climate

MDPI

Article

# Intercomparison of Univariate and Joint Bias Correction Methods in Changing Climate From a Hydrological Perspective

**Olle Räty [1,*], Jouni Räisänen [1], Thomas Bosshard [2] and Chantal Donnelly [2]**

[1] Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, 00100 Helsinki, Finland; jouni.raisanen@helsinki.fi
[2] Swedish Meteorological and Hydrological Institute (SMHI), 60176 Norrköping, Sweden; thomas.bosshard@smhi.se (T.B.); chantal.donnelly@bom.gov.au (C.D.)
* Correspondence: olle.raty@helsinki.fi; Tel.: +358-504-160-500

Received: 26 February 2018; Accepted: 23 April 2018; Published: 26 April 2018


**Abstract:** In this paper, the ability of two joint bias correction algorithms to adjust biases in daily mean temperature and precipitation is compared against two univariate quantile mapping methods when constructing projections from years 1981–2010 to early (2011–2040) and late (2061–2090) 21st century periods. Using both climate model simulations and the corresponding hydrological model simulations as proxies for the future in a pseudo-reality framework, these methods are inter-compared in a cross-validation manner in order to assess to what extent the more sophisticated methods have added value, particularly from the hydrological modeling perspective. By design, bi-variate bias correction methods improve the inter-variable relationships in the baseline period. Cross-validation results show, however, that both in the early and late 21st century conditions the additional benefit of using bi-variate bias correction methods is not obvious, as univariate methods have a comparable performance. From the evaluated hydrological variables, the added value is most clearly seen in the simulated snow water equivalent. Although not having the best performance in adjusting the temperature and precipitation distributions, quantile mapping applied as a delta change method performs well from the hydrological modeling point of view, particularly in the early 21st century conditions. This suggests that retaining the observed correlation structures of temperature and precipitation might in some cases be sufficient for simulating future hydrological climate change impacts.

**Keywords:** regional climate modeling; hydrological modeling; bias correction; multivariate; pseudo reality

## 1. Introduction

In recent years, bias adjustment has become the de facto standard for preprocessing global (GCM) and regional (RCM) climate model simulations for climate change impact studies, hydrological modeling being no exception. The use is driven by practical needs. Due to systematic errors in climate model simulations with respect to the observed climate, GCM and RCM output usually cannot be directly used in impact modeling, as impact models require unbiased, high-resolution information as their input. This is because of non-linear and threshold processes within impact models. For example, a cold bias in forcing data to a hydrological model could lead to an impact result indicating no change in snow depths if the cold bias kept temperatures below 0 degrees.

Numerous methods belonging to so-called model output statistics (MOS) have been developed to adjust biases in temperature and precipitation data from climate models. These range from

simple scaling of time-mean climate to more sophisticated methods addressing biases in the daily variability. This group also covers the widely used quantile mapping techniques of which there are a number of different variations [1–8]. Studies have illustrated that bias correction methods are able to reduce biases in climate model output [6] and also to provide noticeable improvements to hydrological simulations in the present-day climate [9,10]. However, most of these methods are restricted to the independent adjustment of biases in the marginal aspects of GCM-RCM simulations and do not take biases in inter-variable correlation structures into account. For example, studies have indicated that highest precipitation intensities co-occur with high surface temperatures in winter, as indicated by Clausius–Clapeyron relation, while mostly negative relationships between temperature and precipitation have been observed in summer in Europe [11,12]. In case a GCM-RCM has difficulties to reasonably capture such relationships, a bias correction method that does not explicitly take inter-variable correlations into account might not be sufficient for certain applications such as hydrological modeling.

To address this issue, different types of bi- and multivariate bias correction algorithms have been recently proposed [13–17]. These studies have given evidence that jointly bias correcting multiple variables improves the multivariate aspects of bias corrected model simulations when compared against the observed climate, and might outperform their univariate counterparts in further applications, such as in the calculation of Canadian Forest Fire Weather Index [17]. However, most of the intercomparison studies have concentrated on evaluating the relative performance of bias correction methods in the present-day climate, which does not inform on their ability to predict climate variables in changing climatic conditions. In other words, it is not known how well the adjustment of inter-variable correlations, which is inherently constrained by biases in the present-day, is able to capture the inter-variable correlations in the future climate. This information is crucial for reliably assessing potential climate change impacts, particularly as concerns have been expressed on the shortcomings and potentially unjustified use of bias correction in non-stationary conditions [18,19].

Due to the lack of an observational basis, surrogate data emulating the future observations has been proposed to be used as proxy data (hereafter referred to as pseudo-realities) to assess the ability of bias adjustment to improve projections in a changing climate. The pseudo-reality approach has been used in recent studies [20–23] and was also considered in the European Concerted Research Action ES1102 VALUE (Validating and Integrating Downscaling Methods for Climate Change Research) framework as an important, although not sufficient step, when evaluating bias adjustment method performance [24].

Most of the pseudo-reality studies have concentrated on the analysis of the application of bias adjustment directly to climate model output, which does not give direct information on their usability to construct future projections for the purposes of hydrological climate change impact studies. One of the first attempts to extend the pseudo-reality approach to hydrological simulations was made by Velázquez et al. [25], whose study evaluated implications of non-stationarity to bias correction for future conditions and how they affect the estimation of future changes in river discharges. Their results showed that although monthly mean river discharges were improved in some cases after bias correcting the hydrological model input, biases still remained in the results. In their study the pseudo-reality approach was applied without taking pseudo-reality biases in the present-day climate into account. If a hydrological model is sensitive to absolute biases in climate model outputs, the hydrological model behavior and its response to the projected changes might be unrealistic, which would hamper the evaluation of bias adjustment methods in the pseudo-reality framework. Furthermore, the study used only two GCM-RCM combinations and one bias correction method that did not take inter-variable correlations directly into account in the bias correction step. From the hydrological modeling perspective, a physically plausible description of co-variations of temperature and precipitation might be important to reasonably describe the surface fluxes such as evapotranspiration and processes affecting water stored in soil and snow pack, which together regulate the river discharge generation. It is also important to assess the performance of the hydrological modeling of low frequency impacts

(e.g., high and low flows) as these impacts are often of interest to users, but could be subject to different biases than mean flows.

Here, we extend the study of Velázquez et al. [25] to assess the relative performance of four bias adjustment methods from the hydrological impact modeling perspective. More specifically, the aim is to address the following questions:

1. How does the relative performance of bias adjustment methods vary when assessing them from the perspective of the impacts of climate change on different hydrological variables rather than from climate modeling perspective?
2. What is the added value of bias correcting inter-variable relationships between daily mean temperature and precipitation in comparison to the adjustment of their marginal distributions only?

We use five GCM-RCMs produced in the European branch of the Coordinated Regional Climate Downscaling Experiment (EURO-CORDEX) initiative [26]. In addition to the separate adjustment of temperature and precipitation distributions, two methods [14,17] which take biases in their inter-variable relationships into account are compared against univariate quantile mapping to assess the extent to which hydrological simulations benefit from the additional correction of temperature and precipitation correlations in the changing climate. We perform cross-validation tests in a pseudo-reality framework broadly similar to the one used in Velázquez et al. [25] and extend the analysis by taking into account GCM-RCM biases in the calibration data in order to bring the hydrological simulations closer to the observations (i.e., by bias adjusting the pseudo-reality data).

The paper is structured as follows. In the next section we introduce the GCM-RCM simulations used in this study together with the hydrological model used to conduct the hydrological simulations. In addition, the bias adjustment methods, the pseudo-reality framework and the cross-validation statistics used to assess the relative method performance are also discussed in Section 2. The results are shown in Section 3 and the discussion together with conclusions are presented in Section 4.

## 2. Materials and Methods

### *2.1. Reference Data*

WATCH Forcing Data based on ERA-Interim reanalysis (WFDEI) is used as the reference climatology in real-world illustrations as well as in hydrological modeling exercises for both daily mean temperature and precipitation [27]. In this data set the monthly means of daily mean temperature and precipitation have been adjusted for biases in relation to the large scale gridded observations produced by the Climatic Research unit (CRU) and the Global Precipitation Climatology Centre (GPCC) with subsequent elevation corrections applied to both variables. Rust et al. [28] discuss some of the known issues in WFDEI such as spurious jumps between individual months caused by the bias adjustment procedure. These introduced discontinuities might affect the hydrological simulations, for example, by triggering snow melt unrealistically. However, the effect of the spurious jumps is less severe in the European region than in the tropics and cold regions [28], which facilitates the use of WFDEI in this study.

### *2.2. GCM-RCM Data*

Five high-resolution GCM-RCM simulations (Table 1) produced by the European branch of the Coordinated Regional Climate Downscaling Experiment (EURO-CORDEX) initiative were downloaded from the Earth System Grid Federation (ESGF) database [29]. The cross-validation within a pseudo-reality framework (see Section 2.5) requires all projections to be forced by the same representative concentration pathways (RCP). Here, we chose RCP4.5 as a mid-range emission scenario, which corresponds to end-of-21st century radiative forcing of 4.5 $Wm^{-2}$ [30]. Each of the five model simulations has a different GCM and RCM component. One should note that the models were not

selected based on their performance in the present-day conditions but to ensure that the ensemble members are independent from each other to the extent possible.

The model simulations were re-gridded to the regular 0.125° × 0.125° EURO-CORDEX (EUR-11i) grid using nearest neighbor interpolation before bias adjusting them. The grid box closest to each sub-basin was then used to drive the hydrological model. The difference in the spatial resolution to WFDEI was not taken into account, as most of the results are based on comparisons between the GCM-RCM simulations (see Section 3). For both calibration and validation of the bias adjustment methods, we selected three 32-year periods including a 2-year spin-up period for the hydrological model runs. After excluding the spin-up period, years 1981–2010 were used as the baseline period, while years 2011–2040 and 2061–2090 were used to validate the methods in early and late 21st century conditions.

**Table 1.** List of GCM-RCM simulations selected for this study. The first column shows the abbreviation used in the text, the second and third columns show the GCM and RCM part of each model, respectively and the last column shows the name of the providing institution.

| Abbreviation | GCM Component | RCM Component | Institution |
|---|---|---|---|
| CNRM-A | CNRM-CM5 | ALADIN | CNRM |
| CCLM-MPI | MPI-ESM-LR | CCLM4.8.17 | CLM Community |
| RACMO-EC | EC-EARTH | RACMO22E | KNMI |
| RCA4-H | HadGEM | RCA4 | SMHI |
| WRF-I | IPSL-CMA5-MR | WRF | IPSL-INERIS |

### *2.3. Hydrological Simulations*

Hydrological impacts are simulated using sub-models extracted from the European scale hydrological model E-HYPE v2.5 [31]. E-HYPE is an application of the Hydrological Predictions for the Environment (HYPE) model developed in the Swedish Meteorological and Hydrological Institute [32]. The model is process-based, semi-distributed and designed for hydrological modeling at different spatial scales also in ungauged regions. The source code for HYPE (v4.10) is available at [33]. The model calibration and evaluation details together with a list of used topographical and land-use data sets can be found from [31]. E-HYPE was chosen due to its large spatial coverage, distributed nature and also to see how a hydrological model commonly used in impact assessments [34] responds to the bias adjustment step. Because the model is distributed (over sub-basins of median size 215 km$^2$), spatial variability in biases can also be assessed. Hydrological simulations were first conducted using the full 32-year periods. The first two years served to spin-up the hydrological model and have not been included in further analysis.

To sample the varying hydrological conditions in the European region, the model was run with bias adjusted daily mean temperature and precipitation within four sub-models selected from the catchments shown in Figure 1. These catchments have predominantly natural flow conditions which the hydrological model was capable of capturing well in the present-day climate. The northernmost model domain is located in the upper parts of Tornio river catchment, where water stored in snow pack and variations in it strongly regulate the annual cycle of surface hydrology, leading to peak river discharges in late spring and early summer. Two domains with maritime mild climatic conditions, which cover parts of Trent and Ems catchments, are less affected by snow processes and river discharges reach their maximum values in winter months. The southernmost study region is located in the Sava tributary of the Danube river and is characterized by a mixture of Alpine and Mediterranean climates. The seasonal cycle of river discharge has two distinctive peaks here, the first one caused by snow melt in Alpine regions in spring and the latter one by heavy rainfalls in autumn.

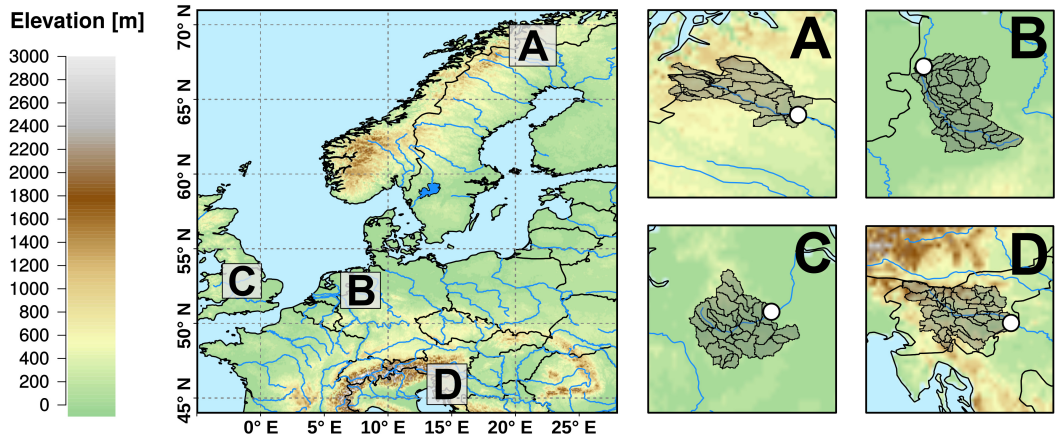


**Figure 1.** Geographical locations of the four sub-models selected for the hydrological simulation tests. The sub-models cover parts of (**A**) Tornio, (**B**) Ems, (**C**) Trent and (**D**) Sava river catchments. White dots denote the locations of the discharge gauging stations for which the reference period statistics are shown in Table 2.

**Table 2.** Statistics for daily time-series of simulated river discharge calculated against the observed discharge at the mouth of each hydrological sub-model (cf. Figure 1). The first column shows the catchment name, the second column Nash-Sutcliffe efficiency coefficient and the third column the relative volume error.

| Sub-Model | NSE | RE (%) |
|---|---|---|
| Tornio | 0.78 | −17.0 |
| Trent | 0.66 | −3.0 |
| Ems | 0.83 | 3.1 |
| Sava | 0.52 | 6.0 |

The ability of HYPE to simulate river flows in the reference period (1981–2010), when WFDEI is used directly as forcing, is briefly illustrated in Table 2, which shows the Nash-Sutcliffe efficiency coefficient (NSE) and relative volume error (RE) in simulated river discharge for the four gauging stations located at the outlets of the selected sub-models. The NSE values vary from 0.83 in Ems to 0.52 in the Sava region. These values are reasonable considering that E-HYPE has been calibrated uniformly for all of Europe to optimize predictions in ungauged regions. The RE values range from −17.0% to 6.0% with largest deviations seen in the Tornio sub-model, where the model tends to underestimate river discharge volume, particularly during the spring season. These differences are at least partially explained by the limitations of the WFDEI data set; the representation of daily precipitation variability is not sufficient in regions with large topographical variations, and subject to gauge undercatch for which the corrections are particularly uncertain in windy, snow dominated regions. Also temperature discontinuities might have a role in explaining the differences to the observations. In addition, the inherent limitations in the HYPE formulation and parameterisation likely explain part of this discrepancy (as would any other hydrological model).

### *2.4. Model Output Statistics*

Non-parametric quantile mapping applied both in the delta change (M1) and bias correction (M2) mode is used to benchmark the ability of joint bias correction methods to adjust temperature and precipitation for biases in the GCM-RCM simulations (see Table 3). In the following, the formulation is shown for the bias correction form of quantile mapping. For given simulated values of daily mean

temperature or precipitation in the scenario period $(s_i)$, the projected values $p_i$ are obtained by transforming $s_i$ according to

$$p_i = F_o^{-1}\left(F_c\left(s_i\right)\right). \quad (1)$$

here $F_c$ denotes the cumulative distribution function of the baseline period simulation and $F_o^{-1}$ its inverse estimated from the observations. Formulation for the delta change form is simply obtained by switching the indexes for the observations ($o$) and the future period simulation ($s$).

Before the transformation shown in Equation (1) is applied, the quantile-quantile relationship between $F_c$ and $F_o$ is smoothed by replacing individual quantiles with a running average taken over a specified quantile range using the approach and numerical values described in Räty [21] and Räty et al. [22]. If the future simulated values are outside the baseline period observations and model simulation, the quantile relationship is extrapolated assuming a constant, additive, relationship above the highest and below the lowest quantile for daily mean temperature. For precipitation, relative values are used. For further implementation details, the reader is referred to Räty [21] and Räty et al. [22].

**Table 3.** List of bias adjustment methods used in this study together with a short description.

| Name | Description | References |
|---|---|---|
| M1 | Univariate delta change: quantile mapping with smoothing | Räisänen and Räty [21], Räty et al. [22] |
| M2 | Univariate bias correction: quantile mapping with smoothing | Räisänen and Räty [21], Räty et al. [22] |
| M3 | Bi-variate bias correction: copula-based, precipitation conditioned on temperature | Li et al. [14], Gennaretti et al. [35] |
| M4 | Bi-variate bias correction: full 2-dimensional distribution using the N-pdft algorithm | Pitié et al. [36], Cannon [17] |

To take biases in the co-variations of daily mean temperature and precipitation into account, two bi-variate bias correction methods were implemented and compared with their univariate counterparts. In the first one (M3), the dependence structure is modeled separately from the marginal (i.e., unconditional) distributions of temperature and precipitation using a copula-based approach as described in Li et al. [14]. The implementation of this method is based on the properties of copula described by Sklar's theorem [37], which states that, given two random variables $X$ and $Y$ such as daily mean temperature and precipitation, their joint cumulative distribution ($H(x, y)$) can be constructed as

$$H(x, y) = C(F(x), G(y)), \quad (2)$$

where $C()$ denotes the cumulative copula distribution and $F(x)$ and $G(y)$ are the cumulative marginal distributions for $X$ and $Y$. Here, it is assumed that the dependence structure of daily mean temperature and precipitation can be reasonably modeled using a Gaussian copula as in Li et al. [14]. The Gaussian copula was chosen due to its relatively simple formulation and its ability to model both positive and negative correlations. Although several other parametric copulas are available for modeling the dependence structure, testing them is beyond the scope of this paper. Marginal distributions were modeled with parametric distributions in a similar manner as in Yang et al. [2] and Li et al. [14]. Temperature is assumed to follow Gaussian distribution, while gamma distribution is used to model precipitation above a wet-day threshold, here defined as 0.1 $\text{mmd}^{-1}$. To improve the performance of M3 at daily temporal scales, separate temperature distributions were fitted both on dry and wet days following a pre-adjustment of the fraction of wet days in model simulations to the observed one [2,35]. In M3, bias correction of the wet-day part of the joint distribution needs to be applied conditionally on either temperature or precipitation, thus offering two approaches to correct the joint distribution. Based on tests with these two alternatives, we decided to bias correct precipitation conditionally on temperature due to the slightly better overall performance of this option (not shown). The correlation parameter values obtained from fitting Gaussian copula to the baseline period simulations are illustrated in the supplementary material (Figure S1). For further details, the reader is referred to an R package available in GitHub [38], which contains implementations of methods M1–M3 written by the authors.

The second bi-variate bias correction method (M4) was recently proposed by Cannon [17] based on the N-pdft algorithm designed by Pitié et al. [36]. In this method the full 2-dimensional distribution structure is adjusted iteratively by reducing the adjustment of the 2-dimensional distribution to a series of 1-dimensional bias corrections of the marginal distributions. First, both temperature and precipitation distributions are normalized and randomly rotated to a new orthogonal coordinate system. Second, quantile mapping is applied to the rotated distributions. The adjusted distributions are then rotated back to the original coordinate system before repeating the described sequence. After several successive iterations it can be shown that the joint distribution converges to the target distribution [36]. As discussed by Cannon [17], the algorithm constructs the joint distribution at each iteration step as a linear combination of the bias corrected marginal distributions, which allows modifying the dependence structure. In the original article quantile delta mapping [4,8] was used to adjust the marginal distributions along the iteration cycles. Here, we use the same quantile mapping algorithm as in M2 instead of quantile delta mapping when bias correcting the marginal distributions of temperature and precipitation. Doing this, the bias corrected temperature and precipitation distributions are identical in M2 and M4. No smoothing was applied to the marginal distributions in the rotation step, as this would had contracted the underlying joint distribution of observations and the control period simulation. The algorithm was terminated after 50 iterations, which should be sufficient for the algorithm to converge to the target distribution, as illustrated by Cannon [17]. The implementation of M4 was based on the R package [39] available in the CRAN repository [40].

To take biases in the annual cycle into account, daily mean temperature and precipitation time series were adjusted on a monthly basis at each sub-model domain. As sampling errors are likely to affect the estimation of simulated changes (M1) and biases (M2–M4) in the GCM-RCM distributions, the effect of increased sample size on method performance was addressed by using both one- and two-month time windows when estimating model biases and simulated changes from GCM-RCM simulations. Using an even larger time window could in principle reduce the sampling noise [41], although with the expense of possibly introducing systematic biases to the future results. Tests with longer time windows did not show significant changes in the results, although a more systematic comparison would be needed to fully assess the potential benefits of reducing sampling noise.

To illustrate how each method represents the dependence structure in the calibration period, Figure 2 shows differences in the empirical copula density for the wet-day values of temperature and precipitation in comparison to the WFDEI copula in winter months of years 1981–2010 in the Tornio sub-model. The copula density has been estimated as a 2-dimensional histogram of the normalized ranks for both variables (see a more detailed description in Section 2.6). The density values can be interpreted as the ratio of the joint probability density to the case, where both variables were independent from each other. For example, values larger than one suggest a larger-than expected joint probability density in this part of the two-dimensional space. Differences in the empirical copula density roughly denote the difference in the strength of dependence for particular cumulative probability values of both temperature and precipitation. The panel for the reference data shows that temperature and precipitation are positively correlated in this example (the highest values slope from bottom-left to top-right). By design, M1 takes the inter-variable relationships directly from the reference. In contrast to the delta change mode, M2 inherits the multivariate dependence structure from the uncorrected GCM-RCM simulation (M0), with some modifications to it due to changes in the fraction of wet days [42]. Therefore, these methods can be thought to give the "limits" in which the multivariate methods can operate on adjusting the inter-variable correlations. Although the dependence structure of temperature and precipitation can be reasonably modeled using Gaussian copula on monthly scales, this might not be as feasible at daily scales, at least in cold climates. From Figure 2, it is immediately seen that although the Pearson correlation coefficient is well captured by M3, the overall copula structure shows noticeable deviations from the target distribution. In this particular case, M3 tends to overestimate the strength of the co-occurrence of low precipitation intensities and medium temperature values. Although the behavior of M3 strongly depends on the selected GCM-RCM, season and location,

this example highlights the importance of evaluating the full multivariate dependence structure to reveal such issues. In contrast to M3, M4 performs very well in capturing the WFDEI dependence structure and differences in the copula density field are small in most parts of the distribution.

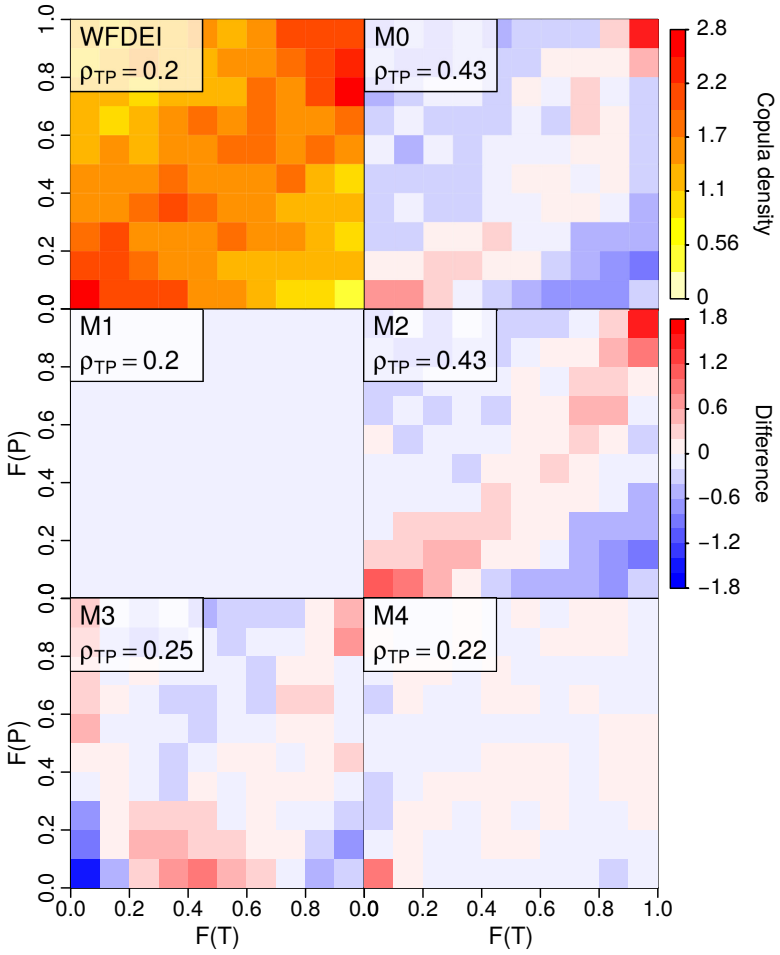


**Figure 2.** Empirical copula density of wet-day ($P > 0.1$ mmd$^{-1}$) precipitation and temperature in winter months (December-January-February) estimated from the reference data (WFDEI) in Tornio river catchment separately for each sub-basin and then averaged over the whole domain. In addition, differences in the estimated densities, when compared against WFDEI, are shown for CNRM-A GCM-RCM without bias adjustments (M0) and after applying each of the four methods (M1–M4) in the baseline period (1981–2010). The sub-model-averaged Pearson correlation coefficient is also shown on the top-left corner of each panel.

As another example, Figure 3 shows how the four methods capture the hydrological conditions in the Tornio sub-model, when adjusting the GCM-RCM simulations against WFDEI in the baseline period (1981–2010). For simplicity, a one-month time window has been used when estimating the simulated changes and biases in the GCM-RCM simulations. As expected, M1 has essentially a perfect correspondence with the observed hydrological conditions. The remaining biases in temperature and precipitation are very similar for M2 and M4 but not identical as small differences arise from the re-shuffling of daily values over the full 32-year period in M4, which is an inherent property of this and many other multivariate bias correction methods. Furthermore, M3 shows a very similar pattern for temperature, while the differences to M2 and M4 are more visible in the remaining precipitation biases. The differences between the methods are also visible in the annual monthly mean cycle of different aspects of the simulated surface hydrology. The differences to WFDEI are largest for M3 in

most cases, which is expected, as the (potentially sub-optimal) parametric marginal distributions used in M3 match the GCM-RCM-simulated temperature and particularly precipitation less accurately with WFDEI than the non-parametric versions used in M2 and M4. Apart from M1, M4 has generally the smallest differences to WFDEI, particularly in total runoff and snow water equivalent, although the remaining evapotranspiration biases are similar for M2 and M4.

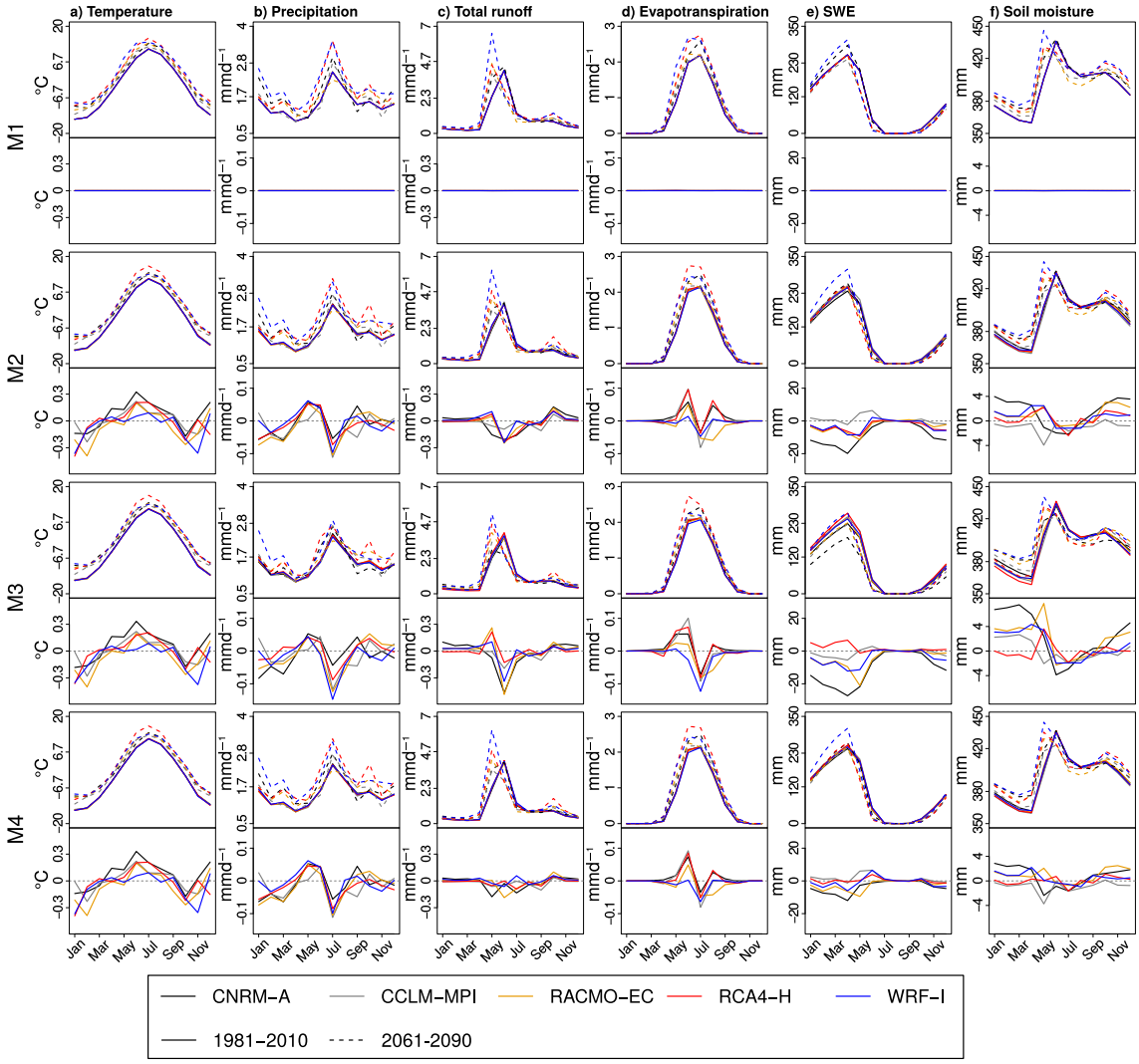


**Figure 3.** A real-world example showing the annual cycle of (**a**) daily mean temperature, (**b**) precipitation, (**c**) total runoff, (**d**) evapotranspiration, (**e**) snow water equivalent and (**f**) soil moisture in the Tornio river catchment in years (solid lines) 1981–2010 and (dashed lines) 2061–2090 separately for each of the GCM-RCMs when adjusted against WFDEI using the four (M1–M4) bias correction and delta change methods. To readily illustrate differences in comparison to WFDEI, the remaining biases in 1981–2010 are shown below the annual cycle panels separately for each method and variable.

### *2.5. Pseudo-Reality Framework*

To make inferences on the potential future performance of the selected univariate and multivariate methods, intermodel cross-validation is performed using the so-called pseudo-reality approach (Figure 4). In the first stage, one GCM-RCM at a time is used as the verifying model (i.e., pseudo-reality) against which the rest of the models are adjusted using the selected methods. The bias adjusted simulations are then compared against the pseudo-reality GCM-RCM with a set of performance measures. The resulting cross-validation statistics are then averaged over all pseudo-realities to

obtain an overall view of the bias correction performance in changing climatic conditions. The same framework is applied to the hydrological simulations to see to what extent the relative performance of the selected MOS methods differs when inspected from the hydrological modeling point of view. To this end, the bias adjusted temperature and precipitation time-series are used as input to the E-HYPE sub models, which are then run to simulate the future hydrological conditions in the selected catchments. Hydrological simulations are cross-validated in a similar manner as the GCM-RCM simulations using complementary performance measures.

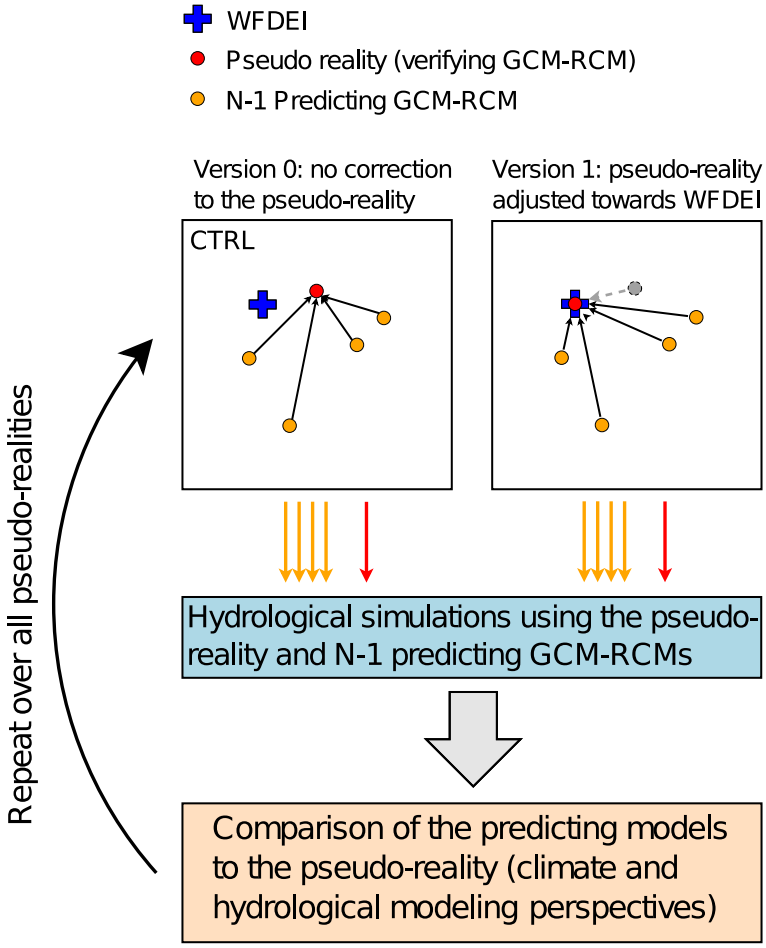


**Figure 4.** An illustration of the pseudo-reality framework procedures in the baseline period, applied both from climate modeling and hydrological modeling perspectives.

In order to improve the applicability of the pseudo-reality approach to hydrological simulations, two ways to construct pseudo-realities were tested (Figure 4): (a) raw GCM-RCM simulations were used as pseudo-realities without taking biases in relation to observations (i.e., WFDEI) into account [25]; (b) the annual cycle of the GCM-RCM acting as pseudo-reality was adjusted to biases in comparison to WFDEI by simply removing the mean bias at each day of the annual cycle using a 30-day sliding window. Daily adjustments were applied instead of monthly ones in order to avoid additional jumps in the annual cycle of the pseudo-reality time series. This shift in the mean values obviously alters the bias between pseudo-reality and the verifying models but leaves the changes in this relatively untouched (see Figure S2 in supplementary material). The motivation for the second approach is apparent: biases in relation to the observed climate are substantial in some of the selected GCM-RCMs, which leads to unrealistic hydrological model behavior both in the pseudo-reality runs and the verifying hydrological simulations. For example, substantial cold biases at high altitude regions in Sava sub-model and during winter in Tornio sub-model cause unrealistic volumes of snow to accumulate throughout the

simulation periods. We argue that without this additional bias adjustment step the use of GCM-RCMs as pseudo-realities when cross-validating bias adjustment methods from hydrological modeling perspective might not be reasonable due to unrealistic shifts in hydrological regimes. One should note that although the intention is to keep the daily variability in the pseudo-reality time series untouched, the multiplicative scaling applied to daily precipitation slightly modifies the spread of precipitation distributions both in the baseline and scenario periods. This also slightly changes the daily variability of hydrological simulations accordingly.

### *2.6. Metrics for GCM-RCM Simulations*

To assess the general similarity between the empirical cumulative probability distributions of the predicting models $F_{\text{pred}}$ and the GCM-RCM acting as pseudo-reality $F_{\text{ver}}$, 2-sample Cramér–von Misés (CM) statistic [43] was calculated according to

$$\text{CM} = A\Big\langle \frac{mn}{(m+n)^2} \Big\{ \sum_{i=1}^{m} \Big[ \hat{F}_{\text{pred}}(x_i) - F_{\text{ver}}(x_i) \Big]^2 + \sum_{j=1}^{n} \Big[ \hat{F}_{\text{pred}}(y_j) - F_{\text{ver}}(y_j) \Big]^2 \Big\} \Big\rangle, \tag{3}$$

where $\hat{(\,)}$ denotes the pooled sample of the four predicting GCM-RCM simulations, while $m$ and $n$ are the numbers of values within the pooled sample (x) and in pseudo-reality (y), respectively. The actual calculations were made for binned data using bin widths of 1 °C and 1 mmd$^{-1}$ and the same number of bins with identical bin boundaries for both predicting GCM-RCMs and the pseudo-reality GCM-RCM. $A\langle\rangle$ indicates an average over 12 months and the area of a sub-model. CM measures the similarity of two empirical distributions in probability space and puts more weight on discrepancies in the tails of the cumulative distributions than the widely used Kolmogorov–Smirnov statistic, which measures the maximum distance between the cumulative probability distributions. Comparison with these statistics did not reveal significant differences, and the results are shown only for CM.

The second statistic, mean absolute error (MAE), was calculated over quantiles $i$ ($i \in [1, ..., 100]$) of the predicting and verifying (i.e., pseudo-reality) model distributions following

$$\text{MAE} = A\langle |\hat{F}^{-1}_{\text{pred}}(i) - F^{-1}_{\text{ver}}(i)| \rangle, \tag{4}$$

where $A\langle\rangle$ encompasses averaging over the distribution quantiles in addition to temporal and spatial averaging. The analysis was also repeated using the mean squared error, but the results did not show substantial differences to MAE. Thus, the relative method performance is illustrated in terms of MAE in the remainder of the paper.

Two statistics measuring errors in inter-variable correlations were calculated. First, to assess to what extent the linear correlation is modified by different methods, MAE in the Pearson correlation coefficient was calculated between the average correlation coefficient of the four verifying models and pseudo-reality, averaged in a similar manner as in Equation (3). Secondly, to evaluate the remaining errors in the full dependence structure, the empirical copula density was approximated from the pseudo-observations $(u, v)$, estimated for the $i$th temperature ($x$) and precipitation ($y$) value as $u = \text{rank}(x_i)/(n+1)$ and $v = \text{rank}(y_i)/(n+1)$, where $n$ is the number of values for both variables. These values were binned 2-dimensionally and normalized such that the histogram approximately corresponds to the copula density. The 2-dimensional binning was done at 0.1 interval. MAE between empirical copula densities of the predicting GCM-RCMs and pseudo-reality was then calculated according to

$$\text{MAE}_{\text{c}} = \frac{1}{n} \sum_{i=1}^{n} |\hat{c}(i)_{\text{pred}} - c(i)_{\text{ver}}|. \tag{5}$$

In Equation (5), $\hat{c}_{\text{pred}}$ denotes the empirical copula density averaged over the four predicting models, $c_{\text{ver}}$ the copula density calculated for pseudo-reality and $n$ is the number of bins used to estimate the copula density. In the following, the subscript c is dropped for brevity. MAE based on kernel density estimates were also tested but the resulting statistics depended substantially on the used

kernel method and the kernel width and thus, were not considered further in this study. To reduce the effect of sampling noise to the results, temperature-precipitation pairs were pooled over the area of each sub-model and season before estimating the empirical copula densities. Identical values were handled using the same approach as in Gennaretti et al. [35]: ranks were first given randomly to identical values before estimating the empirical copula density. This was repeated 10 times and the final copula density was calculated as the average of the randomly ranked estimates. Despite being a simple and not a proper goodness-of-fit measure, this statistic readily illustrates how well each method is capable to adjust the full dependence structure. Gennaretti et al. [35] briefly pointed out that the measured performance depended on whether dry days were included when estimating the empirical copula density. While the focus is here on the copula density including the full time-series, the results for the wet-day copula can be found from the supplementary material (Figure S3).

### *2.7. Metrics for Hydrological Simulations*

An additional set of cross-validation statistics was calculated for the hydrological indexes. First, quantile distributions of river discharge $Q$ (i.e., flow-duration curves) were estimated at the outflow sub-basin of each of the sub-models. The average of the four predicting distributions was then compared against the pseudo-reality distribution using a logarithmic accuracy ratio (LAR10) defined as

$$\mathrm{LAR10} = A\left\langle \left| \log_{10}\left( \frac{\hat{F}^{-1}_{\mathrm{pred}}(i)}{F^{-1}_{\mathrm{ver}}(i)} \right) \right| \right\rangle, \tag{6}$$

where $A\langle\rangle$ has the same meaning as in Equation (4). The statistic is symmetric in the sense that the same value is assigned for under- and overestimation of the same relative magnitude [17]. This alleviates the issue of most other relative accuracy measures penalizing overestimation more strongly than underestimation. In addition to distribution-averaged statistics, LAR10 was also inspected individually for the 5th ($Q_5$) and 99th ($Q_{99}$) percentile of the flow duration curve to see how the relative performance of the selected MOS methods varies in the tails of the monthly flow duration curves.

The analysis of river discharges is complemented by evaluating a set of individual flux and storage elements, which affect the overall water balance and river discharge generation. To this end, MAE was calculated for the monthly mean values of total runoff (R), evapotranspiration (E), soil moisture (S) and snow water equivalent (SWE) in a similar manner as for daily mean temperature and precipitation. MAE in the mean annual maximum SWE ($SWE_{max}$) was also calculated to evaluate how method differences are reflected in the simulation of highest snow pack depths. This allows us to make inferences on how the remaining errors in daily mean temperature and precipitation and in their inter-variable correlations affect the different hydrological elements, as well as to gain insights on the relative performance of the selected bias adjustment methods in terms of hydrological modeling results.

## 3. Results

### *3.1. Distribution-Averaged Statistics for Daily Mean Temperature and Precipitation*

We first inspect the overall performance of the four methods from the GCM-RCM perspective. Figure 5 illustrates the distribution-averaged cross-validation statistics in the two scenario periods (see Figure S4 for the statistics in years 1981–2010). When first concentrating on the results shown for the years 2061–2090, it is seen that bias correction methods M2–M4 slightly outperform M1 in adjusting the temperature distribution in terms of both CM and MAE. On the other hand, both CM and MAE of M3 are very close to M2 and M4, which indicates that temperature can be reasonably modeled using a normal distribution. Although the general picture is mostly similar for precipitation, CM and MAE give a partially contrasting picture about the relative method performance. The CM values are smallest and almost identical for methods M2–M4, while the relative performance of M3

is slightly worse than M2 and M4 in terms of MAE. This might be partially explained by how the fraction of dry days is explicitly taken into account by M3 in its corrections, while using a gamma distribution to model the precipitation distribution might not capture biases in it as efficiently as non-parametric quantile mapping. Using a two-month time window generally reduces errors in both temperature and precipitation distributions, which is in line with the results of Räisänen and Räty [21] and Räty et al. [22]. As expected, all methods have substantially smaller errors in the marginal distributions of temperature and precipitation in comparison to the uncorrected model simulations (M0) in both periods.

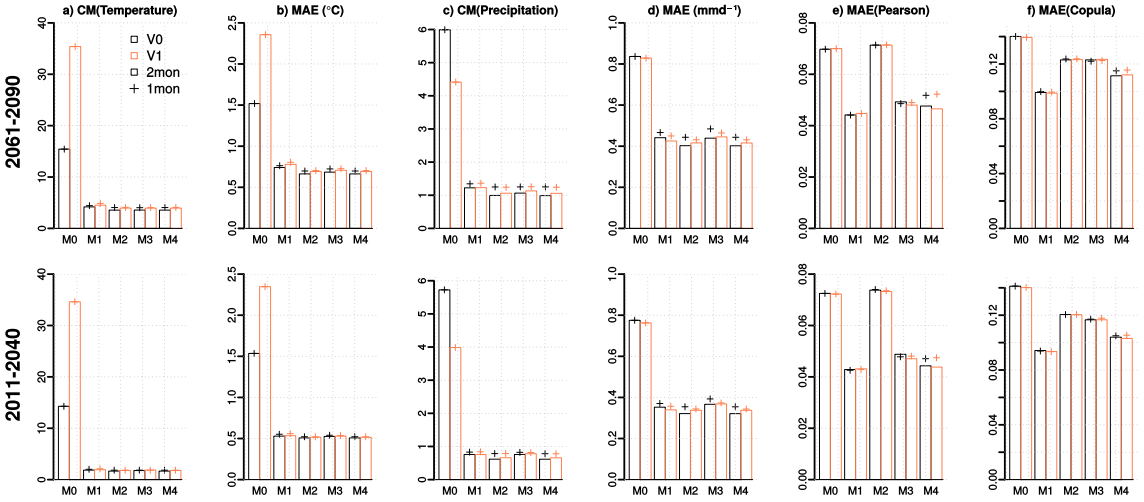


**Figure 5.** Cross-validated CM and MAE for (**a**,**b**) daily mean temperature and (**c**,**d**) daily precipitation distribution in years 2011–2040 (**bottom**) and 2061–2090 (**top**). Also shown are the MAE in (**e**) the Pearson correlation coefficient and (**f**) the empirical copula density. Black color denotes the cross-validation statistics for the pseudo-reality approach without additional adjustments (V0), while the results for the approach where pseudo-realities have been adjusted to biases in relation to WFDEI are shown in red (V1). Furthermore, crosses (bars) indicate the results for the one-month (two-month) time window used to estimate simulated changes or model biases, shown for both V0 and V1. Note that the differences between the one- and two-month time windows are typically small, as indicated by the small differences between the bars and crosses.

The MAE for the Pearson correlation coefficient and the empirical copula density, when calculated over the full monthly time-series of temperature and precipitation, is also shown in Figure 5. The results for the Pearson correlation coefficient show that, although M3 and M4 improve the results in comparison to method M2, M1 performs slightly better in capturing the linear correlation between temperature and precipitation than the other methods. Moreover, M4 seems to be susceptible to the effect of noise, as M3 has a somewhat smaller MAE when the one-month time window is used. The situation is slightly different when the MAE in the empirical copula density is considered. While M1 has again the best performance out of all methods, M2 has now MAE values which are closer to the bi-variate methods. The modest improvement obtained with M4 in comparison to M1 is again at least partially related to the small sample size, as indicated by the reduction in the MAE values for the two-month time window. Yet, this highlights the difficulty to robustly estimate biases in inter-variable correlations in a changing climate. As M1 has a superior performance in terms of both of the two measures regardless of the period considered, this suggests that the inter-variable correlations do not change substantially among the selected models and within the studied regions.

The bottom row shows the cross-validation statistics for the near-term scenario period (2011–2040). As expected, the remaining errors are generally smaller for all methods in this period. The marginal distributions of both temperature and precipitation are slightly better captured by method M1 in comparison to other methods, while the relative performance of other methods does not show marked

differences between the two periods. Furthermore, the MAE in the Pearson correlation coefficient and the copula density indicate a slightly improved performance for M3 and M4 in comparison to M1, although M1 still has the smallest MAE in all cases.

In qualitative terms, the cross-validation statistics are similar for temperature and precipitation regardless of the pseudo-reality approach. By far, the largest differences are shown by method M0 for which the cross-validation statistics calculated for temperature deteriorate when correcting the pseudo-reality GCM-RCM toward WFDEI (V1), while the opposite happens for the precipitation statistics. For temperature, the larger MAE in V1 is explained by the systematic cold bias within the GCM-RCM ensemble. However, for methods M1-M4 the results are mostly similar between the two pseudo-reality approaches, although the cross-validation statistics for the temperature and precipitation distributions tend to be slightly worse for the two-month time window after pseudo-realities have been adjusted against WFDEI (V1). This suggests that, from the climate modeling perspective, the additional adjustment step does not substantially modify the cross-validation statistics apart from the uncorrected model simulations, backing up its use in the hydrological modeling step.

While not the specific target of this study, it should be mentioned that an inherent property of M4 is that in order to obtain correct ranks for each temperature and precipitation pair, both time series need to be temporally re-ordered. This is to a lesser extent an issue in M3, in which only the temporal sequence of precipitation is potentially modified. As the temporal re-ordering might affect the hydrological simulations, a modified version of M4 was tested. First, the time series of uncorrected temperature and precipitation were divided into dry and wet days in a similar manner as in M3. Next, M2 was applied separately on wet-day and dry-day distributions to retain the improved statistics for them, as obtained with M4. Finally, the N-pdft algorithm was applied only on wet-day distributions of temperature and precipitation. Tests with the modified algorithm showed, however, that although the cross-validated MAE of both correlation measures decreased slightly, changes in the cross-validation statistics for hydrological variables in comparison to the original method varied non-systematically depending on the season, region and variable considered (not shown) and, thus, did not offer systematic improvements in comparison to the original algorithm.

### *3.2. Cross-Validation Statistics for Hydrological Simulations*

#### 3.2.1. Spatially Distributed Variables

How are the differences in the ability of the four methods to adjust the joint distribution of temperature and precipitation reflected in the hydrological simulations? Figure 6 shows the average cross-validation statistics for the five hydrological components (R, E, S, SWE and $SWE_{max}$) in the two scenario periods (see Figure S5 for the statistics in years 1981–2010). For clarity, the results are shown only for the two-month time window in the remainder of the paper. When looking at the MAE in monthly mean runoff (R) in the late 21st century period, it is seen that M4 outperforms the other methods in pseudo-reality approach V0, although the differences are small in comparison to M1. In pseudo-reality approach V1, however, MAE is practically identical for M1–M2 and M4. M3 has a somewhat worse performance than the other methods, which is likely caused by the larger remaining errors in the precipitation distribution than for the other methods. When evapotranspiration (E) as simulated by HYPE is considered, bi-variate methods M3 and M4 have the smallest MAE in monthly mean values and M3 actually has the smallest MAE in pseudo-reality approach V0. In addition, the results for E illustrate a side-effect of the additional pseudo-reality adjustment (V1): the MAE in E is systematically larger in V1 for all methods, as the systematic underestimation of temperature in V0 likely leads to too weak evapotranspiration in comparison to real-world hydrological simulations.

We next take a look at the cross-validation statistics for the two storage variables. The MAE of soil moisture (S) is almost identical for all methods, which indicates that it is relatively insensitive to the adjustment of daily-scale inter-variable correlations. The small differences between the four methods tend to follow those seen in the MAE calculated over the precipitation distribution, as the

MAE is smallest and almost identical for methods M4 and for M2. The last two panels in Figure 6 show the MAE of the monthly mean SWE and $SWE_{max}$. These results illustrate the main benefit of pseudo-reality approach V1. As predicted, the adjustment of pseudo-reality GCM-RCMs reduces biases in snow variables, although with the expense of increased MAE for E, as discussed before. This also causes differences in the relative ranking of the correction methods between the two pseudo-reality approaches (V0 and V1); M4 performs slightly worse in relation to M1 in reality approach V0, whereas the opposite is seen after adjusting the pseudo-reality GCM-RCMs towards WFDEI (V1). Overall, these results indicate that the simulation of most hydrological aspects is only marginally improved by joint bias correction and that the accurate adjustment of marginal distributions plays a more important role, at least when only temperature and precipitation are used as input in a hydrological model, such as HYPE.

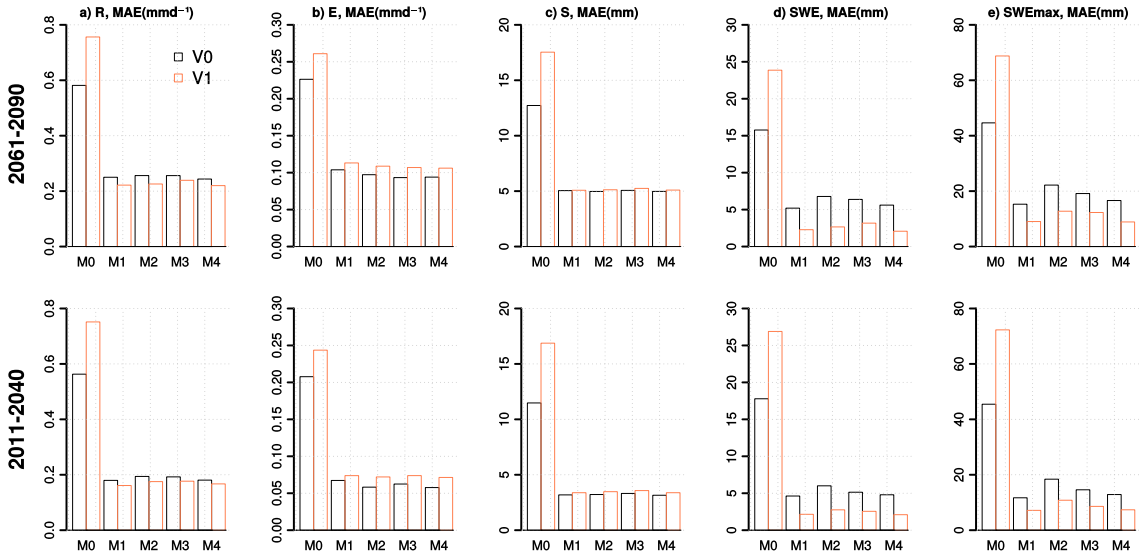


**Figure 6.** Similar to Figure 5 but for the cross-validated MAE of monthly mean (**a**) total runoff, (**b**) evapotranspiration, (**c**) soil moisture, (**d**) snow water equivalent and (**e**) the mean annual maximum of snow water equivalent in years (**bottom**) 2011–2040 and (**top**) 2061–2090.

The cross-validation statistics for the near-term scenario period are in line with the corresponding statistics of temperature and precipitation, with generally smaller errors in all studied hydrological aspects than in the later scenario period. The relatively better performance of M1, when adjusting the joint distribution of temperature and precipitation at that time is to some extent reflected in the hydrological simulations (bottom of row Figure 6), as R, E and S are all better captured by M1 in the near-future period. In contrast to the later scenario period, M2 and M4 have smaller MAE values in monthly mean evapotranspiration than M3. The cross-validation statistics of monthly mean SWE and $SWE_{max}$ show the largest differences between bias adjustment methods also in this period, indicating that method choice is most important for this variable from the studied hydrological aspects.

### 3.2.2. Evaluation of Future River Discharges

The analysis is complemented by illustrating the cross-validated LAR10 for $Q_5$, $Q_{99}$ as well as for the distribution-averaged LAR10 in the two scenario periods (Figure 7). The absolute values of LAR10 vary to some extent between the two pseudo-reality approaches. For example, LAR10 of $Q_5$ is systematically smaller in V1 for all methods (apart from M0) in both periods, while the opposite is seen in the $Q_{99}$ in the early 21st century period. Furthermore, the performance of all methods is extremely consistent when the distribution-averaged LAR10 is considered. Methods M2 and M4 have a marginally smaller LAR10 than M1 and M3, while in the earlier scenario period method M1 performs equally well or even better than M2 and M4. Also the best performing method

depends on the pseudo-reality approach when low flows ($Q_5$) are considered. In V0, method M2 somewhat outperforms the other methods in both periods, while in V1 method M3 has a slightly better performance in comparison to the other methods. On the other hand, M1 has the largest LAR10 values in both periods, which is probably related to the larger errors in temperature and evapotranspiration accordingly. The simulation of $Q_{99}$ seems to marginally benefit from the adjustment of inter-variable correlations, as M4 has the smallest LAR10 among the four methods, particularly in years 2011–2040. Again, the LAR10 is larger for M3 than for the other methods, most likely due to the combination of the aforementioned issues.

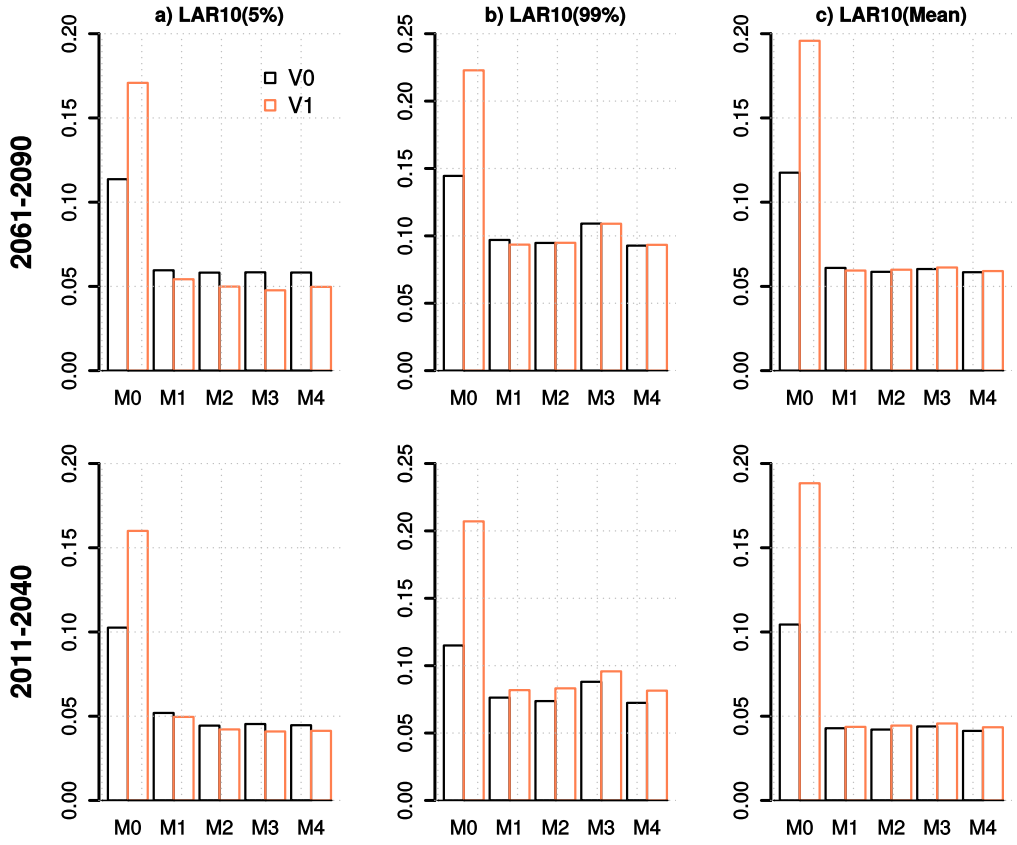


**Figure 7.** Similar to Figure 5 but for the cross-validated LAR10 in (**a**) the 5th and (**b**) 99th percentile of flow duration curves shown together with (**c**) the distribution-averaged LAR10 in years 2011–2040 (**bottom**) and 2061–2090 (**top**).

### *3.3. Temporal and Spatial Variations in the Cross-Validation Statistics*

Seasonal and spatial variations in both the relative performance of the studied methods and contributions of different hydrological processes are reflected in the cross-validation statistics for the hydrological simulations. To better infer the reasons for these variations, Figures 8 and 9 show matrices of cross-validated MAE for different aspects of the joint distribution of temperature and precipitation and the distributed hydrological variables in winter and summer seasons in the four sub-models (see Figures S6 and S7 for the same statistics in spring and autumn seasons). In absolute terms the largest MAE values for temperature are seen in Tornio in both seasons, while for precipitation the MAE is generally largest in Sava. The pattern is less clear for the correlation measures, which show larger variations between the four regions. The numeric values in the panels denote the percentage difference to M1. These values indicate that methods M2–M4 perform better than M1 in most cases in winter, while in summer the temperature is relatively well captured by M1. On the other hand, the statistics for both correlation measures indicate that M2 has systematically poorer performance than the other methods and the relative difference is particularly large in Tornio and Ems. The most striking feature

in Figure 8 is the consistently better performance of M1 in capturing the empirical copula density when compared to the other methods, regardless of region or season; apart from M4, which has a relatively similar performance in winter, M1 outperforms the other methods by a large margin.

When the cross-validation statistics for temperature and precipitation are compared against the statistics for the hydrological variables, it is evident that the relative differences between the methods are mostly explained by their capability to adjust the marginal distributions of temperature and precipitation, particularly in those regions and seasons, where snow processes play a less important role in the hydrological cycle. Backing up the previous conclusions, the added value of the adjustment of temperature and precipitation dependence structure, as indicated by differences between methods M2 and M4, is most visible in Tornio and Sava sub-models, where M4 systematically improves the simulation of SWE. The link between the improved simulation of SWE and improvements in total runoff and soil moisture is apparent, as M4 has a smaller MAE than M2 in both variables. Tornio and Sava also show the largest differences between pseudo-reality approaches V0 and V1, as SWE has substantial errors remaining even in summer in Tornio, when V0 is used.

Results from the previous section suggest that quantile mapping applied as a delta change method (M1) has a relatively robust performance from a hydrological modeling perspective. On sub-model scale this is only partially true, as M2 and M4 tend to have better performance in the northern sub-models, while M1 performs particularly well in Sava catchment. However, the relative differences are small in many cases also in other regions, which suggests that the delta change approach might be a good alternative for bias correction. From individual methods, M3 has largest variations in its relative performance between the four sub-models and two seasons. These variations seem not to be solely due to issues with the marginal distributions, and in Tornio, for example, failures to capture the full dependence structure in winter (cf. Figure 2) might deteriorate the statistics for M3.

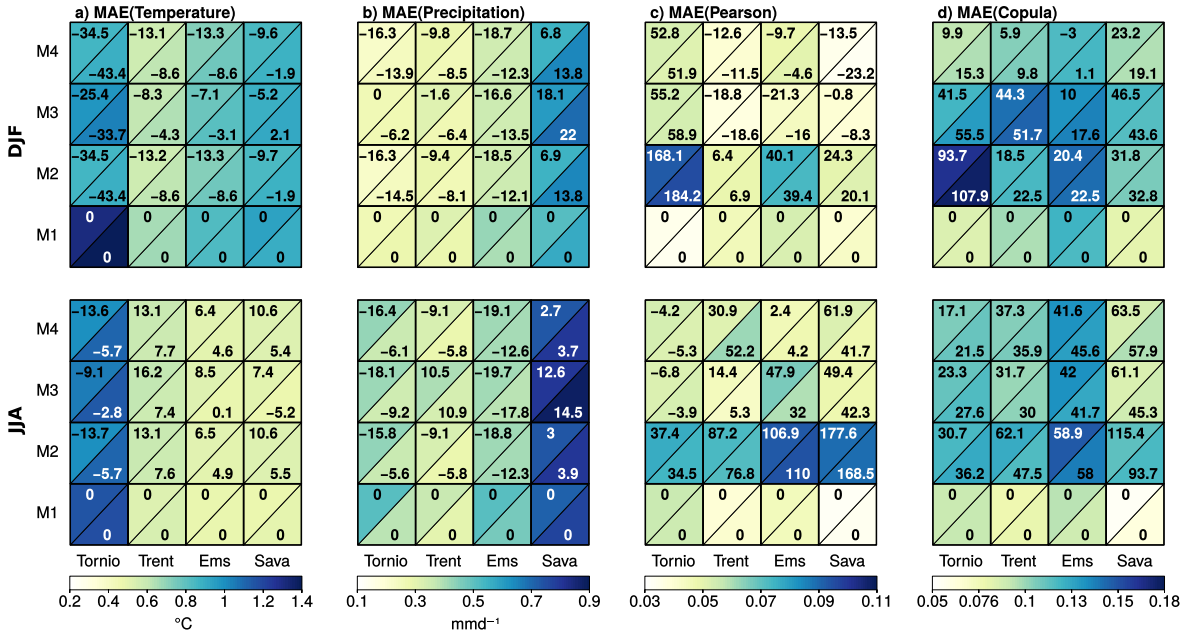


**Figure 8.** Panels showing the cross-validated MAE (colors) of (**a**) daily mean temperature, (**b**) daily precipitation, (**c**) Pearson correlation coefficient and (**d**) empirical copula density separately for each method (panel rows) at each hydrological sub-domain (panel columns) in years 2061–2090, when two-month time window has been used to estimate simulated changes or model biases. Values for the pseudo-reality approach V0 (V1) are plotted in the upper (lower) triangle of each cell and are shown separately for (**top**) winter and (**bottom**) summer months. In addition, percentage differences to M1 are shown as numeric values for each element.

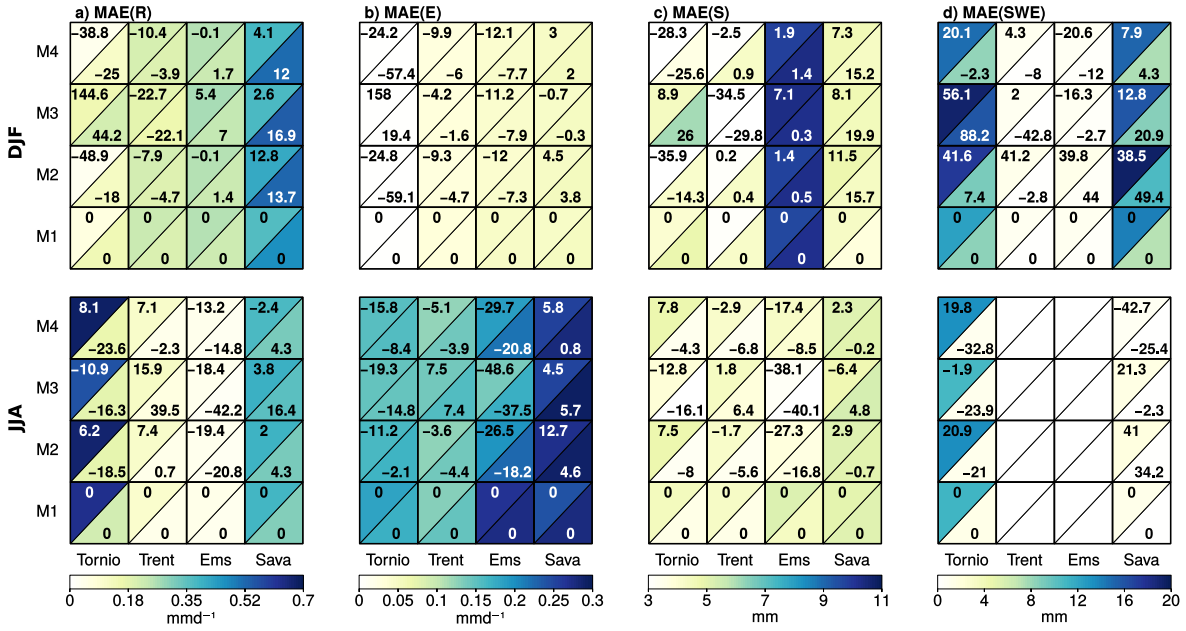


**Figure 9.** As in Figure 8 but shown for monthly mean (**a**) total runoff, (**b**) evapotranspiration, (**c**) soil moisture and (**d**) snow water equivalent.

## 4. Discussion and Conclusions

This paper presents the results from a study in which two joint bias corrections methods applied in the bi-variate mode are compared against quantile mapping applied both traditionally and in the delta change mode using five EURO-CORDEX GCM-RCM simulations as a proxy for the future climate. The evaluation is two-fold: first, cross-validation is performed to obtain quantitative estimates for the relative method performance in the early and late 21st century conditions, when applied to construct future projections of the joint distribution of daily mean temperature and precipitation; second, these projections were fed to a hydrological model to assess, whether or not the bi-variate adjustments improve future hydrological simulations in comparison to univariate quantile mapping.

The main results of these exercises are summarized as follows:

- By design, joint bias correction brings the inter-variable correlations closer to the observed one in the baseline period. In particular, the iterative N-pdft algorithm (M4) reproduces the full dependence structure (as measured by the MAE in the empirical copula density) well in comparison to univariate quantile mapping (M2). The adjustment of inter-variable correlation in M3 might fail in certain situations, as the method tends to have larger remaining biases in the copula structure in winter conditions in HYPE Tornio sub-model.
- Cross-validation statistics in years 2011–2040 and 2061–2090 indicate that although the correlation structure is improved in terms of Pearson correlation, the benefit of bi-variate methods is less clear when the full dependence structure is considered. Part of the modest improvement is likely explained by the limited sample size, which might lead to over-fitting to the present-day climatic conditions. On the other hand, quantile mapping applied in the delta change mode (M1) often performs better than the other methods, which indicates that retaining present-day correlation structures of temperature and precipitation might be sufficient also in future projections.
- The results suggest that the pseudo-reality approach is potentially useful for evaluating the relative performance of bias adjustment methods from hydrological modeling perspective in the future climate. However, to improve the validity of the conclusions in these types of studies, the implementation of pseudo-reality framework needs to be designed on case-by-case basis, for example by first bias adjusting the pseudo-reality GCM-RCMs to avoid unrealistic shifts in hydrological regimes.

- For the hydrological variables, the bi-variate approaches offered no substantial advantage over the univariate methods with M4 often having similar performance to M2. Only marginal improvements in comparison to methods M1 and M2 are seen in the cross-validation statistics for high flows and for the monthly mean and annual maximum snow water equivalent in Tornio and Sava. Although quantile mapping applied as a delta change method (M1) has slightly poorer performance in projecting marginal distributions of temperature and precipitation than quantile mapping-based bias correction (M2) and its bi-variate version (M4), the cross-validation statistics indicate that it has a relatively good ability to capture the future hydrological conditions. Nevertheless, for the hydrological variables studied (apart from snow), there were only small differences in cross-validation statistics between the tested methods, indicating that care should be taken when selecting MOS methods for particular purposes and (ideally) several methods should be used in parallel. Overall, the results highlight the difficulty to illustrate the added value of more complex methods, when applying them in producing projections for daily mean temperature and precipitation.

The main shortcoming of this study is the limited number of GCM-RCM simulations available for cross-validation tests, and optimally a larger set of model simulations should be used. Furthermore, the response to different bias correction and delta change algorithms is likely dependent on the hydrological model and the used parameterisations. For example, earlier studies have shown e.g., [44] that projections for evapotranspiration based on parameterising potential evapotranspiration using only temperature are not suitable for all climatic conditions. Furthermore, snow processes were parameterized in the HYPE simulations using a simple degree-day algorithm, which does not take solar radiation and other meteorological factors into account. More complex parameterizations, which require multiple variables as input, should be evaluated in further studies. If bias correction of a higher dimensional joint distribution were required, more sophisticated bias correction methods could, at least in principle, provide larger improvements in comparison to univariate methods, depending on the available data for robust calibration. This has been demonstrated by Cannon [17], who showed that M4 performs very well when adjusting a higher dimensional distribution for Canadian Forest Fire Weather Index calculations in the present-day climate. Moreover, different implementations of M3 could also be studied in future research. Most importantly, Gaussian copula is unlikely to be the optimal choice for describing the temperature and precipitation dependence structure in some cases and the use of other copulas should be further explored.

The presented framework allows to make some inferences about the ability of bias correction and delta change methods in constructing projections for future climate and their applicability from hydrological modeling perspective, an information of major interest to the impact modeling community. However, these tests are not sufficient alone to determine whether a particular method is suitable for climate change assessments. Additional tests such as those implemented in the VALUE framework [24] should be conducted to obtain a complete picture of benefits and limitations of bias correcting inter-variable correlations. As discussed above, the used approach does not easily allow to evaluate the potential benefits/adverse effects of the modification of temporal sequencing to the hydrological model results, which would require temporally synchronized model simulations and reference data. Furthermore, the effect of errors in the spatial representation of climate model output caused (e.g.,) by differing topography should be studied comprehensively to see how sensitive future hydrological simulations are to the correct representation of spatial fields.

**Supplementary Materials:** Figures S1–S7 are available online at http://www.mdpi.com/2225-1154/6/2/33/s1.

**Author Contributions:** C.D. selected and provided the hydrological sub-models and O.R. ran the hydrological simulations. O.R. and J.R. implemented the bias adjustment methods. All authors contributed to the planning of the study, the analysis of the results and writing the manuscript.

**Acknowledgments:** The study is supported by the Academy funded Center of Excellence (project No. 307331). Olle Räty is funded by the Vilho, Lauri and Yrjö Väisälä Foundation and by NordForsk through project number 74456 "Statistical Analysis of Climate Projections" (eSACP). Further funding was received from the projects HazardSupport and AQUACLEW. HazardSupport is financed by the Swedish Civil Contingencies Agency, MSB (grant No. 2015-3631). AQUACLEW is part of ERA4CS, an ERA-NET initiated by JPI Climate and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES) and ANR (FR) with co-funding by the European Union (Grant 690462). We would also like to thank the climate modeling groups participating the European branch of Coordinated Regional Climate Downscaling Experiment (EURO-CORDEX) for producing and making their model output publicly available.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Déqué, M. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Glob. Planet. Chang.* **2007**, *57*, 16–26. [CrossRef]
2. Yang, W.; Andréasson, J.; Phil Graham, L.; Olsson, J.; Rosberg, J.; Wetterhall, F. Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrol. Res.* **2010**, *41*, 211–229. [CrossRef]
3. Olsson, J.; Berggren, K.; Olofsson, M.; Viklander, M. Applying climate model precipitation scenarios for urban hydrological assessment: A case study in Kalmar City, Sweden. *Atmos. Res.* **2009**, *92*, 364–375. [CrossRef]
4. Li, H.; Sheffield, J.; Wood, E.F. Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. *J. Geophys. Res. Atmos.* **2010**, *115*. [CrossRef]
5. Maraun, D.; Wetterhall, F.; Ireson, A.M.; Chandler, R.E.; Kendon, E.J.; Widmann, M.; Brienen, S.; Rust, H.W.; Sauter, T.; Themeßl, M.; et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **2010**, *48*, RG3003. [CrossRef]
6. Themeßl, M.J.; Gobiet, A.; Leuprecht, A. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.* **2011**, *31*, 1530–1544. [CrossRef]
7. Hempel, S.; Frieler, K.; Warszawski, L.; Schewe, J.; Piontek, F. A trend-preserving bias correction –the ISI-MIP approach. *Earth Syst. Dyn.* **2013**, *4*, 219–236. [CrossRef]
8. Cannon, A.J.; Sobie, S.R.; Murdock, T.Q. Bias Correction of GCM Precipitation by Quantile Mapping: How Well Do Methods Preserve Changes in Quantiles and Extremes? *J. Clim.* **2015**, *28*, 6938–6959. [CrossRef]
9. Teutschbein, C.; Seibert, J. Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *J. Hydrol.* **2012**, *456–457*, 12–29. [CrossRef]
10. Chen, J.; Brissette, F.P.; Chaumont, D.; Braun, M. Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resour. Res.* **2013**, *49*, 4187–4205. [CrossRef]
11. Trenberth, K.E.; Shea, D.J. Relationships between precipitation and surface temperature. *Geophys. Res. Lett.* **2005**, *32*, L14703. [CrossRef]
12. Berg, P.; Haerter, J.O.; Thejll, P.; Piani, C.; Hagemann, S.; Christensen, J.H. Seasonal characteristics of the relationship between daily precipitation intensity and surface temperature. *J. Geophys. Res. Atmos.* **2009**, *114*. [CrossRef]
13. Piani, C.; Haerter, J.O. Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophys. Res. Lett.* **2012**, *39*. [CrossRef]
14. Li, C.; Sinha, E.; Horton, D.E.; Diffenbaugh, N.S.; Michalak, A.M. Joint bias correction of temperature and precipitation in climate model simulations. *J. Geophys. Res. Atmos.* **2014**, *119*, 153–162. [CrossRef]
15. Mehrotra, R.; Sharma, A. A Multivariate Quantile-Matching Bias Correction Approach with Auto- and Cross-Dependence across Multiple Time Scales: Implications for Downscaling. *J. Clim.* **2016**, *29*, 3519–3539. [CrossRef]
16. Cannon, A.J. Multivariate Bias Correction of Climate Model Output: Matching Marginal Distributions and Intervariable Dependence Structure. *J. Clim.* **2016**, *29*, 7045–7064. [CrossRef]

17. Cannon, A.J. Multivariate quantile mapping bias correction: An N-dimensional probability density function transform for climate model simulations of multiple variables. *Clim. Dyn.* **2017**, *50*, 31–49. [CrossRef]
18. Ehret, U.; Zehe, E.; Wulfmeyer, V.; Warrach-Sagi, K.; Liebert, J. HESS Opinions "Should we apply bias correction to global and regional climate model data?". *Hydrol. Earth Syst. Sci.* **2012**, *16*, 3391–3404. [CrossRef]
19. Maraun, D. Bias Correcting Climate Change Simulations—A Critical Review. *Curr. Clim. Chang. Rep.* **2016**, *2*, 211–220. [CrossRef]
20. Maraun, D. Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums. *Geophys. Res. Lett.* **2012**, *39*, L06706. [CrossRef]
21. Räisänen, J.; Räty, O. Projections of daily mean temperature variability in the future: cross-validation tests with ENSEMBLES regional climate simulations. *Clim. Dyn.* **2013**, *41*, 1553–1568. [CrossRef]
22. Räty, O.; Räisänen, J.; Ylhäisi, J.S. Evaluation of delta change and bias correction methods for future daily precipitation: intermodel cross-validation using ENSEMBLES simulations. *Clim. Dyn.* **2014**, *42*, 2287–2303. [CrossRef]
23. Van Schaeybroeck, B.; Vannitsem, S. Assessment of calibration assumptions under strong climate changes. *Geophys. Res. Lett.* **2016**, *43*, 1314–1322. [CrossRef]
24. Maraun, D.; Widmann, M.; Gutiérrez, J.M.; Kotlarski, S.; Chandler, R.E.; Hertig, E.; Wibig, J.; Huth, R.; Wilcke, R.A. VALUE: A framework to validate downscaling approaches for climate change studies. *Earth's Future* **2015**, *3*, 1–14. [CrossRef]
25. Velázquez, J.A.; Troin, M.; Caya, D.; Brissette, F. Evaluating the Time-Invariance Hypothesis of Climate Model Bias Correction: Implications for Hydrological Impact Studies. *J. Hydrometeorol.* **2015**, *16*, 2013–2026. [CrossRef]
26. Jacob, D.; Petersen, J.; Eggert, B.; Alias, A.; Christensen, O.B.; Bouwer, L.M.; Braun, A.; Colette, A.; Déqué, M.; Georgievski, G.; et al. EURO-CORDEX: New high-resolution climate change projections for European impact research. *Reg. Environ. Chang.* **2014**, *14*, 563–578. [CrossRef]
27. Weedon, G.P.; Balsamo, G.; Bellouin, N.; Gomes, S.; Best, M.J.; Viterbo, P. The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.* **2014**, *50*, 7505–7514. [CrossRef]
28. Rust, H.W.; Kruschke, T.; Dobler, A.; Fischer, M.; Ulbrich, U. Discontinuous Daily Temperatures in the WATCH Forcing Datasets. *J. Hydrometeorol.* **2015**, *16*, 465–472. [CrossRef]
29. Federated ESGF-CoG Nodes. Available online: https://esgf.llnl.gov/nodes.html (accessed on 27 April 2016).
30. Moss, R.H.; Edmonds, J.A.; Hibbard, K.A.; Manning, M.R.; Rose, S.K.; van Vuuren, D.P.; Carter, T.R.; Emori, S.; Kainuma, M.; Kram, T.; et al. The next generation of scenarios for climate change research and assessment. *Nature* **2010**, *463*, 747–756. [CrossRef] [PubMed]
31. Donnelly, C.; Andersson, J.C.; Arheimer, B. Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrolog. Sci. J.* **2016**, *61*, 255–273. [CrossRef]
32. Lindström, G.; Pers, C.; Rosberg, J.; Strömqvist, J.; Arheimer, B. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrol. Res.* **2010**, *41*, 295–319. [CrossRef]
33. HYPE Open Source Code. Available online: http://hypecode.smhi.se/ (accessed on 13 December 2016).
34. Donnelly, C.; Greuell, W.; Andersson, J.; Gerten, D.; Pisacane, G.; Roudier, P.; Ludwig, F. Impacts of climate change on European hydrology at 1.5, 2 and 3 degrees mean global warming above preindustrial level. *Clim. Chang.* **2017**, *143*, 13–26. [CrossRef]
35. Gennaretti, F.; Sangelantoni, L.; Grenier, P. Toward daily climate scenarios for Canadian Arctic coastal zones with more realistic temperature-precipitation interdependence. *J. Geophys. Res. Atmos.* **2015**, *120*, 862–877. [CrossRef]
36. Pitié, F.; Kokaram, A.C.; Dahyot, R. Automated colour grading using colour distribution transfer. *Comput. Vis. Image Underst.* **2007**, *107*, 123–137. [CrossRef]
37. Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **1959**, *8*, 229–231.
38. BCUH: University of Helsinki bias adjustment tools. Available online: https://github.com/RatyO/BCUH (accessed on 9 April 2018).
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.

40. Cannon, A. MBC: Multivariate Bias Correction of Climate Model Outputs. Available online: https://CRAN.R-project.org/package=MBC (accessed on 20 April 2017).
41. Reiter, P.; Gutjahr, O.; Schefczyk, L.; Heinemann, G.; Casper, M. Does applying quantile mapping to subsamples improve the bias correction of daily precipitation? *Int. J. Climatol.* **2017**, *38*, 1623–1633. [CrossRef]
42. Rajczak, J.; Kotlarski, S.; Schär, C. Does Quantile Mapping of Simulated Precipitation Correct for Biases in Transition Probabilities and Spell Lengths? *J. Clim.* **2016**, *29*, 1605–1615. [CrossRef]
43. Anderson, T.W. On the distribution of the two-sample Cramer-von Mises criterion. *Ann. Math. Stat.* **1962**, *33*, 1148–1159. [CrossRef]
44. Hagemann, S.; Chen, C.; Haerter, J.O.; Heinke, J.; Gerten, D.; Piani, C. Impact of a Statistical Bias Correction on the Projected Hydrological Changes Obtained from Three GCMs and Two Hydrology Models. *J. Hydrometeorol.* **2011**, *12*, 556–578. [CrossRef]