Hydrology and
Earth System
Sciences

Cross-validation of bias-corrected climate simulations is misleading

Douglas Maraun[1] **and Martin Widmann**[2]

[1]Wegener Center for Climate and Global Change, University of Graz, Brandhofgasse 5, 8010 Graz, Austria
[2]School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

Correspondence: Douglas Maraun (douglas.maraun@uni-graz.at)

Abstract. We demonstrate both analytically and with a modelling example that cross-validation of free-running bias-corrected climate change simulations against observations is misleading. The underlying reasoning is as follows: a cross-validation can have in principle two outcomes. A negative (in the sense of not rejecting a null hypothesis), if the residual bias in the validation period after bias correction vanishes; and a positive, if the residual bias in the validation period after bias correction is large. It can be shown analytically that the residual bias depends solely on the difference between the simulated and observed change between calibration and validation periods. This change, however, depends mainly on the realizations of internal variability in the observations and climate model. As a consequence, the outcome of a cross-validation is also dominated by internal variability, and does not allow for any conclusion about the sensibility of a bias correction. In particular, a sensible bias correction may be rejected (false positive) and a non-sensible bias correction may be accepted (false negative). We therefore propose to avoid cross-validation when evaluating bias correction of free-running bias-corrected climate change simulations against observations. Instead, one should evaluate non-calibrated temporal, spatial and process-based aspects.

1 Introduction

Bias correction is a widely used approach to postprocess climate model simulations before they are applied in impact studies (e.g. Gangopadhyay et al., 2011; Hagemann et al., 2013; Girvetz et al., 2013; Warszawski et al., 2014). A wide range of different correction methods has been developed, ranging from simple additive or multiplicative corrections to quantile-based approaches. For reviews of bias correction see

Teutschbein and Seibert (2012), Maraun (2016) and the book by Maraun and Widmann (2018).

The performance of a bias correction is typically evaluated against independent observational data, which have not entered the calibration of the correction function. For instance, Piani et al. (2010a, b), Li et al. (2010) and Dosio and Paruolo (2011) apply the holdout method, i.e. they calibrate the method on a calibration period and evaluate it on a non-overlapping validation period. Some authors even apply a full cross-validation, most often by permuting calibration and validation period (a 2-fold cross-validation; Gudmundson et al., 2012).

Cross-validation is a well-known and widely used statistical concept to assess the skill of predictive statistical models (Stone, 1974; Efron and Gong, 1983). It has been successfully applied in the atmospheric sciences, e.g. in weather forecasting (Jolliffe and Stephenson, 2003; Mason, 2008; Wilks, 2006) and perfect predictor experiments of downscaling methods (Themeßl et al., 2011; Maraun et al., 2015).

In climate change applications, however, the setting is typically different from a weather forecasting or perfect predictor setting: here, the model is running free, i.e. only external forcings are common to observation and simulation. Internal climate variability on all scales is independent and not synchronized. In this setting, the aim is not to assess predictive power, e.g. on a day-by-day or season-by-season basis, as in weather forecasting – in fact, by construction it cannot be assessed. Importantly, observed and simulated long-term trends may also differ substantially, just because of different random realizations of long-term modes of variability. Prominent examples of such modes are the Pacific Decadal Oscillation (PDO, Mantua et al., 1997) and the Atlantic Multidecadal Oscillation (AMO, Schlesinger and Ramankutty, 1994).

These differences have crucial implications for the application of cross-validation or any evaluation on independent data. Our results build upon a recent study by Maraun et al. (2017), who demonstrated that in a climate change setting cross-validation of marginal aspects is not able to identify bias correction skill. Here, we additionally show that the outcome of a cross-validation is essentially random and independent of the sensibility of the bias correction. We will demonstrate these consequences for the holdout method, but they can of course be generalized to any type of cross-validation.

We will discuss the specific context of climate change simulations in Sect. 2. An analytical derivation of the cross-validation problem will be given in Sect. 3, a modelling example in Sect. 4. We will close with a discussion of the implications of our findings.

2   Cross-validation in the climate context

Cross-validation was developed to quantify the predictive skill of statistical models in the 1930s, and has become widely used with the advent of modern computers (Stone, 1974; Efron and Gong, 1983). It has become a standard tool in weather and climate forecasting (Michaelsen, 1987; Jolliffe and Stephenson, 2003; Wilks, 2006; Mason, 2008).

The first major aim of cross-validation is to eliminate artificial skill: if the statistical model is evaluated on the same data that are used for calibration, the performance to predict new data will almost certainly be lower than the estimated skill. Hence, the model is calibrated only on a subset of the data, and evaluated on another – ideally independent – subset of the data. This so-called holdout method, however, uses each data point only either for calibration or validation and thus suffers from relatively high sampling errors.

The second major aim of cross-validation is therefore to use the data optimally. To this end, the holdout method, i.e. training and validation, is repeated on different subsets of the data. The simplest approach is the so-called split sample method, where the data are just split once into two subsets. More advanced k-fold cross-validation splits the data set into k non-overlapping blocks; in each fold, k–1 blocks are used for calibration and the remaining block is used for validation.

In weather and climate predictions, the aim is to predict the weather, i.e. internal variability, with a given lead time (say, 3 days or a season) at a desired timescale (say, 6 h or a season). A typical evaluation assesses how well certain meteorological aspects are predicted: in weather forecasting, one may for instance be interested in the overall prediction accuracy, measured by the root-mean squared error between predicted and observed daily time series. In a seasonal prediction, one may be interested in the bias of the predicted mean, or in the bias of the predicted wet-day frequency over a season. In this context, a cross-validation makes perfect sense

if the validation blocks are long compared to the prediction lead time (and process memory).

Downscaling and bias correction methods are typically tested in perfect predictor or perfect boundary condition experiments (Maraun et al., 2015), where predictors or boundary conditions are taken from reanalysis data. The aim of the downscaling in this context is not to predict internal variability ahead into the future, but rather to predict the local weather conditional on the state of the large-scale weather (i.e. to simulate the correct local long-term weather statistics). Still, in such a setting cross-validation makes perfect sense: the choice of reanalysis data as predictors/boundary conditions synchronizes simulated and observed local variability on timescales beyond a few weeks, such that the evaluation framework is similar to the case of seasonal prediction.

In free-running climate simulations, however, the situation is fundamentally different: here, any predictive power results only from external (e.g. anthropogenic) forcing at very long timescales, but internal variability is not synchronized at any timescale. Yet long-term modes of internal climate variability, such as the PDO (Schlesinger and Ramankutty, 1994) and the AMO (Schlesinger and Ramankutty, 1994), often mask forced climate trends even at multidecadal timescales (Deser et al., 2012; Maraun, 2013b). Thus, much of the difference between observed and simulated trends is not caused by model errors, but rather by random fluctuations of the climate system. This fact has strong implications for the evaluation of simulated trends (Bhend and Whetton, 2013; van Oldenborgh et al., 2013; Laprise, 2014), but it is also the reason why cross-validation of bias correction fails in this context.

As any cross-validation consists of repeated holdout evaluations, we will in the following only consider the holdout method. In Sect. 5 we will discuss how the following results generalize to a full cross-validation.

3   Analytical derivation

Consider a simulated time series x_i and an observed time series y_i. Assume that an evaluation addresses the representation of some statistic such as the long-term mean. Over the calibration period, we denote the simulated and observed means as $\overline{x}_{cal}$ and $\overline{y}_{cal}$, respectively. Correspondingly, we denote them as $\overline{x}_{val}$ and $\overline{y}_{val}$ over the validation period. Then an estimate for the bias over the calibration period is given as

$$\text{BIAS} = \overline{x}_{cal} - \overline{y}_{cal}. \tag{1}$$

Applying the bias estimate to the validation period, one obtains an estimate of the corrected mean over the validation period:

$$\overline{x}_{val}^{corr} = \overline{x}_{val} - \text{BIAS} = \overline{x}_{val} - \overline{x}_{cal} + \overline{y}_{cal}. \tag{2}$$

The remaining residual bias is then

$$\text{BIAS}_{res} = \overline{x}_{val}^{corr} - \overline{y}_{val} = \overline{x}_{val} - \overline{x}_{cal} + \overline{y}_{cal} - \overline{y}_{val}. \tag{3}$$

This residual bias can be expressed in terms of the observed and simulated climate change signals. The change signal from calibration to validation period is defined as

$$\Delta x = \overline{x}_{\text{val}} - \overline{x}_{\text{cal}} \tag{4}$$

for the model and

$$\Delta y = \overline{y}_{\text{val}} - \overline{y}_{\text{cal}} \tag{5}$$

for the observations. Thus, the residual bias is given as

$$\text{BIAS}_{\text{res}} = \Delta x - \Delta y. \tag{6}$$

For variables such as precipitation, one often considers relative changes. Here a corresponding derivation holds. The relative error is defined as

$$\text{RE} = \overline{x}_{\text{cal}}/\overline{y}_{\text{cal}}, \tag{7}$$

and the corrected mean over the validation period is given as

$$\overline{x}_{\text{val}}^{\text{corr}} = \overline{x}_{\text{val}}/\text{RE} = \overline{x}_{\text{val}} \cdot \overline{y}_{\text{cal}}/\overline{x}_{\text{cal}}. \tag{8}$$

The residual relative error results in

$$\text{RE}_{\text{res}} = \overline{x}_{\text{val}}^{\text{corr}}/\overline{y}_{\text{val}} = \frac{\overline{x}_{\text{val}} \cdot \overline{y}_{\text{cal}}}{\overline{x}_{\text{cal}} \cdot \overline{y}_{\text{val}}}. \tag{9}$$

The relative change signal from calibration to validation period is defined as

$$\Delta x = \overline{x}_{\text{val}}/\overline{x}_{\text{cal}} \tag{10}$$

for the model and

$$\Delta y = \overline{y}_{\text{val}}/\overline{y}_{\text{cal}} \tag{11}$$

for the observations. Hence, the residual relative error is

$$\text{RE}_{\text{res}} = \Delta x/\Delta y. \tag{12}$$

The residual bias or relative error could further be tested for significance, i.e. whether the bias-corrected statistic $\overline{x}_{\text{val}}^{\text{corr}}$ is significantly different from the observed statistic $\overline{y}_{\text{val}}$ over the validation period. Thus, a holdout evaluation will yield a positive result (in the sense of rejecting the null hypothesis, i.e. a non-zero residual bias) if the simulated change Δx is different from the observed change Δy, and a negative result (i.e. a residual bias compatible with zero) if simulated and observed changes are indistinguishable.

Assume now that a given bias correction may or may not be sensible. Note in this context that it is completely irrelevant to explicitly define what constitutes a sensible bias correction (but for a brief discussion see Sect. 4). Thus, in principle four cases are possible.

1. True negative: the bias correction is sensible, and the (bias-corrected) climate model simulates a trend closely resembling the observed trend.

2. False positive: the bias correction is sensible, but due to internal climate variability, the (bias-corrected) climate model simulates a trend different from the observed trend.

3. False negative: the bias correction is not sensible, but the (bias-corrected) climate model for some reason simulates a trend similar to the observed trend. This case corresponds to the example given in Maraun et al. (2017).

4. True positive: the bias correction is not sensible, and the (bias-corrected) climate model simulates a trend different from the observed trend.

The crucial point is that for typical record lengths, much of the difference between simulated and observed changes Δx and Δy will be caused by internal climate variability. Thus the result of a cross-validation, i.e. which of the four cases occurs, is mostly random and says only little about the sensibility of the cross-validation.

Maraun et al. (2017) considered case 3: as the difference between simulated and observed trends on typical timescales of a few decades is dominated by internal variability, the holdout method is not suitable to identify a non-sensible bias correction. The reverse conclusion is that the holdout method – and consequently also a cross-validation – is not able to corroborate whether a bias correction is sensible.

Yet the discussion above implies an even stronger conclusion: because case 2 might randomly occur, a sensible bias correction may be rejected by a cross-validation. Thus, even more importantly, cross-validation in the given context is not just useless, but even misleading.

4 Empirical demonstration

To further illustrate the analytic findings, we will give examples of the four cases in an exaggerated modelling example. We consider mean summer (JJA) precipitation at four locations. As observational reference we select the E-OBS data set (Haylock et al., 2008). As calibration period we use 1956–1980, as evaluation period 1981–2005.

We need to select two examples where the given bias correction is sensible, and two where it is not. Finding a convincing example of a sensible bias correction has to rely on process understanding (Maraun et al., 2017). A major precondition is that the climate model simulates a realistic present climate and a credible climate change (Maraun and Widmann, 2018). The former condition mainly involves a realistic representation of the large-scale circulation (Maraun et al., 2017). We therefore consider the following set-up: as examples of a sensible bias correction, we consider summer mean precipitation at two locations in Norway. Summer mean precipitation in Norway is dominated by large-scale precipitation, which can sensibly be assumed to be
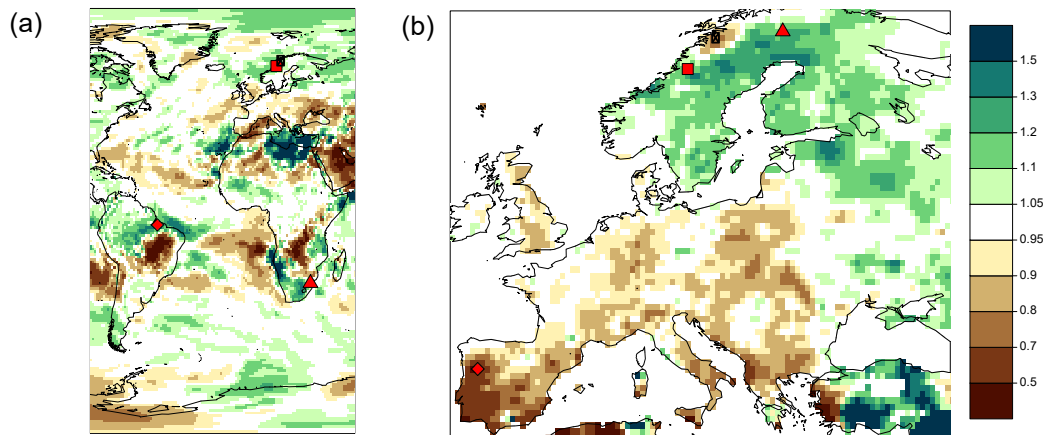
Figure 1. Maps of relative changes in boreal mean summer (JJA) precipitation, 1981–2005 relative to 1956–1980. **(a)** EC-EARTH, **(b)** E-OBS, version 15.0. Square: case 1 (true negative); circle: case 2 (false positive); triangle: case 3 (false negative); diamond: case 4 (true negative).

realistically simulated by current-generation general circulation models (GCMs). Specifically we choose a transient simulation of EC-EARTH (Hazeleger et al., 2010), a model which has been demonstrated to suffer from minor biases in the synoptic-scale atmospheric circulation over Europe only (Zappa et al., 2013). We assume that other potential problems such as mislocations (Maraun and Widmann, 2015) or scale gaps (Maraun, 2013a) are negligible for the considered locations and timescales. In this setting, we argue that a bias correction is in principle sensible. Two slightly different locations have been selected (Fig. 1): in the Børgefjell region north-east of Trondheim observed and simulated trends are very similar (case 1). Further north, around the town of Bodø, the two trends are very different as observed precipitation has decreased whereas simulated precipitation has increased. It is reasonable to assume that the observed negative trend is caused by internal climate variability and not a forced change. Thus, in principle the bias correction is sensible even though the validation rejects it (case 2). Here, the key point is not whether the bias correction in this particular example is really sensible or not, but that case 2 is indeed a possible – and misleading – outcome of a bias correction.

In the following we show two examples where a bias correction is not sensible. A discussion about the question when a bias correction makes no sense would go very much beyond the scope of this piece. Therefore, we follow the logic of Maraun et al. (2017) and select examples where model simulation and observation are taken from geographically far away and climatically rather different regions. The underlying idea is that for such cases, the model does not represent the target variable such that a bias correction is without doubt not sensible. Specifically, we consider the following two cases (see Fig. 1): first, mapping simulated boreal summer mean precipitation from the sub-tropical Maputo area (Mozambique, close to the South African border) to the Taiga region of the

Norwegian–Finnish boarder. Here, observed and simulated trends are randomly similar. This example is equivalent to that given in Maraun et al. (2017), where a non-sensible bias correction is not identified by the validation (case 3). Second, we map summer mean precipitation from the tropical climate at Belén in the Amazon delta to the mild and maritime climate of northern Portugal. Here, observed and simulated trends are randomly very different: positive in the Amazon delta, negative in Portugal. The non-sensible bias correction is thus correctly identified (case 4). In these two examples, it was a priori obvious that a bias correction is not sensible. In real applications, of course, such a priori reasoning will be much more difficult and has to rest upon process understanding (see discussion above). The key point, again, is that case 3 is a possible outcome of a bias correction.

Figure 2 shows observed and simulated time series, the latter before and after bias correction, for the four cases we considered. Figure 2a and b show the sensible examples, Fig. 2c and d the non-sensible examples. In Fig. 2a and c, observed and simulated trends randomly agree, and in Fig. 2b and d they randomly disagree. As shown analytically in Sect. 3, the residual bias vanishes in cases (a) and (c), where the relative trends in observations and simulations are similar, and it does not vanish in (b) and (d), where the relative trends in observations and simulations disagree. The relevant cases are (b) and (c): in the former, the bias correction is in principle sensible, but the holdout method would suggest that it was not sensible (false positive). In the latter, the bias correction is not sensible, but the holdout method would suggest that it was sensible (false negative). These examples clearly illustrate our previous reasoning: the holdout method, and thus also cross-validation, yields misleading results when it is used to assess the sensibility of bias-corrected climate change simulations against observations.
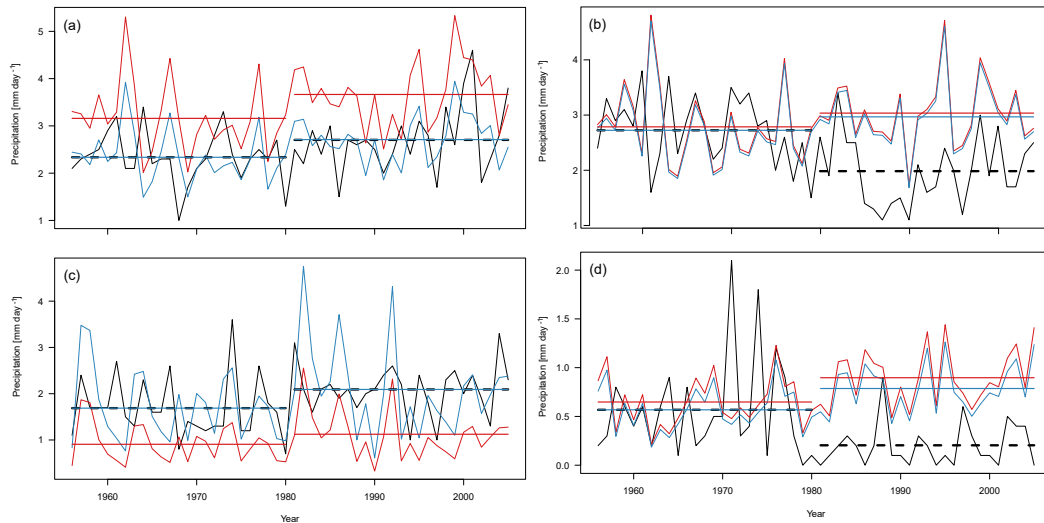
Figure 2. Time series of boreal summer (JJA) precipitation. **(a)** case 1 (true negative); **(a)** case 2 (false positive); **(c)** case 3 (false negative); **(d)** case 4 (true negative). Black: E-OBS; red: raw EC-EARTH; blue: bias-corrected EC-EARTH. The straight horizontal lines depict the long-term means over calibration and validation periods, respectively.

5 Conclusions

We have demonstrated both analytically and with a modelling example that cross-validation of free-running bias-corrected climate change simulations against observations is misleading. The underlying reasoning is as follows: the result of a cross-validation – a significant or non-significant residual bias in the validation period – depends on the difference between observed and simulated changes between calibration and validation periods. For typical lengths of calibration and validation periods, these differences depend mostly on the realizations of internal variability in the observations and climate model. These differences therefore do not allow for conclusions about the sensibility of a bias correction. As in any setting of significance testing, four cases are possible: true negative, false positive, false negative and true negative. The actual outcome in a given application is mostly random.

The relevance of internal variability in the discussed cross-validation context depends on the relative strength of internal variability and forced trends, and the length of the calibration and validation period compared to the periodicity of the dominant modes of internal climate variability. In tropical climates, interannual variability such as that of El Niño–Southern Oscillation dominates climate variability. Thus, relatively short periods of a few decades may suffice to obtain stable estimates of forced changes between calibration and validation period, and therefore to assess whether a bias correction performs well given the observed changes. In mid-latitude climates, however, the dominant modes of internal variability have periodicities of several decades (Mantua et al., 1997; Schlesinger and Ramankutty, 1994) such that stable estimates of forced changes will in general not be obtained for the typical calibration and validation periods of

a few decades only. The strength of internal variability depends also on the chosen variable (e.g. higher for precipitation than for temperature) and index (higher for extremes than for long-term means). Spatial aggregation will reduce regional short-term internal variability (Maraun, 2013b), but will not affect long-term modes of variability, which typically have coherent spatial patterns.

We have derived these conclusions for the mean and the holdout method, where the bias correction is calibrated against one part of the data and validated against its complement. Yet the results can in principle be transferred to other statistics such as variances or individual quantiles, and to a full cross-validation. The residual mean bias, however, is always zero in a full cross-validation, as long as the individual folds have the same length. The reason is that changing the calibration and validation period changes the sign of the residual bias. When averaging the residual bias across the different folds, it cancels out. For the variance or similar statistics, the outcome depends on the way the cross-validation is carried out: if the residual bias is calculated for each fold separately and then averaged (as suggested in the classical literature), the behaviour is as for the mean. If the residual bias is calculated over a concatenated cross-validated time series (as is typically done in the atmospheric sciences), the bias correction in cases (b) and (d) will yield extremely high residual biases (because the shift in the mean is not removed in the variance calculation).

The consequence of these findings is that cross-validation should not be used when evaluating bias correction of free-running climate simulations against observations. In fact, a framework for evaluating bias correction of climate simulations is still missing and not trivial. As discussed in Maraun et al. (2017), we propose to evaluate non-calibrated tempo-

ral, spatial and process-based aspects of the simulated time series. Whether a bias correction makes sense for climate change projections depends in particular on the realism and the credibility of the underlying climate model simulation. A process-based evaluation is thus a key prerequisite for a successful bias correction (Maraun et al., 2017).

Author contributions. DM and MW developed the idea for the study. DM conducted the analytical derivations, carried out the analysis, and wrote the manuscript. MW commented on the manuscript. DM and MW discussed the results.

References

Bhend, J. and Whetton, P.: Consistency of simulated and observed regional changes in temperature, sea level pressure and precipitation, Clim. Change, 118, 799–810, 2013.

Deser, C., Knutti, R., Solomon, S., and Phillips, A.: Communication of the role of natural variability in future North American climate, Nat. Clim. Change, 2, 775–779, 2012.

Dosio, A. and Paruolo, P.: Bias correction of the ENSEMBLES high resolution climate change projections for use by impact models: Evaluation on the present climate, J. Geophys. Res. Atmos., 116, D16106, https://doi.org/10.1029/2011JD015934, 2011.

Efron, B. and Gong, G.: A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Stat., 37, 36–48, 1983.

Gangopadhyay, S., Pruitt, T., Brekke, L., and Raff, D.: Hydrologic projections for the Western United States, EOS, 92, 441–442, 2011.

Girvetz, E., Maurer, E., Duffy, P., Ruesch, A., Thrasher, B., and Zganjar, C.: Making climate data relevant to decision making: the important details of spatial and temporal downscaling, The World Bank, 2013.

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

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss, F., and Wiltshire, A. J.: Climate change impact on available water resources obtained using multiple global climate and hydrology models, Earth Syst. Dynam., 4, 129–144, https://doi.org/10.5194/esd-4-129-2013, 2013.

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res., 113, D20119, https://doi.org/10.1029/2008JD010201, 2008.

Hazeleger, W., Severijns, C., Semmler, T., Ştefănescu, S., Yang, S., Wang, X., Wyser, K., Dutra, E., Baldasano, J., Bintanja, R., Bougeault, P., Caballero, R., Ekman, A., Christensen, J., van den Hurk, B., Jimenez, P., Jones, C., Kållberg, P., Koenigk, T., Mc Grath, R., Miranda, P., van Noije, T., Palmer, T., Parodi, J., Schmith, T., Selten, F., Storelvmo, T., Sterl, A., Tapamo, H., Vancoppenolle, M., Viterbo, P., and Willen, U.: EC-Earth: a seamless earth-system prediction approach in action, B. Am. Meteorol. Soc., 91, 1357–1363, 2010.

Jolliffe, I. and Stephenson, D., eds.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley, 2003.

Laprise, R.: Comment on "The added value to global model projections of climate change by dynamical downscaling: A case study over the continental US using the GISS-ModelE2 and WRF models" by Racherla et al., J. Geophys. Res., 119, 3877–3881, 2014.

Li, H., Sheffield, J., and Wood, E.: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching., J. Geophys. Res., 115, D10101, https://doi.org/10.1029/2009JD012882, 2010.

Mantua, N. J., Hare, S., Zhang, Y., Wallace, J., and Francis, R.: A Pacific interdecadal climate oscillation with impacts on salmon production, B. Am. Meteorol. Soc., 78, 1069–1079, 1997.

Maraun, D.: Bias Correction, Quantile Mapping and Downscaling: Revisiting the Inflation Issue, J. Climate, 26, 2137–2143, 2013a.

Maraun, D.: When will trends in European mean and heavy daily precipitation emerge?, Env. Res. Lett., 8, 014004, https://doi.org/10.1088/1748-9326/8/1/014004, 2013b.

Maraun, D.: Bias Correcting Climate Change Simulations – a Critical Review, Curr. Clim. Change Rep., 2, 211–220, https://doi.org/10.1007/s40641-016-0050-x, 2016.

Maraun, D. and Widmann, M.: The representation of location by a regional climate model in complex terrain, Hydrol. Earth Syst. Sci., 19, 3449–3456, https://doi.org/10.5194/hess-19-3449-2015, 2015.

Maraun, D. and Widmann, M.: Statistical Downscaling and Bias Correction for Climate Research, Cambridge University Press, Cambridge, 2018.

Maraun, D., Widmann, M., Gutierrez, J., Kotlarski, S., Chandler, R., Hertig, E., Wibig, J., Huth, R., and Wilcke, R.: VALUE: A Framework to Validate Downscaling Approaches for Climate Change Studies, Earth's Future, 3, 1–14, 2015.

Maraun, D., Shepherd, T., Widmann, M., Zappa, G., Walton, D., Hall, A., Gutierrez, J. M., Hagemann, S., Richter, I., Soares, P., and Mearns, L.: Towards process-informed bias correction of climate change simulations, Nat. Clim. Change, 7, 764–773, https://doi.org/10.1038/NCLIMATE3418, 2017.

Mason, S.: Understanding forecast verification statistics, Meteorol. Appl., 15, 31–40, 2008.

Michaelsen, J.: Cross-validation in statistical climate forecast models, J. Clim. Appl. Meteorol., 26, 1589–1600, 1987.

Piani, C., Haerter, J., and Coppola, E.: Statistical bias correction for daily precipitation in regional climate models over Europe, Theor. Appl. Climatol., 99, 187–192, 2010a.

Piani, C., Weedon, G., Best, M., Gomes, S., Viterbo, P., Hagemann, S., and Haerter, J.: Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models, J. Hydrol., 395, 199–215, 2010b.

Schlesinger, M. and Ramankutty, N.: An oscillation in the global climate system of period 65–70 years, Nature, 367, 723–726, 1994.

Stone, M.: Cross-validatory choice and assessment of statistical predictions, J. Roy. Stat. Soc. B, 32, 111–147, 1974.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, J. Hydrol., 456, 12–29, 2012.

Themeßl, M. J., Gobiet, A., and Leuprecht, A.: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models, Int. J. Climatol., 31, 1530–1544, 2011.

van Oldenborgh, G., Doblas Reyes, F.-J., Drijfhout, S., and Hawkins, E.: Reliability of regional climate model trends, Environ Res. Lett., 8, 014055, https://doi.org/10.1088/1748-9326/8/1/014055, 2013.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework, Proc. Nat. Acad. Sci., 111, 3228–3232, https://doi.org/10.1073/pnas.1312330110, 2014.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press/Elsevier, 2 Edn., 2006.

Zappa, G., Shaffrey, L., and Hodges, K.: The ability of CMIP5 models to simulate North Atlantic extratropical cyclones, J. Climate, 26, 5379–5396, 2013.