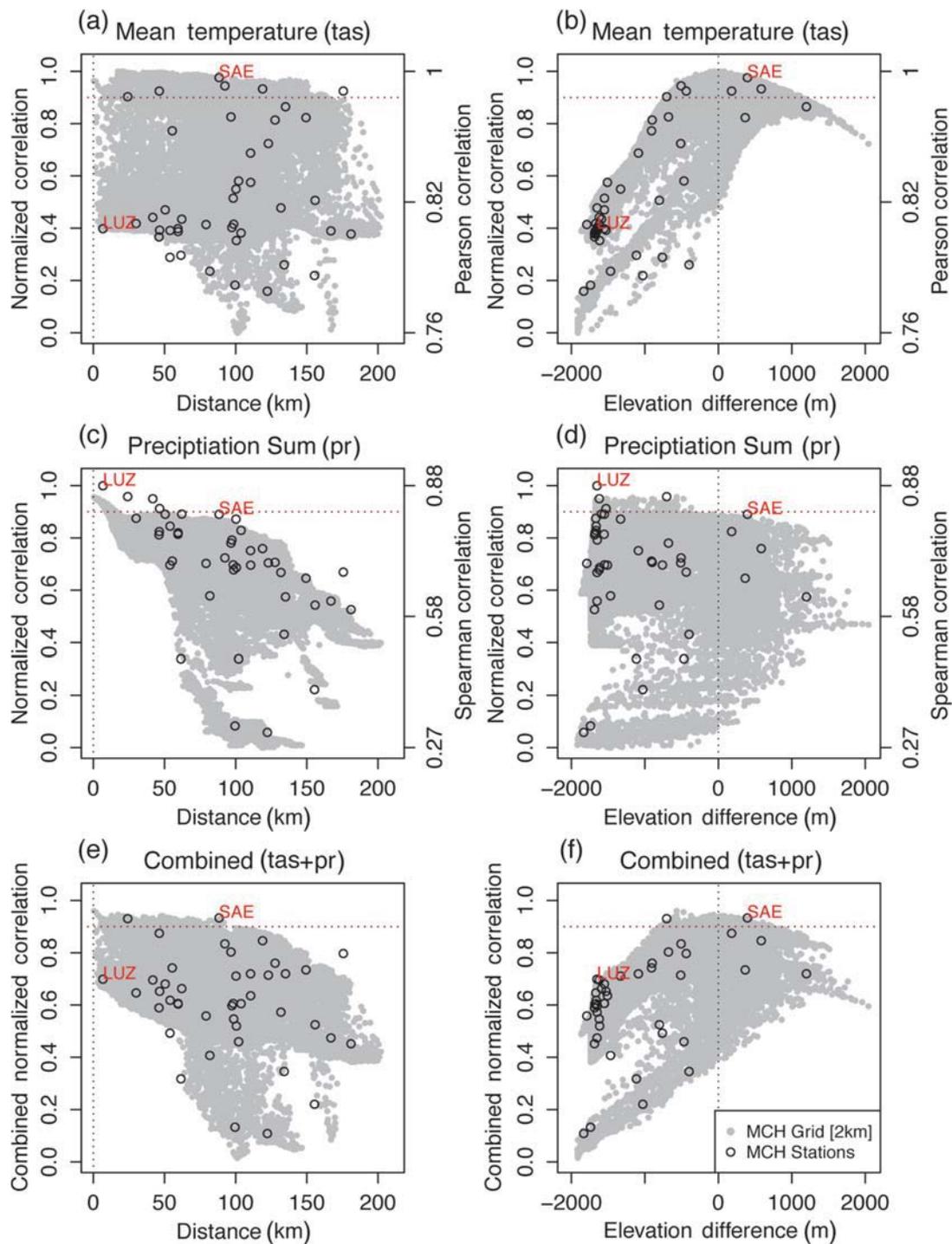


Figure 6. Detailed analysis of objective selection procedure. The left column depicts results in dependence of the distance to the exemplary target site PIL and the right panel results in dependence of elevation difference with respect to PIL. The top row depicts results for temperature, the middle row for precipitation and the bottom row the combined results. The closest (LUZ) and the most similar (SAE) stations are marked in red. The analysis is based on daily values in period 1981–2012.

westerly (moist) atmospheric flow show the largest correlations. This kind of analysis comprises large uncertainties in summer in which daily precipitation occurrence can be highly localized due to the mostly convective type of precipitation.

Figure 5 (right) shows the spatial patterns of representativeness based on the combination of temperature and

precipitation (Equation (3)). The station (grid cells) with the largest value(s) is (are) selected as the most representative station (grid cells) for the target site. Ultimately, the selection based on a mixture of similarity in temperature and precipitation finds a station within a similar altitudinal setting (vertical difference), which is not too distant (horizontal distance). In this particular example, the station

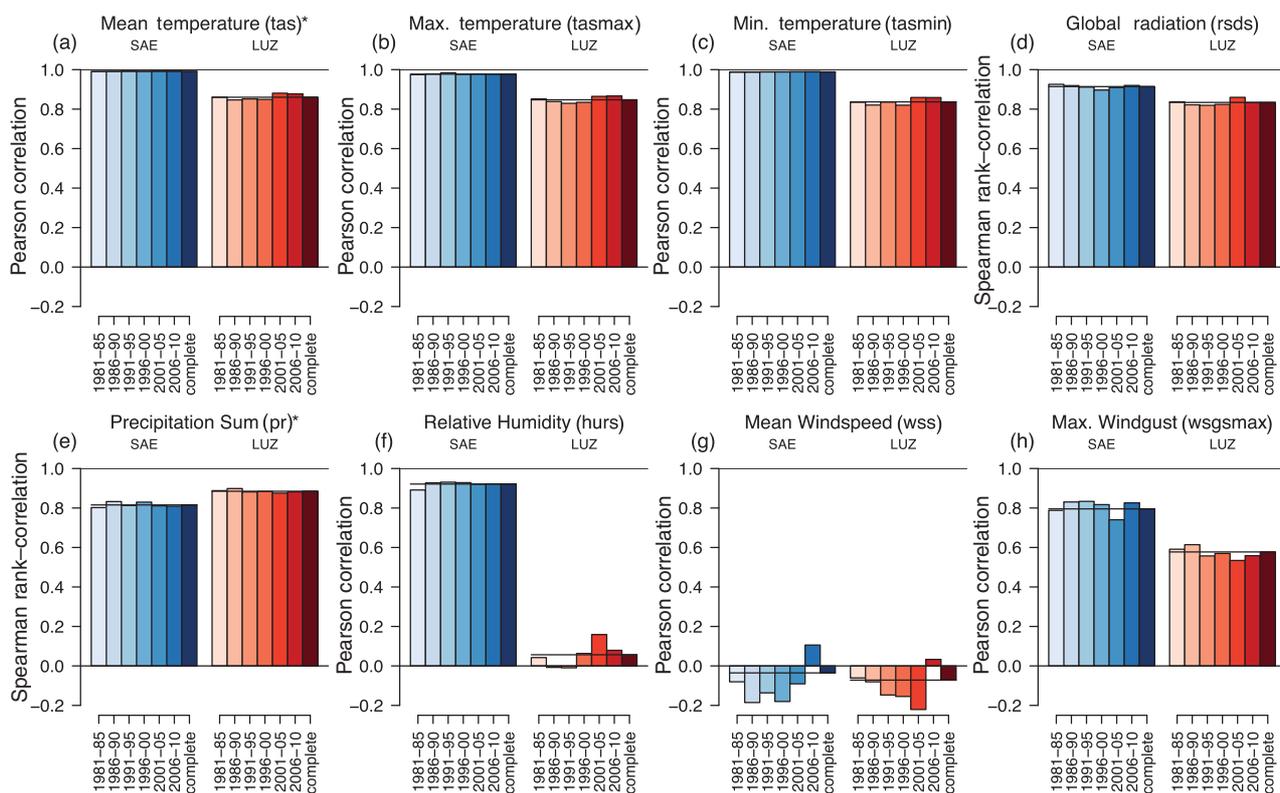


Figure 7. Temporal stability and multivariate analysis of correlations between target site PIL and the most similar (SAE, blue) and the closest station (LUZ, red). x-axis tick marks denote the analysed period, whereas ‘complete’ represents the 32-year period 1981–2012 and all the other bars 5-year periods, respectively.

SAE is found to be the MRS for the exemplary target site PIL. The horizontal distance of SAE to PIL is 88.4 km and the vertical difference is +396 m. The proposed selection procedure is well supported by results of an analysis that groups similar stations from the MCH network (Begert, 2008).

In Figure 7, we provide additional information on the correlation between PIL and both the most similar station SAE and the closest station LUZ. Information on the temporal stability of the found inter-station relations is provided for eight variables. Statements are based on results from consecutive blocks of 5 years in the period 1981–2012. The results indicate that multivariate station interrelations found in a short period of 5 years are stationary and virtually equal to the results based on an inter-comparison using 32 years of data (‘complete’). This indicates that only few years of pairwise data are sufficient to characterize long-term inter-station relations in an adequate manner.

Figure 7 also indicates that the MRS (SAE) exhibits more similar conditions for all parameters except precipitation when compared with the closest station LUZ. In terms of mean wind speed, no relationship between PIL and both sites of comparison can be found. This can be attributed to the spatially highly complex specification of wind speed, predominantly governed by topography and local effects that in turn drive the associated thermodynamic processes and channelling effects. As maximum wind speed is governed by the weather

situation, inter-station similarities can be found for daily maximum wind gusts, with better agreement between the two high-altitude sites SAE and PIL. Relative humidity shows very large correlation between SAE and PIL and, in contrast, no relation between LUZ and PIL.

#### 4.2. Spatial transfer

We validate the performance of the spatial transfer to the selected target site PIL (2106 m) for eight variables in a comparative manner from the closest station LUZ (454 m, 6.7 km distance, 1652 m elevation difference) and the MRS SAE. We focus on the period 1981–2012 (Figure 8) and discuss the results for mean temperature and precipitation in detail, but provide figures for the remaining six variables in the supplementary material.

##### 4.2.1. Temperature

For daily mean temperature, presented in Figure 9, the transfer from SAE to PIL distinctly outperforms the transfer from LUZ. This is true for virtually all cases and validation strategies. The spread between different realizations (number of years) of the limited data approach (LDA) and hence the uncertainties are also distinctly smaller when transferring data from SAE, instead of LUZ, to PIL. While the outcomes for time-integrating measures (*bias* and *PS*) are reasonably good for both stations, even for small sample sizes, clear differences can be seen in the other three measures that depend on the daily correspondence.

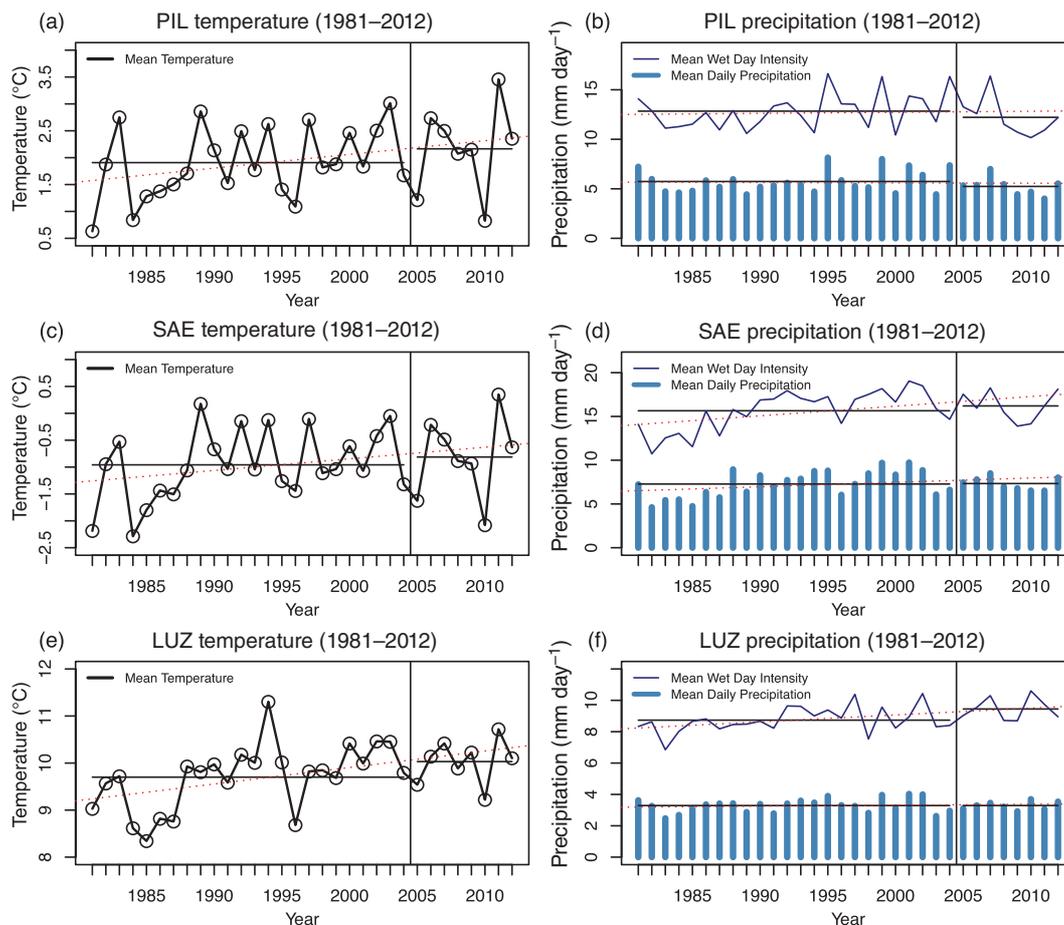


Figure 8. Evolution of annual mean temperature (left column) and annual mean daily precipitation and wet day intensity (right column) in the calibration period 1981–2012 at the exemplary target site PIL (top row), the most similar station SAE (middle row) and the closest station LUZ (bottom row). Horizontal continuous lines depict time slice average values for the first three quarters and the last quarter of the period. Vertical horizontal lines separate the last from the three preceding quarters as they are used for validation and calibration in methods (b) and (c). Stippled red lines indicate linear trends.

The *MAE* is systematically larger for LUZ ( $>2.5^{\circ}\text{C}$ ) and the differences can be more than five to six times larger than those for SAE ( $\sim 0.5^{\circ}\text{C}$ ). Results are remarkable for SAE as the *MAE* amounts to only approximately  $0.5^{\circ}\text{C}$  even when using only very few years. The spatial transfer from SAE to PIL shows very good skill for the index *FDD*, which is very well reproduced with only few years of data used for calibration. In contrast, the transfer from LUZ to PIL shows an extraordinary weak skill for *FDD* and an overestimation of up to 20% even for long calibration periods ( $>20$  years). Throughout all validation strategies and skill scores, it is obvious that a better skill is achieved when transferring data from SAE to PIL. In general, a spatial transfer from SAE to PIL based on only few years outperforms most long-term calibrations from LUZ to PIL.

#### 4.2.2. Precipitation

Figure 10 depicts the validation for precipitation. In contrast to mean temperature, better skill can be found for the transfer from the closest station LUZ. However, differences between SAE and LUZ are not as distinct as for temperature and uncertainties are considerably larger for both stations. For SAE and LUZ, the length of a calibration

period should preferably be long, as uncertainties and biases from actually observed values remain large, even for increasing calibration period lengths.

The evaluation shows similar skill for time-integrating measures (*bias*, *PS*, *fre*) for both stations, but larger uncertainty for SAE. For instance, only one year of calibration using SAE can potentially result in an overestimation (underestimation) of precipitation by 65% ( $-25\%$ ). For LUZ, the respective number is 30% ( $-25\%$ ). For wet day frequency, both stations show very good skill. Even when considering a small sample size, the number of rainy days is captured very well. Skill measures that focus on the daily agreement between prediction and observation consistently show better skill for LUZ, which is an implication of its proximity to the target site. An improvement for long calibration periods, however, is not clearly visible and uncertainties remain large.

Obviously, precipitation is the variable associated with the largest uncertainties and the weakest skill in the framework of matching observations at a potential target site. Owing to the large spatial and temporal variability of daily precipitation, it surely is the most critical variable within the presented approach (Fig. 4, S2). However, in

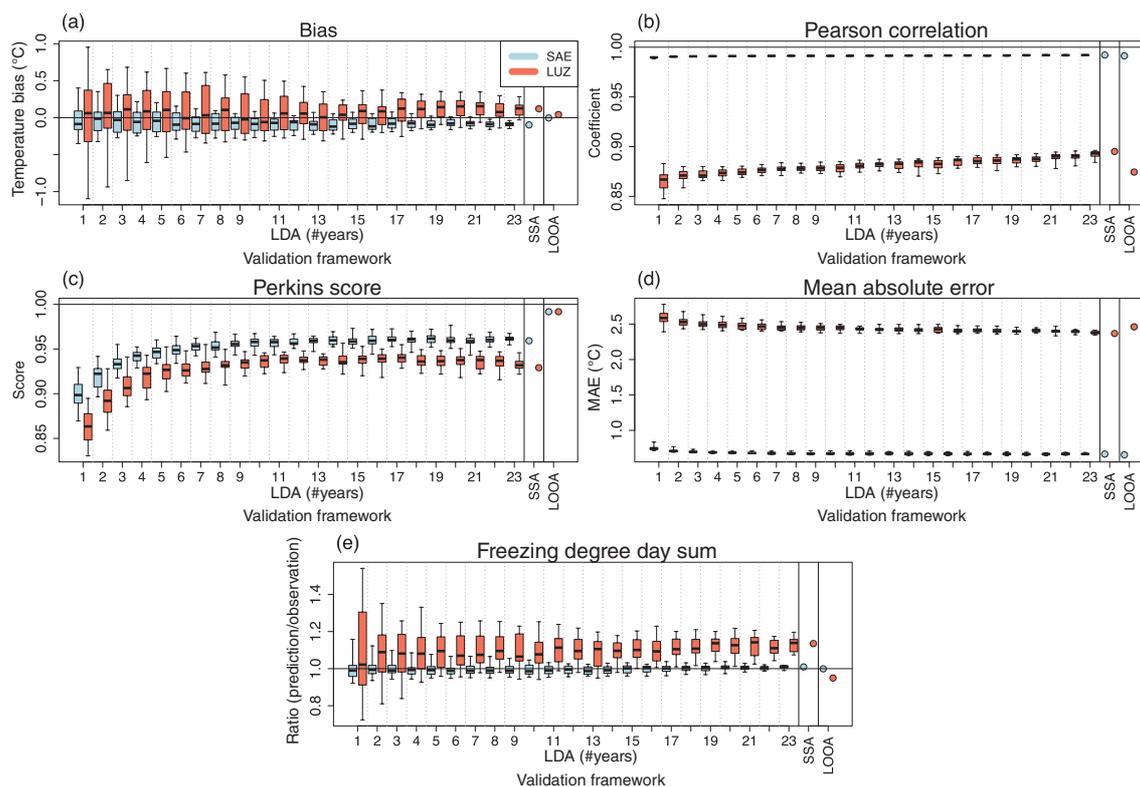


Figure 9. Validation of spatial transfer of daily 2-m temperature based on quantile mapping from stations SAE (most similar, blue) and LUZ (closest, red) to the target site PIL. The x-axis denotes different validation techniques (Section 3.3.). LDA (limited data approach) with the number of considered years for calibration (#years) increasing from left (one) to right (23). The boxplots indicate the range of 24 re-sampling estimates. SSA (split sample approach) and LOOA (leave-one-out approach). The individual panels present different validation scores (Section 3.3.1.): (a) bias, (b) correlation, (c) Perkins score, (d) mean absolute error and (e) ratio of predicted and observed freezing degree-day sum. For all approaches and scores, the calibration period is 1981–2004 and the validation period is 2005–2012.

a climate scenario context, the exact match of the daily correspondence (represented by  $r$ ,  $MAE$ ) between two sites is not important compared to the ability to match time-integrated characteristics (represented by  $bias$ ,  $fre$ ,  $q95$ ,  $PS$ ). Nonetheless, the results indicate clear drawbacks with respect to data reconstruction. The associated uncertainties are large at the daily timescale, but decrease when considering monthly, seasonal or annual values (not shown).

## 5. Application of the approach to a high-altitude permafrost research site

We exemplify some results from an end user–tailored application based on the presented two-step approach. Results are shown for the site Schilthorn, which is one of the 12 TEMPS sites (Fig. 1). Within the framework of the present study, scenarios have been generated for all of the 12 sites and will contribute to a follow-up study that investigates the future evolution of mountain permafrost in Switzerland.

The site Schilthorn is situated at 2913 m and measurements consist of almost 14 years of data in the period 1999–2013. Depending on the variable, missing values occur frequently. For temperature, humidity and radiation ~12% missing values are found, for mean wind speed

~28% and for daily maximum wind gusts ~89%. Precipitation observations for the period 1981–2010 are taken from the 2 km grid cell of the observational RhiresD (see section 2) grid located over Schilthorn. Related uncertainties, discussed in Section 2, should be carefully taken into account.

Previous studies focusing on the Schilthorn site have addressed the sensitivity of permafrost to climate (Marmy *et al.*, 2013), and the 21<sup>st</sup> century evolution of permafrost using statistically downscaled climate model projections (Scherler *et al.*, 2013). Scherler *et al.* (2013) used elevation-corrected, delta-change-based (Bosshard *et al.*, 2011; CH2011, 2011) scenarios from the MCH station Interlaken (577 m, 12 km distance).

### 5.1. Most representative MCH station

The objective selection of the most representative operational long-term MCH station for the target site Schilthorn is presented in Figure 11. Results are shown in a similar manner as in Figure 5. For temperature, stations at a similar topographical situation appear most representative for Schilthorn. It is noticeable that the station Weissfluhjoch (WFJ, 152.9 km distance, 223 m elevation difference) shows the largest association, followed by other high-elevation stations. In terms of precipitation, stations within proximity and the same climatological region

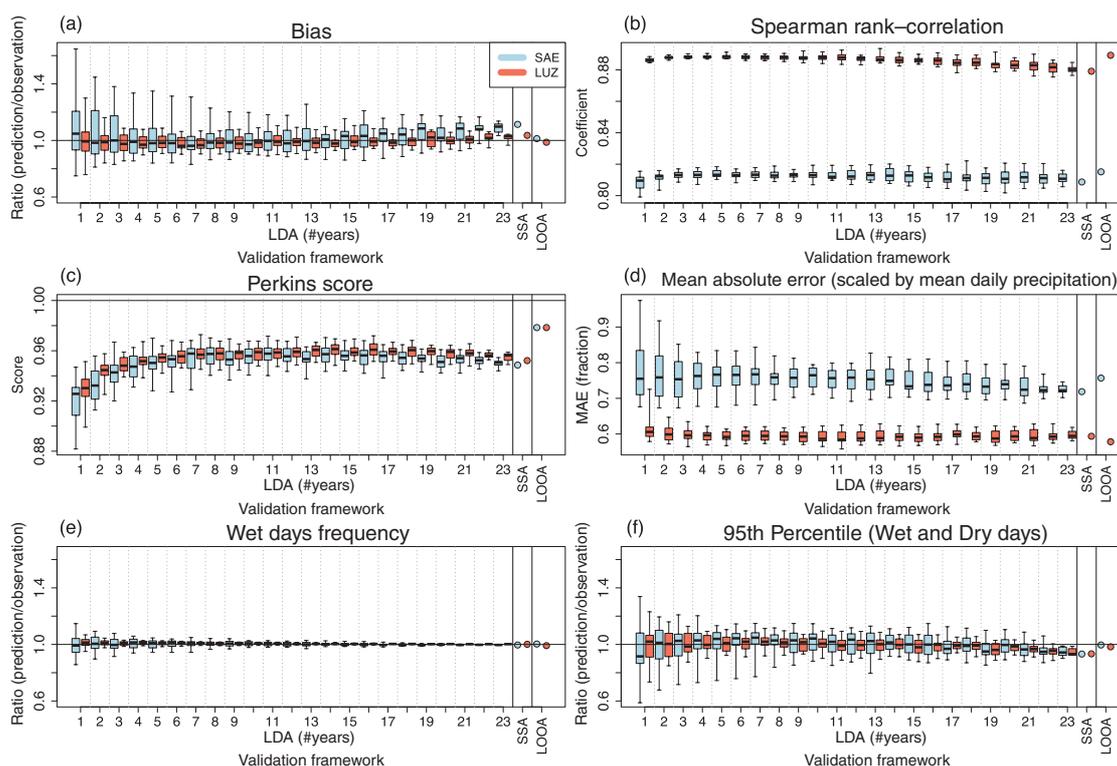


Figure 10. Validation of spatial transfer of daily precipitation sum based on quantile mapping from stations SAE (most similar, blue) and LUZ (closest, red) to the target site PIL. The x-axis denotes different validation techniques (Section 3.3.). LDA (limited data approach) with the number of considered years for calibration (#years) increasing from left (one) to right (23). The boxplots indicate the range of 24 re-sampling estimates. SSA (split sample approach) and LOOA (leave-one-out approach). The individual panels present different validation scores (Section 3.3.1.): (a) bias, (b) correlation, (c) Perkins score, (d) mean absolute error, (e) ratio of predicted and observed wet day frequency and (f) ratio of predicted and observed 95th percentile of precipitation (wet and dry days). For all approaches and scores, the calibration period is 1981–2004 and the validation period is 2005–2012.

appear most representative and the nearby station Interlaken (12 km) reveals most similar observations. The combination of both (Equation (3)) yields Pilatus (PIL) as the MRS for Schilthorn. The station PIL is at a distance of 56.8 km from Schilthorn, at 2106 m elevation and is ranked 5<sup>th</sup> (out of 43) in terms of temperature and 4<sup>th</sup> (out of 43) in terms of daily precipitation similarities. Both, PIL and Schilthorn represent mountain peak locations along the northwestern Alpine slopes.

## 5.2. Climate scenarios

### 5.2.1. Temperature

Multi-model ensemble projections of impact-relevant temperature indices for the period 1960–2099 are presented in Figure 12, which depicts the ensemble spread and also highlights one individual bias-corrected RCM simulation (ETH-HadCM3Q0). In addition, observed values at the target site are shown. The ensemble spread is an approximate estimate of model uncertainty, highlighting the importance of using several instead of only one single ensemble member in impact studies. For instance, mean temperature (Fig. 12a) is projected to lie within a range between  $-1.3^{\circ}\text{C}$  and  $+2.4^{\circ}\text{C}$  in the period 2080–2099. As a consequence of warming, all indices experience distinct trends in the course of the 21<sup>st</sup> century, except for frost transition days ( $tasmin < 0^{\circ}\text{C}$  &  $tasmax > 0^{\circ}\text{C}$ ).

Referring to 20-year-average best estimates and comparing the period 2080–2099 (2040–2059) with 1990–2009, the annual mean temperatures are projected to rise by  $3.0^{\circ}$  ( $1.5^{\circ}$ ) from  $-2.8^{\circ}\text{C}$  to  $0.2^{\circ}\text{C}$  ( $-1.3^{\circ}\text{C}$ ). The freezing degree-day sum reduces by  $-40\%$  ( $-22\%$ ). While the number of days with frost transitions is not projected to change, the number of ice days ( $tasmax < 0^{\circ}\text{C}$ ) reduces from 174 to 130 (151) at the expense of ice-free days ( $tasmin > 0^{\circ}\text{C}$ ), which increase from 80 to 131 (107) days per year. Regarding extremes, subjectively defined for Schilthorn as a day with  $tasmax > 15^{\circ}\text{C}$ , which, at present conditions (1990–2009) appears once every three years, a massive increase to a number of 14 (4) days per year is projected.

It is noticeable, that all the present-day observations lie within the range of model simulations for present-day conditions. This proves that the approach produces reasonable estimates at the local scale, even for impact-relevant indices.

### 5.2.2. Precipitation

Projections of seasonal precipitation characteristics are visualized in Figure 13. At seasonal scale, the figure presents changes in different precipitation indices: mean precipitation, the frequency of wet days and the 99<sup>th</sup> percentile of (all day) precipitation. Changes are expressed

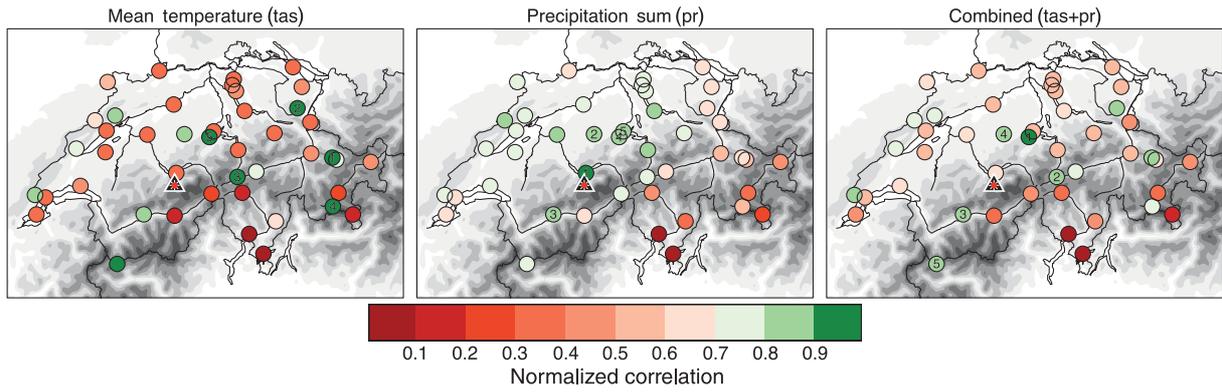


Figure 11. Objective selection of a most similar MCH station for the permafrost research site Schilthorn (black triangle with red star). The right panel shows the combined result, the left panel shows the results for temperature and the middle panel, results for precipitation. The most similar stations appear in green colour and are numbered according to the degree or representativeness.

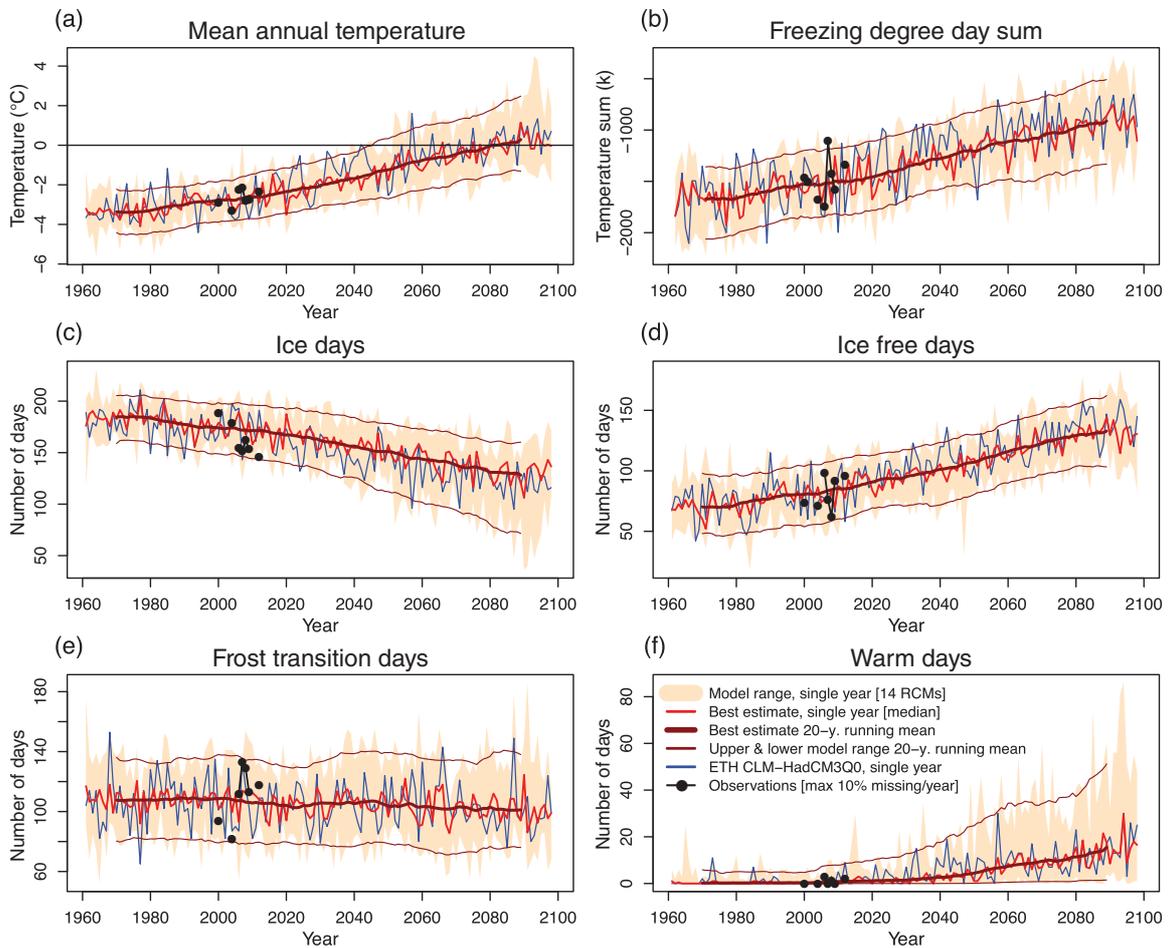


Figure 12. Multi-model climate scenario for temperature-based indices for site Schilthorn. Fourteen RCM simulations are taken into account and MCH station Pilatus (PIL) is used for the intermediate step. Black points denote observed values at Schilthorn.

in percentage, extracted from the bias-corrected local scenario, considering 30-year moving averages with respect to the reference period 1980–2009. As for temperature (Fig. 12), characteristics in terms of not only projected change but also projection uncertainty in a multi-model framework can be deduced. For precipitation, uncertainties, expressed as inter-model spread, are obviously larger than for temperature.

Models agree on mean precipitation (left column) to decrease in summer (JJA) and to slightly increase in spring (MAM). There are no apparent signals in winter (DJF) and fall (SON). Long-term signals in fall tend towards a reduction of precipitation totals. However, uncertainties in fall are larger than in other seasons. For instance, the median change signal for the period centred around 2040 (2080) relative to the one centred around 1994 is  $-11\%$

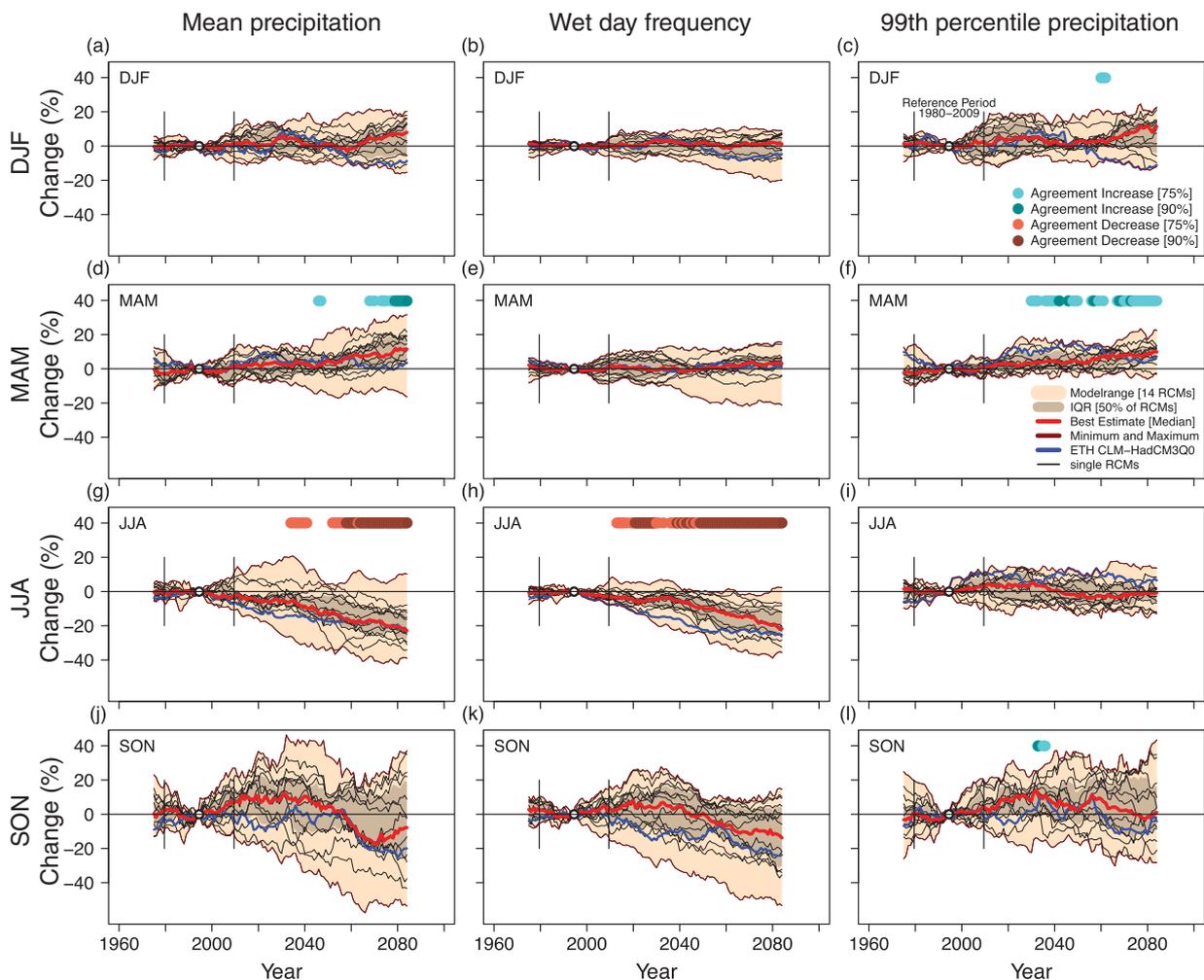


Figure 13. Multi-model climate scenario for precipitation-based indices (columns, left to right) at seasonal scale (rows, top to bottom) for site Schilthorn. Fourteen RCM simulations are taken into account and MCH station Pilatus (PIL) is used for the intermediate step. Changes are expressed as percentage change of a 30-year moving window with respect to reference period 1980–2009. Coloured points at the top of each panel indicate model agreement on a projected sign of change.

(+10%), with individual ensemble members ranging from  $-50\%$  to  $+33\%$  ( $-34\%$  to  $+42\%$ ). Changes in precipitation totals can be attributed to a complex interplay between projected changes in precipitation frequency and intensity (here, presented by the 99<sup>th</sup> percentile). Overall, the results are consistent with the domain-mean analysis presented by Rajczak *et al.* (2013). In summer (and to a certain degree also in fall), models agree on a prominent decrease in the number of wet days. While in winter and spring, the number of wet days is projected to remain stable and precipitation tends to intensify.

## 6. Discussion and conclusion

We presented a method that allows for the generation of robust climate scenarios at sites with sparse and short observational data coverage. The approach relies on two consecutive applications of the quantile mapping (QM) procedure, with climate model simulations first being bias corrected to long-term measurement stations and then spatially transferred to the target site (see Fig. 4). Both steps

aim for the most appropriate translation of climate model simulations to match the desired accuracy and scales. In addition, the presented method also incorporates the possibility for approximately reconstructing observations in unobserved periods.

The basic concept of improving the robustness of climate projections at a sparsely observed site by taking an intermediate station with long-term measurements into account is transferable to other empirical-statistical downscaling (ESD) methods and can serve future developments in impact research.

The approach facilitates (1) the construction of robust GCM-RCM-chain based local-scale climate scenarios, (2) a more accurate calibration of impact models and (3) an assessment of the sensitivity of systems towards climate variability and change at locations that usually suffer from a lack of data (e.g. mountain areas).

The proposed method is validated using a set of multivariate long-term measurements from operational weather stations in Switzerland. Exemplary results from an end user application targeting the evolution of mountain

permafrost in Switzerland are also shown, which underline the applicability of the approach in climate impact research. Validation results reveal a satisfying performance in a multivariate and challenging framework in a region of complex topography and large spatial climate variability. Further, the results show that the method can also be applied at extreme locations.

Uncertainties with regard to the application of the presented method depend on a number of factors. Beyond the well-known uncertainties relevant for climate projections (Hawkins and Sutton, 2009), the employed bias correction and spatial transfer approaches entail additional sources of uncertainty. First, they can be related to possible non-stationarities of transfer functions under future or changing climatic conditions (Christensen *et al.*, 2008; Boberg and Christensen, 2012; Maraun, 2012; Bellprat *et al.*, 2013). Second, climate change signals are imposed by the location of the MRS considered for RCM bias correction. Finally, the spatial transfer from one point-scale observation to another (S2, Fig. 4) also bears uncertainties, which are of particular relevance for the use of reconstructed data. While methodological uncertainties are small for homogenous variables such as temperature, considerably larger uncertainties arise for spatially and temporally heterogeneous variables such as precipitation and wind speed. The spatial transfer involves uncertainties that approximately scale with the considered timescale and are large for daily data and smaller for temporally aggregated data (e.g. monthly, seasonal, annual). Moreover, uncertainties can be expected to increase when transferring between stations with a distinctly different character, for instance, from a dry to a very wet location. Nonetheless, as time-integrating measures are matching well even for potentially critical variables such as precipitation, the mentioned uncertainties do not degrade the quality of GCM-RCM-based products, but rather that of reconstructed products that aim for an adequate daily correspondence.

In this respect, the objective selection of an MRS also nicely demonstrates that for some cases, in particular for multivariate and temperature-based applications, the closest station might not always be the most appropriate choice of reference.

In summary, the specified two-step QM approach offers the potential to improve climate projections at sparsely observed sites, and it appears useful for impact studies across various research fields. Its application is versatile and not restricted to alpine environments. Further, it offers the potential to reconstruct data in non-measured periods.

## Acknowledgements

We acknowledge the RCM data sets from the EU-FP6 project ENSEMBLES (<http://ensembles-eu.metoffice.com>). We also acknowledge MeteoSwiss for providing observational data, Christoph Frei (MeteoSwiss) for plotting functions in R and Ingo Völsch (University of Zurich) for access and provision of data from

permafrost research sites. This research was partly funded by the Swiss National Science Foundation through the SNSF Sinergia project CRSII2\_136279 “The Evolution of Mountain Permafrost in Switzerland” (TEMPS).

## Supporting Information

The following supporting information is available as part of the online article:

**Figure S1.** Validation of spatial transfer of daily 2-m minimum temperature (tasmin) based on quantile mapping from stations SAE (most similar, blue) and LUZ (closest, red) to the target site PIL. The x-axis denotes different validation techniques (Section 3.3.). LDA (limited data approach) with the number of considered years for calibration (# years) increasing from left (one) to right (23). The boxplots indicate the range of 24 re-sampling estimates. SSA (split sample approach) and LOOA (leave-one-out approach). The individual panels present different validation scores (Section 3.3.1.): (a) bias, (b) correlation, (c) Perkins score, (d) mean absolute error. For all approaches and scores, the calibration period is 1981–2004 and the validation period is 2005–2012.

**Figure S2.** Same as Figure S1 but for daily 2-m maximum temperature (tasmax).

**Figure S3.** Same as Figure S1 but for daily mean relative humidity (hurs).

**Figure S4.** Same as Figure S1 but for daily mean wind-speed (wss).

**Figure S5.** Same as Figure S1 but for daily maximum windgust (wsgsmax).

**Figure S6.** Same as Figure S1 but for daily mean radiation (rsds).

## References

- Addor N, Seibert J. 2014. Bias correction for hydrological impact studies – beyond the daily perspective. *Hydrol. Processes* **28**(17): 4823–4828, doi: 10.1002/hyp.10238.
- Alexander LV, Zhang X, Peterson TC, Caesar J, Gleason B, Klein Tank AMG, Haylock M, Collins D, Trewin B, Rahimzadeh F, Tagipour A, Rupa Kumar K, Revadekar J, Griffiths G, Vincent L, Stephenson DB, Burn J, Aguilar E, Brunet M, Taylor M, New M, Zhai P, Rusticucci M, Vayquez-Aguirre JL. 2006. Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res.* **111**(D5): D05109, doi: 10.1029/2005JD006290.
- Begert M. 2008. Die repräsentativität der stationen im Swiss National Basic Climatological Network (Swiss NBCN). Arbeitsberichte der MeteoSchweiz No. 217, MeteoSwiss, Zurich: Switzerland.
- Bellprat O, Kotlarski S, Lüthi D, Schär C. 2013. Physical constraints for temperature biases in climate models. *Geophys. Res. Lett.* **40**(15): 4042–4047, doi: 10.1002/grl.50737.
- Boberg F, Christensen JH. 2012. Overestimation of Mediterranean summer temperature projections due to model deficiencies. *Nat. Clim. Change* **2**(6): 433–436.
- Boé J, Terray L, Habets F, Martin E. 2007. Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *Int. J. Climatol.* **27**(12): 1643–1655, doi: 10.1002/joc.1602.
- Bosshard T, Kotlarski S, Ewen T, Schär C. 2011. Spectral representation of the annual cycle in the climate change signal. *Hydrol. Earth Sys. Sci.* **15**(9): 2777–2788, doi: 10.5194/hess-15-2777-2011.
- Bosshard T, Carambia M, Goergen K, Kotlarski S, Krahe P, Zappa M, Schär C. 2013. Quantifying uncertainty sources in an ensemble of hydrological climate-impact projections. *Water Resour. Res.* **49**(3): 1523–1536, doi: 10.1029/2011WR011533.

- CH2011. 2011. Swiss Climate Change Scenarios CH2011, C2SM, MeteoSwiss, ETH, NCCR Climate, and OcCC, Zurich, Switzerland.
- Christensen J, Christensen O. 2007. A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Clim. Change* **81**(1): 7–30, doi: 10.1007/s10584-006-9210-7.
- Christensen JH, Boberg F, Christensen OB, Lucas-Picher P. 2008. On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophys. Res. Lett.* **35**(20): L20709, doi: 10.1029/2008GL035694.
- Collins M, Knutti R, Arblaster J, Dufresne J-L, Fichet T, Friedlingstein P, Gao X, Gutowski WJ, Johns T, Krinner G, Shongwe M, Tebaldi C, Weaver AJ, Wehner M (eds). 2013. Long-term climate change: projections, commitments and irreversibility. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: Cambridge, UK and New York, NY.
- Déqué M. 2007. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. *Glob. Planet. Change* **57**(1–2): 16–26, doi: 10.1016/j.gloplacha.2006.11.030.
- Donat MG, Alexander LV, Yang H, Durre I, Vose R, Dunn RJH, Willet KM, Aguilar E, Brunet M, Caesar J, Hewitson B, Jack C, Klein Tank AMG, Kruger AC, Marengo J, Peterson TC, Renom M, Oria Rojas C, Rusticucci M, Sailinger J, Elrayah AS, Sekele SS, Srivastava AK, Trewin B, Villarreal C, Vincent LA, Zhai P, Zhang X, Kitching S. 2013. Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: the HadEX2 dataset. *J. Geophys. Res.* **118**(5): 2098–2118, doi: 10.1002/jgrd.50150.
- Flato G, Marotzke J, Abiodun B, Braconnot P, Chou SC, Collins W, Cox P, Driouech F, Emori S, Eyring V, Forest C, Gleckler P, Guilyardi E, Jakob C, Kattsov V, Reason C, Rummukainen M (eds). 2013. Valuation of climate models. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: Cambridge, UK and New York, NY.
- Fowler HJ, Blenkinsop S, Tebaldi C. 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.* **27**(12): 1547–1578, doi: 10.1002/joc.1556.
- Frei C. 2014. Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. *Int. J. Climatol.* **34**(5): 1585–1605, doi: 10.1002/joc.3786.
- Gudmundsson L, Bremnes JB, Haugen JE, Engen-Skaugen T. 2012. Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods. *Hydrol. Earth Syst. Sci.* **16**(9): 3383–3390, doi: 10.5194/hess-16-3383-2012.
- Gutjahr O, Heinemann G. 2013. Comparing precipitation bias correction methods for high-resolution regional climate simulations using COSMO-CLM. *Theor. Appl. Climatol.* **114**(3–4): 511–529, doi: 10.1007/s00704-013-0834-z.
- Hartmann DL, Klein Tank AMG, Rusticucci M, Alexander LV, Brönnimann S, Charabi Y, Dentener FJ, Dlugokencky EJ, Easterling DR, Kaplan A, Soden BJ, Thorne PW, Wild M, Zhai PM (eds). 2013. Observations: atmosphere and surface. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: Cambridge, UK and New York, NY.
- Hawkins E, Sutton R. 2009. The potential to narrow uncertainty in regional climate predictions. *Bull. Am. Meteorol. Soc.* **90**(8): 1095–1107, doi: 10.1175/2009BAMS2607.1.
- Keller DE, Fischer AM, Frei C, Liniger MA, Appenzeller C, Knutti R. 2014. Stochastic modelling of spatially and temporally consistent daily precipitation time-series over complex topography. *Hydrol. Earth Syst. Sci. Discuss.* **11**(7): 8737–8777, doi: 10.5194/hessd-11-8737-2014.
- Kirtman B, Power SB, Adedoyin JA, Boer GJ, Bojariu R, Camilloni I, Doblans-Reyes FJ, Fiore AM, Kimoto M, Meehl GA, Prather M, Sarr A, Schär C, Sutton R, van Oldenborgh GJ, Vecchi G, Wang HJ (eds). 2013. Near-term climate change: projections and predictability. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: Cambridge, UK and New York, NY.
- Kotlarski S, Keuler K, Christensen OB, Colette A, Déqué M, Gobiet A, Goergen K, Jacob D, Lüthi D, van Meijgaard E, Nikulin G, Schär C, Teichmann C, Vautard R, Warrach-Sagi K, Wulfmeyer V. 2014. Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model Dev.* **7**: 1297–1333, doi: 10.5194/gmd-7-1297-2014.
- van der Linden P, Mitchell J. 2009. ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. EU FP6 Contract number 505539, Met Office Hadley Centre, Exeter, UK, 160 pp.
- Maraun D. 2012. Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophys. Res. Lett.* **39**(6): L06706, doi: 10.1029/2012GL051210.
- Maraun D. 2013. Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. *J. Clim.* **26**(6): 2137–2143, doi: 10.1175/JCLI-D-12-00821.1.
- Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themeßl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiery-Eich I. 2010. Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **48**(3): RG3003, doi: 10.1029/2009RG000314.
- Marmy A, Salzmänn N, Scherler M, Hauck C. 2013. Permafrost model sensitivity to seasonal climatic changes and extreme events in mountainous regions. *Environ. Res. Lett.* **8**(3): 035048, doi: 10.1088/1748-9326/8/3/035048.
- Mearns LO, Gutowski W, Jones R, Leung R, McGinnis S, Nunes A, Qian Y. 2009. A regional climate change assessment program for North America. *EOS Trans. AGU* **90**(36): 311, doi: 10.1029/2009EO360002.
- Meehl GA, Covey C, Taylor KE, Delworth T, Stouffer RJ, Latif M, McAvaney B, Mitchell JFB. 2007. THE WCRP CMIP3 multimodel dataset: a new era in climate change research. *Bull. Am. Meteorol. Soc.* **88**(9): 1383–1394, doi: 10.1175/BAMS-88-9-1383.
- MeteoSwiss. 2014. *RhiresD – gridded precipitation data set*. <http://www.meteoschweiz.admin.ch/home/klima/vergangenheit/klimanormwerte/normwert-karten.subpage.html/de/data/products/2014/raeumliche-daten-niederschlag.html> (accessed 20 January 2015).
- Nakicenovic N, Swart R (eds). 2000. *Special Report on Emissions Scenarios*. Cambridge University Press: Cambridge, UK.
- Perkins SE, Pitman AJ, Holbrook NJ, McAneney J. 2007. Evaluation of the AR4 climate models’ simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *J. Clim.* **20**(17): 4356–4376, doi: 10.1175/JCLI4253.1.
- Räisänen J, Räty O. 2013. Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations. *Clim. Dyn.* **41**(5–6): 1553–1568, doi: 10.1007/s00382-012-1515-9.
- Rajczak J, Pall P, Schär C. 2013. Projections of extreme precipitation events in regional climate simulations for Europe and the Alpine Region. *J. Geophys. Res.* **118**(9): 3610–3626, doi: 10.1002/jgrd.50297.
- Räty O, Räisänen J, Ylhäisi J. 2014. Evaluation of delta change and bias correction methods for future daily precipitation: inter-model cross-validation using ENSEMBLES simulations. *Clim. Dyn.* **42**(9–10): 2287–2303, doi: 10.1007/s00382-014-2130-8.
- Salzmänn N, Frei C, Vidale PL, Hoelzle M. 2007. The application of regional climate model output for the simulation of high-mountain permafrost scenarios. *Glob. Planet. Change* **56**: 188–202, doi: 10.1016/j.gloplacha.2006.07.006.
- Scherler M, Hauck C, Hoelzle M, Salzmänn N. 2013. Modeled sensitivity of two alpine permafrost sites to RCM-based climate scenarios. *J. Geophys. Res.: Earth Surf.* **118**(2): 780–794, doi: 10.1002/jgrf.20069.
- von Storch H. 1999. On the use of “inflation” in statistical downscaling. *J. Clim.* **12**(12): 3505–3506, doi: 10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2.
- Taylor KE, Stouffer RJ, Meehl GA. 2011. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**(4): 485–498, doi: 10.1175/BAMS-D-11-00094.1.
- Teutschbein C, Seibert J. 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies: review and evaluation of different methods. *J. Hydrol.* **456–457**(0): 12–29, doi: 10.1016/j.jhydrol.2012.05.052.
- Themeßl M, Gobiet A, Leuprecht A. 2011. Empirical-statistical downscaling and error correction of daily precipitation from regional

- climate models. *Int. J. Climatol.* **31**(10): 1530–1544, doi: 10.1002/joc.2168.
- Thiemeßl M, Gobiet A, Heinrich G. 2012. Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. *Clim. Change* **112**(2): 449–468, doi: 10.1007/s10584-011-0224-4.
- Wilcke R, Mendlik T, Gobiet A. 2013. Multi-variable error correction of regional climate models. *Clim. Change* **120**(4): 871–887, doi: 10.1007/s10584-013-0845-x.
- Wong G, Maraun D, Vrac M, Widmann M, Eden M, Kent T. 2014. Stochastic model output statistics for bias correcting and downscaling precipitation including extremes. *J. Clim.* **27**: 6940–6959, doi: 10.1175/JCLI-D-13-00604.1.
- Wood AW, Leung LR, Sridhar V, Lettenmaier DP. 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim. Change* **62**(1–3): 189–216, doi: 10.1023/B:CLIM.0000013685.99609.9e.