Responses by the author (in green) to comments (in black) by anonymous referee #1

1. General Comments

[…]

My main concern is that the author provides some improvement to the description of results, particularly in terms of figures. I am aware that comparison among 8*3 methods, adopting different parameters over LW and SW radiation fields separately, requires a challenging effort in terms of clarity and conciseness. In some parts of the manuscript I found difficult to benchmark arguments described in the text with the mentioned figures. I will be more specific in the next section.

The results section has been almost completely rewritten and the figures have been made clearer, see my responses to your specific comments.

Another aspect that I think might be improved is a discussion of the implications of using a deterministic parametric method, rather than a stochastic one, for bias correction when a downscaling/upscaling is made necessary. A reference to Maraun, 2013 (JCLI) might be helpful in this respect. Related to this, a further appendix may be suitable, not only including such a discussion but also a basic description of the quantile mapping methodology for those who are not familiar with it. In the current draft, this is left to references although, as far as I could check, none of the mentioned papers explicitly addresses for the quantile mapping methodology.

I appreciate that not every reader is familiar with the quantile mapping (QM) methodology. Since also anonymous referee #2 asked for it, I have added Appendix A that includes a general description of QM and touches on parametric versus non-parametric as well as deterministic versus stochastic QM.

2. Specific Comments

Figure 2: it was very difficult to me to distinguish among the various lines shown in the panels. The dotted red and dashed blue lines are almost indistinguishable (particularly in (b) and (c)) and the light blue line in (a) can hardly be seen. I would suggest to split this figure in two, separately showing the beta and advanced distributions respectively, with the related parameters. As for the caption, I would suggest to explain in first place on which data the computation of the distributions and their parameters is based.

As to the caption, I followed your suggestion. I did not want to split the figure in two as suggested because the figure is supposed to illustrate similarities and differences between the different QM methods and that would be difficult if different methods were shown in different figures. However, I have simplified the plot by removing the lowermost and uppermost dotted red and dashed blue lines as these were a mere bonus (they just showed that the distribution fitting works well). Also, I have made the light blue line green and added the following sentence to the figure caption: “Note that the basic and advanced estimates of mean values and standard deviations only differ in panel (c) near the beginning and end of polar night (cf. Table 1).” This should clarify that it is not a bug but a feature that the dotted red and dashed blue lines are mostly indistinguishable.

Table 2: I wonder if one could improve the notation for distribution parameters and arrange it with a more mathematically appropriate symbols. Rather than plain text and footnotes, you may want to introduce a consistent notation with brackets and apostrophes to indicate means, running means and variances, as well as apexes and subscripts referring to the length of the window and the amount of

years to be considered.

Thank you very much for this suggestion. I have introduced such a mathematical notation in the revised manuscript.

l. 32-33, p. 9: it may be worth mentioning here how the common factor for the aggregation of bias-corrected values in the SRB-grid cell is chosen.

I have rewritten this paragraph such that it is now clearer how the common factor is determined.

l. 22, p. 10: As far as I understood the common factor f(i,j) is not the same as for the aggregation to the SRB-grid cell, given that it depends on whether the bias correction is applied on the lower or higher resolution. If it is not the case, it is once again not clear to me how the value of this common factor is chosen (see previous comment).

I have also adjusted this part such that it should be clear how f_{ij} is calculated.

l. 33-34 p. 12: the limits of parametric methods are here correctly mentioned. As stated in the General Comments section, this is a critical issue, and I think it would be worthwhile a few more arguments. If it is not too much work, I wonder if it would be possible to apply a non-parametric quantile mapping (e.g. using a cubic spline empirical CDF) to be compared with these parametric methods.

The number of QM methods compared in this study is already quite large. Also testing non-parametric QM methods is beyond the scope of the article. However, I have added a paragraph to Section 5 that discusses potential benefits of using non-parametric QM methods compared to the parametric QM methods tested here.

l. 15-16 p. 14: looking at Figure 6 is very hardly distinguishable that the BCvmp1 at the daily time scale outperforms the same methods at the monthly time scale. This is in my opinion because Figure 6, as well as Figure 2, contains too much information that prevents from emphasizing the key points that are described in the text. The uncertainty range masks the differences among the bars. Furthermore, having five bars for every months makes very difficult to distinguish them, particularly the ones in lighter colours (BCvmp1 methods). I would suggest to split the figures in order at least to separately consider original and bias corrected p-values.

Again, I think that it would not help to split the figure as suggested because plotting p-values before and after bias correction using the same scale is needed in order to illustrates the effect of the bias correction. Yet I appreciate that there are quite many box-whisker plots in the figure, so I have reduced the plot's temporal resolution from monthly to seasonal. Also, I have reduced the range of the y-axis from [-14, 0] to [-10, 0], which has made differences between the individual box-whisker plots more easily distinguishable.

l. 9-11 p. 16 and l. 1-2 p. 18: I found very challenging to carve out the important information from Figures 7-8 and link it with the arguments in the text. It seems to me that the only clear information that can be driven from them is that BCvdax methods outrank BCvdax at the daily resolution for what concerns rlds, and the other way round for what concerns rsds and rlds in the monthly mean. The author refers to a tropical/extratropical asymmetry that to my best effort is barely distinguishable. Furthermore the seasonal dependence (if any) is not mentioned in the text, still making the clarity of the two figures even more arguable. I would suggest either to restructure the layout of Figures 7 and 8 or removing this part, since it does not add much to the discussion of results.

Since referee #2 also revealed several substantial shortcomings in this part of the manuscript, the entire validation against BSRN observations has been removed from the revised manuscript.

3. Technical comments

l. 6 p. 7 (and elsewhere in the text): replace “Sect.” with “Appendix”, when you reference to appendices.

I have done as suggested.

l. 5 p. 9: correct “it”.

I have done as suggested.

l. 8 p. 10: maybe “be” is needed between “to” and “made”.

I have done as suggested.

l. 11 p. 14: “that” is repeated twice.

I have substituted “this” for the second “that”.

Responses by the author (in green) to comments (in black) by anonymous referee #2

1. General comments

A first concern is the focus of the paper: is the focus the evaluation of different methods or the quantitatively correct bias correction rsds and rlds in an absolute sense? Overall, the paper seems to suggest the former (comparison of methods). However, the use of BSRN data as an independent quantitative check points to the later (quantitatively correct rsds and rlds in an absolute sense). If the latter is indeed part of the goal, more work has to go into ascertaining the quantitative correctness of the SRB data used for bias adjustment.

Many thanks to referee #2 for her comprehensive criticism of the validation agains independent surface observations. After carefully consulting the concerns presented and literature provided by the referee I have decided to completely remove this part of the manuscript. Indeed, the validation was a secondary goal of the paper, which clearly benefits from focusing on its main goal, which is the evaluation of the different quantile mapping (QM) methods.

A second major point is the overall clarity of the manuscript. The methods used are complex, the figures shown are (too) packed with interesting information. However, explanations and descriptions come in often (very) long sentences, with lots of details, making it difficult to grasp the essentials. More focused and shorter sentences would help, as would some more information (possibly equations) on the parametric methods. The reason for specific choices (e.g. why comparing these methods, why using these metrics?) are not given. Conclusions read in wide parts more like an extensive summary.

Since referee #1 also pointed to too packed figures, I have reduced their information content to some extent in the revised manuscript. Also, I have almost entirely rewritten the results and conclusions sections using shorter sentences. These parts are now better structured, more focused and concise. Reasons for choices of methods and metrics are now better motivated.

Ideally, the statement that there are two best methods (one for rsds the other for rlds, and measured in terms of cross-validation) would be further embedded. Can these methods be used for bias correction of the entire E2OBS period without introducing artifacts? Could the methods be further improved? Are the other methods just slightly or clearly worse?

The methods can definitely be used for bias correction of the entire E2OBS period, see my response to your specific comment below. The relative performances of the different methods are now better described in the conclusions section.

2. Specific comments

p.3, l.27: Why use to different versions of SRB for rlds and rsds?

These are the latest available versions of the SRB dataset. The version numbers differ between rlds and rsds. This is now explained.

p.4, l.9: "If deviations of SRB from SRBQC data quantify methodological uncertainty inherent to SRB data then these findings justify the bias correction of E2OBS rlds and rsds using SRB data over land at least." Two points here. For rsds, one may argue on the same ground that wide parts of the oceans also need adjustment. More generally, you assume here that SRB is correct (at least more correct than E2OBS). How can you be sure? For example, how does SRB compare to CERES data?

Or to global mean estimates of rsds and rlds? A number of papers, e.g. by Trenberth et al., give numbers for the latter. An alternative may be to focus only on the methods and not argue at all about the quality of the SRB data.

In the revised manuscript I have focused only on the methods and do not argue at all about the quality of the SRB data.

Figure 1: Which of the differences are statistically significant?

This figure has been removed from the manuscript, in line with focusing on the methods.

Table 1: How about the altitude dependence of short wave radiation? (See e.g. Marty, Philipona, Frohlich, Ohmura, Theor. Appl. Climatol. 2002)

Also this table has been removed from the manuscript (and along with it the question of how shortwave radiation changes with altitude), in line with focusing on the methods.

p.6, l.6: What do you mean by bilinear interpolation from coarse (SRB) to fine (E2OBS) grid? Copying? Same question on p.11, l.18.

I have changed “bilinearly interpolated” to “spatially bilinearly interpolated” in both cases. I think this is a standard term, which does not need further explanation.

p.6, l.8: "For the BCvtp2 methods, the sub-SRB-grid scale spatial structure of the original E2OBS data is imposed upon spatially disaggregated SRB data prior to bias correction at the E2OBS grid." Please try to clarify. I think I understood much later, in Section 3.2.1, that you adjust the mean and variance of E2OBS data on the E2OBS grid with mean and variance of SRB data on the corresponding, coarser SRB gird. True?

I have changed this sentence to “the BCvtp2 methods adjust mean values and variances at the E2OBS grid such that mean values and variances of spatial aggregates to the SRB grid match the corresponding SRB estimates while the sub-SRB-grid scale spatial structure of mean values and variances present in the original E2OBS data is retained.”

p.6, l.14: "... of the underlying four E2OBS values." The two grids thus are such that four E2OBS cells correspond to one SRB cell? They are not shifted against each other?

Correct. I have added the sentence “ Every SRB grid cell contains exactly four E2OBS grid cells.” to the data description section.

p.6, l.16: It would be helpful if you added some information, possibly equations, on transfer functions, target distributions, estimation of means and variances of beta functions etc. in an appendix, as these are absolutely central to your study. Currently, the reader has to know all this or has to check out the references. After all, you even devote an appendix to explaining Kolmogorov-Smirnov.

Thank you for pointing this out. Such an appendix has been added to the revised manuscript.

Figure 2d: Why are the colored lines so far away from the black and gray lines?

Because my estimates of the upper bounds of monthly mean radiation are calculated based on the upper bounds to the corresponding daily mean radiation. The resulting upper bounds are typically

much larger than observed maximum monthly mean radiation because 31 consecutive days of daily mean radiation at its physical upper limit are very unlikely to occur in reality. I have added this explanation to Sect. 3.1.2.

p.7, l.8: What do you mean by "The rsdt climatology at a given latitude is rescaled such that it sits just above the multi-year maximum..."? Why do that?

To answer your questions, I have rewritten the beginning of this paragraph as follows: “The BCsda1 method employs the climatology of daily mean shortwave insolation at the top of the atmosphere (rsdt; see Appendix B for how rsdt is calculated in this study) for the upper bound estimation. This is motivated by rsds being limited by rsdt in most locations and seasons, which suggests that the annual cycle of the upper bound of daily mean rsds has a similar shape as the climatology of daily mean rsdt. Therefore, method BCsda1 uses a rescaled daily mean rsdt climatology as the upper bound climatology of daily mean rsds (solid blue line in Fig. 1c). The rescaling is done with the smallest possible factor which guarantees that the resulting upper bounds are greater than or equal to the multi-year maximum values of daily mean rsds on all days of the year with rsdt ≥ 50 W m-2. An extension of this guarantee to days of the year with lower rsdt would inflate the rescaling factor because during dusk and dawn of polar night, rsds can exceed rsdt due to diffuse radiation coming in from lower latitudes. Therefore, on days of the year with rsdt < 50 W m-2, the maximum of the rescaled rsdt and the empirical multi-year maximum daily mean rsds is used as the upper rsds bound.”

p.10, l.9: "... one possibility to define ..." What would other possibilities be? Why your choice?

Another possibility would be to follow the BCvtp0 approach, i.e. to use interpolated data. The motivation of my choice is that it solves the problem illustrated and discussed in Sect. 4.2. I have rephrased the sentence as follows: “With target distributions fixed at the SRB grid, target distributions at the E2OBS grid can be defined such that the bias-corrected data have the SRB-grid scale target distributions and the sub-SRB-grid scale structure of the original E2OBS data.”

p.10, Eq. 1: Where does the equation come from? Can you give a reference? The explanation following eq. 1 reads rather lengthy but not too clearly.

This does not need any reference. It is the standard formula for the variance of a linear combination of random variables. I have however inserted one intermediate step using covariances in the equation to make its derivation easier to understand.

p.11, l.9: How often does this "99%" condition kick in?

For longwave (shortwave) radiation, this "99%" condition kicks in over four (11% of all) grid cells and there on 15% (5%) of all days of the year. I have added this information to the revised manuscript version.

p.11, l.16: How often does this "40%" condition kick in?

The “40%” condition is never met for longwave radiation whereas for shortwave radiation it kicks in over 14% of all E2OBS grid cells and there on 2% of all days of the year. I have added this information to the revised manuscript version.

p.11, l.27: "Metrics used..." Why these? Why, for example, skewness? What do I learn from this measure? And why a Kolmogorov-Smirnov test? Why not a test that gives more weight to tails, e.g. Anderson-Darling? More generally, when do you say that your bias adjustment is good? When the

adjusted E2OBS distribution is identical (mean, variance, skewness...) to the SRB distribution? Why then adjust at all and not just take the SRB data? Can you use your method to adjust E2OBS data beyond the time span where SRB data is available?

The skewness is included because it is the first distribution moment which is not explicitly adjusted by my parametric QM methods. It is included here to illustrate this conceptual imperfection of my methods. I have included this motivation in the revised results section. You are right about the KS test and the relatively low weight it gives to tails. In the revised section 4.1, I have included Kuiper's test as one that (like the suggested AD test) gives the same weight to CDF differences at all quantiles. Qualitatively, however, the Kuiper's test results are the same as those of the KS test. You are right that I (and, as far as I know, everybody else who cross-validates bias correction methods) consider a bias adjustment good if the adjusted distributions are identical to the target distributions. I have included this definition of (overall) performance in the revised section 4.1. In the ISIMIP framework, there are two reasons for doing the bias adjustment of E2OBS to SRB data and not just using the SRB data directly: It (i) promises a higher inter-variable consistency (e.g. consistency of temperature and longwave radiation) in the EWEMBI dataset and (ii) produces radiation data that cover a longer time span. Applying the methods to E2OBS data beyond the time span where SRB data are available is fine since the 1979-2013 period is in fact not much larger than the 1983-2007 period, so that the former is expected to be sufficiently well represented by the latter.

p.12, l.2: Does the remark about CVCC imply that your method cannot be used to correct E2OBS data outside the SRB period (1983-2007)?

No, it does not, see my response to your previous comment.

p.12, l.11: "In the following, cross-validation results are only shown and discussed for the BCvtp0 and BCvtp1 methods, since results for the corresponding BCvtp1 and BCvtp2 are virtually identical." What do you mean? That the difference between BCvtp0 and BCvtp1 is similar as between BCvtp1 and BCvtp2? And, consequently, BCvtp0 and BCvtp2 differ more?

No, I mean that cross-validation results for the BCvtp1 and BCvtp2 methods are virtually identical. In order to make this clearer I have rewritten the statement as follows: "Please note that results for BCvtp2 are not shown or discussed in this section because BCvtp1 and BCvtp2 produce virtually identical data at the SRB-grid scale."

p.12, l.17: "... overall performance ..." What do you mean by overall performance?

This is now better explained in the new section 4.1.3.

p.12, l.24: Why now looking at relative differences?

Why not? For standard deviations, I think this makes more sense than to look at absolute differences.

Figure 3: I guess a good bias correction in your metrics results in a white map. True? The color / hue coding may be better explained upon first use.

Not true. White means low agreement in bias direction (positive or negative bias) over months and validation data samples. In order to make this clearer I have added the following sentence to the caption of this figure: "More saturated colours indicate higher statistical significance of biases remaining after bias correction."

Figures 4 and 5: Why are the quantities shown of interest? And, again, what is good and what is bad? If white means "good", then none of the methods performs well here?

The cross-validation of multi-year maximum values shall reveal if it is worthwhile and if so, then how to explicitly adjust upper radiation bounds. This is now better explained in section 4.1.2. For why skewness is of interest, see my response to your comment on p.11, l.27. As to the significance of Figure 5, see my answer to your next question. In terms of what white means, see my answer to your previous comment. The methods are clearly not perfect but I also did not expect that. It does not make sense to make an absolute statement such as “this shows that the method performs well.” The only sensible question is if one method performs better than another one. Figures 2 to 4 (formerly 3 and 5) quantify the magnitude of biases of selected statistics that remain after bias correction with different methods.

p.14, l.15: Why should bias adjustment on monthly timescales outperform daily bias adjustment with subsequent monthly averaging?

Because of what is shown in Figure 4 (formerly 5). I have revised the explanation of Figure 4 earlier in the text as follows in order to answer your question: “By design, the BCvdpx and BCvmpx methods are equally good at correcting multi-year mean values of daily mean radiation. However, both day-to-day and year-to-year variability are expected to be differently well corrected by the methods operating at different time scales. Since day-to-day variability is (not) explicitly adjusted by the methods operating at the daily (monthly) time scale the BCvdpx methods are expected to perform better at the daily time scale than the BCvmpx methods. The year-to-year variability, on the other hand, is explicitly corrected by the BCvmpx methods and it is not by the Bcvdpx methods because daily data from different years are pooled before quantile mapping is carried out at the daily time scale. Consequently, biases in interannual standard deviations of monthly mean radiation are much larger after bias correction with BCvda1 than with BCvma1 (Fig. 4), and the BCvmpx methods are generally expected to perform better at the monthly time scale than the BCvdpx methods.”

p.15, l.3: "Rather, the p-value distributions depicted in Fig. 6b,d suggest that if sampling errors are taken into account then the BCvdp1 methods correct the distributions of monthly mean values almost as well as the BCvmp1 methods." I do not see this point from the text and / or figure.

I have removed this sentence from the revised manuscript.

p.15, l.7: "For BCvdp1, this is linked to an insufficient adjustment of third-and higher-order moments..." Not sure what you mean. That you should use another parametric method that takes into account higher moments? At what point do you start to "overfit" if you do this?

I mean that my parametric methods explicitly adjust mean values and variances. Higher-order moments are only implicitly (and therefore most likely not perfectly) adjusted through the distribution fitting. In fact, with my methods you cannot overfit in your sense because both the normal and the beta (provided its bounds have been fixed) distribution only have two parameters, which are fixed once two moments have been fixed. Therefore, they cannot adjust more than two moments explicitly. An alternative would be to use non-parametric QM methods. I think that all of this is now clearer thanks to the appendix about QM and downscaling that has been appended to the revised manuscript.

p.15, l.11: "... correct the upper tail of the rlds and rsds distributions." Can you say this if you use Kolmogorov-Smirnov, which focuses on the center of the distribution?

Kuiper's test results confirm this. I have adjusted the statement accordingly.

Section 4.2: Comparison with BSRN data. Here you compare point data with area mean data. This comes with potentially quite some uncertainty. See e.g. papers by M.Z. Hakuba et al. 2013 / 2014 / 2016 or N.A.J. Schutgens et al. 2016. Part of your disagreement could have its roots there. More generally, you are looking here more into how good your SRB data is than how good your bias adjustment is. If this is of interest, you should also consider other data, e.g. CERES or global mean estimates for rlds and rsds, e.g. by Trenberth et al. In its current form, the comparison with BSRN data is rather confusing than helping, I think.

I agree (see above). I have removed this section from the manuscript.

Figures 7 and 8: What is the colored rectangle to the lower left in each panel?

These figures have been removed from the revised manuscript.

p.18, l.1: "... and differences between standard deviation biases generated by BCdsdp0, BCsdp1 and BCsdp2 are in line with cross-validation results." What do you mean?

Irrelevant now that this part has been removed from the revised manuscript.

p.18, l.5: "... which again suggests that biases relative to BSRN after bias correction using SRB data depend more on the corresponding SRB data biases than on the method used for the bias correction." So the BSRN comparison does not make sense?

Irrelevant now that this part has been removed from the revised manuscript.

p.18, l.8: I do not understand this paragraph.

Irrelevant now that this part has been removed from the revised manuscript.

p.19, l.1 to 14: I think much of what you are describing here has to do with the fact that you are comparing point measurements with area means. See the above mentioned papers by Hakuba, Schutgens, and references therein.

Maybe. Has been removed from the revised manuscript.

p.19, l.26: Why use a staggered grid?

Smaller differences between RMSDs of adjusted E2OBS data from SRB-grid cell and staggered SRB-grid cell mean values are considered to indicate a better bridging of the E2OBS-to-SRB spatial scale gap. Ideally, there would be no such difference and it would therefore be impossible to tell from this analysis if the target distributions of the bias correction were defined on the SRB or staggered SRB grid. I have included this explanation in the revised manuscript.

Figure 10: The figure seems to suggest that variability is strongly enhanced (red areas) by the bias adjustment. True?

True.

Appendix C: What is the take home message? Figure C2 seems to suggest that the window length is irrelevant. True?

True.

List of all relevant changes made in the manuscript

- the old Table 1 and the old Figures 1, 7, 8 and 9 have been removed
- all text related to the validation against BSRN data has been removed
- Figures 1 (formerly 2) and 5 (formerly 6) have been simplified
- cross-validation using Kuiper's two-sample test has been added (new Figure 6)
- the results and conclusions sections have been almost entirely rewritten
- a new Appendix A on quantile mapping and statistical downscaling has been added
- the individual reasons for testing the different bias correction methods are better explained
- a more mathematical notation has been introduced to define the bias correction methods in Table 1 (formerly 2)

Bias correction of surface downwelling longwave and shortwave radiation for the EWEMBI dataset

Stefan Lange[1]

[1]Potsdam Institute for Climate Impact Research, Telegraphenberg A 31, 14473 Potsdam, Germany

Correspondence to: Stefan Lange (slange@pik-potsdam.de)

Abstract. Many meteorological forcing datasets include bias-corrected surface downwelling longwave and shortwave radiation (rlds and rsds). Methods used for such bias corrections range from multi-year monthly mean value scaling to quantile mapping at the daily time scale. An additional downscaling is necessary if the data to be corrected have a higher spatial resolution than the observational data used to determine the biases. This was the case when EartH2Observe (E2OBS; Calton et al., 2016) rlds and rsds were bias-corrected using more coarsely resolved Surface Radiation Budget (SRB; Stackhouse Jr. et al., 2011) data for the production of the meteorological forcing dataset EWEMBI (Lange, 2016). This article systematically compares various parametric quantile mapping methods designed specifically for this purpose, including those used for the production of EWEMBI rlds and rsds. The methods vary in the time scale at which they operate, in their way of accounting for physical upper radiation limits, and in their approach to bridging the spatial resolution gap between E2OBS and SRB. It is shown how temporal and spatial variability deflation related to bilinear interpolation and other deterministic downscaling approaches can be overcome by downscaling the target statistics of quantile mapping from the SRB to the E2OBS grid such that the sub-SRB-grid scale spatial variability present in the original E2OBS data is retained. Cross-validations at the daily and monthly time scale reveal that it is worthwhile to take empirical estimates of physical upper limits into account when adjusting either radiation component and that, overall, bias correction at the daily time scale is more effective than bias correction at the monthly time scale if sampling errors are taken into account. ~~A validation against independent ground observations from the Baseline Surface Radiation Network (BSRN; König-Langlo et al., 2013) suggests that the bias correction of E2OBS surface downwelling radiation using SRB data that was done for the production of EWEMBI had a positive and neutral overall effect on rlds and rsds, respectively. Using any of the other methods tested here would have given similar results as the biases relative to BSRN remaining after bias correction are dominated by the corresponding SRB data biases.~~

1 Introduction

High-quality observational datasets of surface downwelling radiation are of interest in many fields of climate science, including energy budget estimation (Kiehl and Trenberth, 1997; Trenberth et al., 2009; Wild et al., 2013) and climate model evaluation

(Garratt, 1994; Ma et al., 2014; Wild et al., 2015). As part of so-called climate or meteorological forcing datasets such as those generated within the Global Soil Wetness Project (GSWP; Zhao and Dirmeyer, 2003), at Princeton University (Sheffield et al., 2006), and within the WATer and global CHange project (WATCH; Weedon et al., 2011), the longwave and shortwave components of surface downwelling radiation (abbreviated as rlds and rsds or just longwave and shortwave radiation in the following) are used to, e.g., correct model biases in climate model output (Hempel et al., 2013; Iizumi et al., 2017; Cannon, 2017) and drive simulations of climate impacts (Müller Schmied et al., 2016; Veldkamp et al., 2017; Chang et al., 2017; Krysanova and Hattermann, 2017; Ito et al., 2017).

These meteorological forcing datasets are global, long-term meteorological reanalysis datasets such as those produced by the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR; Kalnay et al., 1996; Kistler et al., 2001) and the European Centre for Medium-Range Weather Forecasts (ECMWF; Uppala et al., 2005; Dee et al., 2011), refined by bias correction using global, gridded observational data. For the components of surface downwelling radiation, such a bias correction is often necessary ~~as~~ because observations of these variables are not assimilated in the reanalyses, which makes them subject to modelling biases of, e.g., land-atmosphere interactions and cloud processes (Kalnay et al., 1996; Ruane et al., 2015).

Different approaches are adopted in order to carry out these bias corrections. Weedon et al. (2011, 2014) apply indirect corrections at the monthly time scale using near-surface air temperature observations for rlds and observations of atmospheric aerosol loadings and cloudiness for rsds. Sheffield et al. (2006) directly rescale rlds and rsds to match observed multi-year monthly mean values. Ruane et al. (2015) directly adjust distributions of daily mean rsds. The observational dataset commonly used for such direct adjustments of rlds and rsds is the Surface Radiation Budget (SRB) dataset assembled by the National Aeronautics and Space Administration (NASA) and the Global Energy and Water EXchanges project (GEWEX; Stackhouse Jr. et al., 2011).

Another meteorological forcing dataset, the EarthH2Observe, WFDEI and ERA-Interim data Merged and Bias-corrected for ISIMIP (EWEMBI; Lange, 2016), was recently assembled to be used as the reference dataset for bias correction of global climate model output within the Inter-Sectoral Impact Model Intercomparison Project phase 2b ~~(ISIMIP2b; Frieler et al., 2016)~~ (ISIMIP2b; Frieler et al., 2017). The surface downwelling longwave and shortwave radiation data included in EWEMBI are based on daily rlds and rsds from the climate forcing dataset compiled for the EarthH2Observe project (E2OBS; Calton et al., 2016). In order to reduce deviations of E2OBS rlds and rsds statistics from the corresponding SRB estimates in particular over tropical land (Dutra, 2015), for EWEMBI, the former were bias-adjusted to the latter at the daily time scale using two newly developed parametric quantile mapping methods.

These methods are conceptually similar to the Ruane et al. (2015) method, which fits beta distributions to reanalysed and observed daily mean rsds for every calendar month, thereby accounting for upper and lower physical limits of rsds using the multi-year monthly maximum value as the upper and zero as the lower limit of the distribution, and then uses quantile mapping to adjust the distributions. In contrast to Ruane et al. (2015), the methods developed to adjust E2OBS rlds and rsds for EWEMBI applies moving windows to estimate beta distribution parameters for every day of the year. This precludes discontinuities at the turn of the month (Rust et al., 2015; Gennaretti et al., 2015) and promises a better bias correction where the seasonality of

radiation is very pronounced such as for rsds at high latitudes. Also, the new methods estimate the physical upper limits of rlds and rsds differently, acknowledging that these limits are necessarily greater than or equal to the greatest value ~~measured over~~ observed during any fixed period. Lastly, while Ruane et al. (2015) linearly interpolate SRB rsds from its natural horizontal resolution of 1.0° to the 0.5° reanalysis grid prior to bias correction, the new methods aggregate the E2OBS data from their original 0.5° grid to the 1.0° SRB grid, where the bias correction is ~~done~~then carried out, and disaggregates these aggregated and bias-corrected data back to the E2OBS grid. Depending on the disaggregation method, this approach promises to generate bias-corrected data with more realistic temporal as well as spatial variability.

The new methods are comprehensively described and cross-validated in this article~~, and in order to assess the value added by the bias correction, the E2OBS and EWEMBI rlds and rsds are compared to independent ground observations from the Baseline Surface Radiation Network (BSRN; König-Langlo et al., 2013)~~. Moreover, several modifications of the new methods are tested here that differ in how they handle the spatial resolution gap between the E2OBS and SRB grids, and how they account for the physical upper limits of rlds and rsds. Also ~~tested~~ included are bias correction methods that operate at the monthly time scale ~~as it is unclear a priori~~ in order to test if bias correction of daily or monthly mean values yields better ~~validation resultsat either time scale~~overall cross-validation results. The lessons learned from these analyses shall benefit bias corrections of surface downwelling radiation to be carried out in future generations of climate forcing datasets.

2 Data

2.1 E2OBS

The EarthH2Observe (E2OBS; Dutra, 2015; Calton et al., 2016) daily mean rlds and rsds data bias-corrected for EWEMBI cover the whole globe on a regular 0.5° × 0.5° latitude-longitude grid and span the 1979–2014 time period. Over the ocean, E2OBS rlds and rsds are identical to bilinearly interpolated ERA-Interim (ERAI; Dee et al., 2011) rlds and rsds. Over land, they are identical to WATCH Forcing Data methodology applied to ERA-Interim reanalysis data (WFDEI; Weedon et al., 2014) rlds and rsds. WFDEI rlds, in turn, is identical to bilinearly interpolated ERAI rlds, adjusted for elevation differences between the ERAI and Climatic Research Unit (CRU; Harris et al., 2013) grids. WFDEI rsds is identical to bilinearly interpolated ERAI rsds bias-corrected at the monthly time scale using CRU TS3.1/3.21 mean cloud cover and considering effects of interannual changes in atmospheric aerosol optical depths (Weedon et al., 2010, 2011, 2014).

2.2 SRB

The ~~E2OBS data are bias-corrected using the~~ observational data used for the bias correction of E2OBS daily mean rlds and rsds for EWEMBI were the NASA-GEWEX Surface Radiation Budget (SRB; Stackhouse Jr. et al., 2011) primary-algorithm estimates of daily mean rlds and rsds from ~~SRB Release~~ the latest SRB releases available at the time, which were release 3.1 ~~and~~ for rlds and release 3.0 ~~, respectively~~for rsds. These data cover the whole globe on a regular 1.0° × 1.0° latitude-longitude grid and span the 07/1983–12/2007 time period. For bias correction and cross-validation, a 24-year subsample of these data ~~is~~

~~used which~~ was used and is used here that spans the 12/1983–11/2007 time period. Additional data from the adjacent months 11/1983 and 12/2007 are employed for computations of running mean values. The SRB estimates of rlds and rsds are based on satellite-derived cloud parameters and ozone fields, reanalysis meteorology and a few other ancillary datasets. Due to a lack of satellite coverage during most of the 07/1983–06/1998 time period over an area centred at 70°E, SRB data artefacts are present over the Indian Ocean (https://gewex-srb.larc.nasa.gov/common/php/SRB_known_issues.php; cf. ~~Fig. 1).~~

~~Deviations of E2OBS from SRB (**left**) and SRB from SRBQC (**right**) 12/1983–11/2007 mean longwave (**top**) and shortwave (**bottom**) radiation. Root-mean-square deviations (RMSDs) over all ocean and all land grid cells are given at the bottom of each panel.~~

~~Deviations of~~ Figs. 2–4, 7). Every SRB grid cell contains exactly four E2OBS ~~from SRB long-term mean rlds and rsds are shown in Fig. 1, together with corresponding deviations of SRB from SRB Release 3.0 quality-check (SRBQC)products. The SRB and SRBQC products were produced with different algorithms (Stackhouse Jr. et al., 2011). Since the primary-algorithm products are more reliable than the quality-check products (Zhang et al., 2015; Stackhouse Jr. et al., 2011) the former were used for the bias correction of E2OBS rlds and rsds for EWEMBI. Over land, differences in long-term mean radiation between E2OBS and SRB are greater in magnitude than those between SRB and SRBQC. Over the ocean, the differences are of similar magnitude. If deviations of SRB from SRBQC data quantify methodological uncertainty inherent to the SRB data then these findings justify the bias correction of E2OBS rlds andrsds using SRB data over land at least.~~ grid cells.

2.3 ~~BSRN~~

~~Observations made at the following 54 BSRN stations are used in this study. In order to adjust rlds for elevation differences between BSRN stations and E2OBS-grid cells, prior to data comparison, BSRN rlds values are offset by the values listed in the rightmost column, based on the formula proposed by Stackhouse Jr. et al. (2011; see text).~~ ~~station latitude longitude offset ALE 82.451 -62.508 -6.580 ASP -23.798 133.888 -4.256 BAR 71.323 -156.607 0.000 BER 32.267 -64.667 -0.112 BIL 36.605 -97.515 -0.560 BON 40.060 -88.370 -0.280 BOU 40.048 -105.007 -9.772 BRB -15.601 -47.713 -0.504 CAB 51.971 4.927 -0.056 CAM 50.217 -5.317 0.588 CAR 44.083 5.059 -14.840 CLH 36.905 -75.713 0.896 CNR 42.816 -1.601 -3.500 COC -12.193 96.835 0.140 DAA -30.665 23.993 0.476 DAR -12.425 130.891 0.812 DOM -75.100 123.383 0.534 DRA 36.626 -116.018 -3.780 EUR 79.980 -85.930 -5.740 FLO -27.533 -48.517 -9.324 FPE 48.310 -105.100 -1.204 FUA 33.582 130.375 -1.092 GCR 34.255 -89.873 -0.112 GOB -23.561 15.041 -4.788 GVN -70.650 -8.250 -0.097 ILO 8.533 4.566 2.492 ISH 24.337 124.163 -0.504 station latitude longitude offset IZA 28.500 -16.300 57.316 KWA 8.720 167.731 0.252 LAU -45.045 169.689 -7.420 LER 60.140 -1.185 1.232 LIN 52.210 14.122 1.764 MAN -2.058 147.425 -0.952 MNM 24.288 153.983 0.000 NAU -0.521 166.916 -0.084 NYA 78.925 11.950 -3.388 PAL 48.713 2.208 1.932 PAY 46.815 6.944 -6.076 PSU 40.720 -77.930 0.028 PTR -9.069 -40.320 0.504 REG 50.205 -104.713 0.336 SAP 43.060 141.328 -4.200 SBO 30.860 34.779 -1.764 SMS -29.443 -53.823 1.932 SON 47.054 12.958 39.424 SOV 24.910 46.410 -3.640 SPO -89.983 -24.799 0.290 SXF 43.730 -96.620 -0.056 SYO -69.005 39.589 -14.012 TAM 22.790 5.529 -1.120 TAT 36.058 140.126 -0.924 TIK 71.586 128.919 -1.484 TOR 58.254 26.462 -0.028 XIA 39.754 116.962 0.280~~

~~Ground observations of longwave downward and shortwave downward (global) radiation made at 54 stations of the Baseline Surface Radiation Network (BSRN; Table 1; König-Langlo et al., 2013) are used as independent validation data for rlds and rsds, respectively. BSRN measurements began at a few stations in 1992. The latest measurements included here are from 2014. Daily mean values of BSRN measurements, which are taken every minute or every few minutes, depending on the station, are computed in two steps. First, gaps no longer than 467/11 minutes in the original rlds/rsds time series are filled by linearly interpolation between values right before the beginning and after the end of a gap, as suggested by Schild (2016; for statistics of BSRN data gaps see Roesch et al., 2011). Daily mean values are then calculated for days that are fully covered by these gap-filled values. Prior to data comparison, the resulting BSRN data availability masks are applied to the original and bias-corrected E2OBS time series from the respective E2OBS-grid cells. Additionally, BSRN rlds values are adjusted for elevation differences between BSRN stations and E2OBS-grid cells as proposed by Stackhouse Jr. et al. (2011). For elevations z_{BSRN} of BSRN stations and z_{E2OBS} of E2OBS-grid cells, BSRN rlds values are offset by 0.028 $(z_{\mathrm{BSRN}} - z_{\mathrm{E2OBS}})$ (cf. Table 1).~~

3 Methods

For the reader who is is not familiar with the concepts of quantile mapping and/or statistical downscaling, a short introduction including definitions of relevant terms is given in Appendix A. The parametric quantile mapping methods introduced in the following are named according to the scheme $\mathrm{BC}vtpx$, where v, t, p are used to distinguish between methods for longwave and shortwave radiation ($v = \mathrm{l}, \mathrm{s}$) operating at the daily and monthly time scale ($t = \mathrm{d}, \mathrm{m}$) using basic and advanced distribution types or parameter estimation techniques ($p = \mathrm{b}, \mathrm{a}$). Index $x = 0, 1, 2$ is used for variants of these methods that differ in how they handle the spatial resolution gap between the SRB and E2OBS ~~. The~~ grids. For the $\mathrm{BC}vtp0$ methods~~correct E2OBS data directly at~~ , the SRB data are spatially bilinearly interpolated to the E2OBS grid ~~using bilinearly~~ and the E2OBS data are then bias-corrected using these interpolated SRB data; this is to mimic the Ruane et al. (2015) approach. For bias correction with the $\mathrm{BC}vtp1$ methods, E2OBS data are spatially aggregated to the SRB grid, the aggregated data are then bias-corrected and the resulting data disaggregated back to the E2OBS grid~~. For~~ ; this approach was used to produce the EWEMBI radiation data. Lastly, the $\mathrm{BC}vtp2$ methods ~~, the~~ adjust mean values and variances at the E2OBS grid such that mean values and variances of spatial aggregates to the SRB grid match the corresponding SRB estimates while the sub-SRB-grid scale spatial structure of mean values and variances present in the original E2OBS data is ~~imposed upon spatially disaggregated SRB data prior to bias correction at the E2OBS grid. The bias correction of E2OBS rlds and rsds for EWEMBI was done with methods BClda1 and BCsda1, respectively.~~ retained; this is to overcome the variability deflation induced by the other two approaches. Since the $\mathrm{BC}vtp0$ and $\mathrm{BC}vtp2$ methods are based on the $\mathrm{BC}vtp1$ methods, the latter are introduced first. Readers who are merely interested in how the EWEMBI radiation data were produced are informed that methods BClda1 and BCsda1 were used for that purpose.

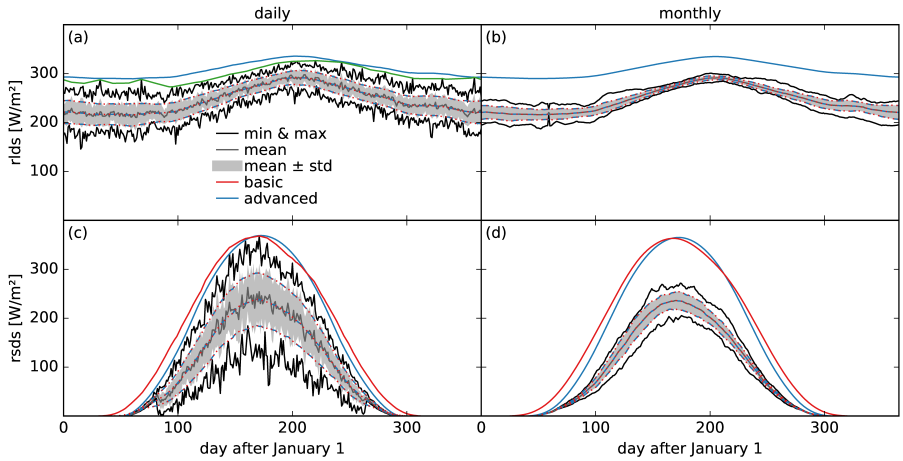

Figure 1. ~~Parameters~~ Estimation of ~~climatological distributions~~ parameters of quantile mapping methods used for the bias correction of longwave (**top**) and shortwave (**bottom**) radiation at the daily (**left**) and monthly (**right**) time scale. This example is based on SRB daily mean rlds and rsds data from 79.5°N, 12.5°E and the 12/1983–11/2007 time period. Climatological distribution parameters are estimated based on empirical 24-year mean values (dark grey), standard deviations (light grey range around mean values) and minimum and maximum values (black) of daily mean (**left**) and 31-day running mean (**right**) radiation computed ~~individually~~ for every day of the year. The distribution parameters estimated for the basic (red) and advanced (blue) bias correction methods (cf. Table 1) include mean values and standard deviations (dotted red, dashed blue), and upper bounds (solid red, solid blue) where beta distributions are used. Note that the basic and advanced estimates of mean values and standard deviations only differ in panel (**c**) near the beginning and end of polar night (cf. Table 1). The ~~light-blue~~ green line in panel (**a**) represents 25-day running mean values of 25-day running maximum values of 24-year maximum values of daily mean rlds, which are used to estimate the upper bounds of the climatological beta distributions used by the BClda1 method (solid blue ~~; see text~~line in panel (**a**)). ~~Please note that the~~ The lower bounds of all climatological beta distributions are set to zero.~~The lowermost and uppermost dotted red and dashed blue lines are the medians of sample minimum and maximum values of random samples of length 24 drawn from the estimated climatological distributions. This plot is based on SRB daily mean rlds and rsds data from 79.5, 12.5and the 12/1983–11/2007 time period.~~

3.1 Bias correction at the ~~SRB grid~~SRB-grid scale

For the BCvtp1 methods, daily mean E2OBS rlds and rsds are first aggregated to the SRB grid using a first-order conservative remapping scheme (Jones, 1999). ~~This~~ The conservative remapping ensures that each aggregated value is the grid-cell area-weighted mean of the underlying four E2OBS values. ~~In the following, the~~ The methods of bias correction of these aggregated values are described ~~.~~ in the following. The method used for the subsequent disaggregation to the E2OBS grid is described in Sect. 3.1.3.

The BCvtp1 methods use parametric transfer functions of the form $F_{vtp}^{\mathrm{SRB}^{-1}}(F_{vtp}^{\mathrm{E2OBS}}(\cdot))$, where F_{vtp}^{E2OBS} and F_{vtp}^{SRB} are climatological cumulative distribution functions (CDFs) of aggregated E2OBS and SRB data, respectively~~, estimated at daily temporal resolution for each~~ . The CDFs are estimated individually for every SRB-grid cell ~~individually~~ and day of the year

Table 1. Distribution types and parameter estimation methods of bias correction methods BC$vtp1$ for day d of the year (cf. Fig. 1). Please note that the lower bounds of all climatological beta distributions are set to zero and that 24-year statistics are replaced by 12-year statistics for cross-validation.

method	distribution type	mean value μ_d	variance σ_d^2	upper bound b_d
BCldb1	normal	~~rm25ym24~~[1] $\langle\langle x_{ij}\rangle_{i24}\rangle_{j25d}$	~~rm25ys24~~[4] $\langle\{x_{ij}\}_{i24}\rangle_{j25d}$	—
BClda1	beta	~~rm25ym24~~[1] $\langle\langle x_{ij}\rangle_{i24}\rangle_{j25d}$	~~rm25ys24~~[4] $\langle\{x_{ij}\}_{i24}\rangle_{j25d}$	~~rm25rx25yx24-rm25ym24~~[7] $A\langle\langle x_{ij}\rangle_{i24}\rangle_{j25d}+B$
BClmb1	normal	~~ym24rm31~~[2] $\langle\langle x_{ij}\rangle_{j31d}\rangle_{i24}$	~~ys24rm31~~[5] $\{\langle x_{ij}\rangle_{j31d}\}_{i24}$	—
BClma1	beta	~~ym24rm31~~[2] $\langle\langle x_{ij}\rangle_{j31d}\rangle_{i24}$	~~ys24rm31~~[5] $\{\langle x_{ij}\rangle_{j31d}\}_{i24}$	~~rm31lda~~[8] $\langle b_j^{\mathrm{lda1}}\rangle_{j31d}$
BCsdb1	beta	~~rm25ym24~~[1] $\langle\langle x_{ij}\rangle_{i24}\rangle_{j25d}$	~~rm25ys24~~[4] $\langle\{x_{ij}\}_{i24}\rangle_{j25d}$	~~rm25rx25yx24~~[9] $\langle[[x_{ij}]_{i24}]_{j25k}\rangle_{k25d}$
BCsda1	beta	~~rm25*ym24~~[3] $\langle\langle x_{ij}\rangle_{i24}\rangle_{j25d^*}$	~~rm25*ys24~~[6] $\langle\{x_{ij}\}_{i24}\rangle_{j25d^*}$	~~yx24-rsdt~~[10] $C\,\mathrm{rsdt}_d$
BCsmb1	beta	~~ym24rm31~~[2] $\langle\langle x_{ij}\rangle_{j31d}\rangle_{i24}$	~~ys24rm31~~[5] $\{\langle x_{ij}\rangle_{j31d}\}_{i24}$	~~rm31sdb~~[11] $\langle b_j^{\mathrm{sdb1}}\rangle_{j31d}$
BCsma1	beta	~~ym24rm31~~[2] $\langle\langle x_{ij}\rangle_{j31d}\rangle_{i24}$	~~ys24rm31~~[5] $\{\langle x_{ij}\rangle_{j31d}\}_{i24}$	~~rm31sda~~[12] $\langle b_j^{\mathrm{sda1}}\rangle_{j31d}$

~~[1] 25-day running mean value of 24-year daily mean values~~

~~[2] 24-year daily mean value of 31-day running mean values, with February 29 value replaced by average of February 28 and March 1 values~~

~~[3] 25-or-fewer-day running mean value of 24-year daily mean values (see text)~~

~~[4] 25-day running mean value of 24-year daily variances~~

~~[5] 24-year daily variance of 31-day running mean values, with February 29 value replaced by average of February 28 and March 1 values~~

~~[6] 25-or-fewer-day running mean value of 24-year daily variances (see text)~~

~~[7] affine transformation of mean value climatology of BClda1 that sits just above the 25-day running mean values of 25-day running maximum values of 24-year maximum values of daily mean rlds (see text)~~

~~[8] 31-day running mean value of upper bounds of BClda1 method~~

~~[9] 25-day running mean value of 25-day running maximum values of 24-year maximum values of daily mean rsds~~

~~[10] rescaled rsdt climatology that sits just above 24-year maximum values of daily mean rsds (see text)~~

~~[11] 31-day running mean value of upper bounds of BCsdb1 method~~

~~[12] 31-day running mean value of upper bounds of BCsda1 method~~ x_{ij} is the daily mean rlds (for BCl$tp1$) or rsds (for BCs$tp1$) on day j of year i.

Brackets $\langle\cdot\rangle$, $\{\cdot\}$, and $[\cdot]$ denote the calculation of sample mean values, variances, and maximum values, respectively.

Bracket subscripts $i24$, $j25d$, and $j25d^*$ indicate that these sample statistics are calculated over years $i\in\{1,\ldots,24\}$, over days $j\in\{d-12,\ldots,d+12\}$, and over days $j\in\{d-n,\ldots,d+n\}$ with $n=\min\{12,\max\{n\geq 0\colon\forall j\in\{d-n,\ldots,d+n\}\colon \mathrm{rsdt}_j\geq 0\}\}$, respectively.

Constants A, B, and C are determined by $\arg\min_{A,B'}\sum_{l=1}^{365}(\langle[[x_{ij}]_{i24}]_{j25k}\rangle_{k25l}-A\langle\langle x_{ij}\rangle_{i24}\rangle_{j25l}+B')^2$, $\min\{B\geq 0\colon\forall l\in\{1,\ldots,365\}\colon A\langle\langle x_{ij}\rangle_{i24}\rangle_{j25l}+B\geq\langle[[x_{ij}]_{i24}]_{j25k}\rangle_{k25l}\}$, and $\min\{C\geq 0\colon\forall j\in\{1,\ldots,365\}\colon C\,\mathrm{rsdt}_j\geq[x_{ij}]_{i24}\}$, respectively.

(Fig. 1). In order to quantify the extent to which bias correction results benefit from explicitly accounting for physical radiation limits, the basic and advanced methods BCl$tb1$ and BCl$ta1$ for longwave radiation use normal and beta distributions, respectively. For shortwave radiation, the relevance of physical limits is less questionable, given that the lower limit of zero matters at least during polar night, and that the solar radiation incident upon land and ocean surfaces is limited by the solar radiation incident upon the top of the atmosphere (cf. Fig. 1). Therefore, all BCs$tp1$ methods use beta distributions and the basic and advanced methods only differ in how they estimate the beta distribution parameters (cf. Fig. 1, Table 1).

3.1.1 Bias correction at the daily time scale

The parameters of the climatological CDFs F_{vdp}^{E2OBS} and F_{vdp}^{SRB} are estimated based on empirical multi-year mean values, variances and maximum values of daily mean radiation from the 12/1983–11/2007 time period. Data from the whole period were used for the production of EWEMBI rlds and rsds. Data from some half of the period (cf. Sect. 4.1) are used for cross-validation in this study.

For shortwave radiation, the basic daily bias correction method is designed to resemble the method outlined by Ruane et al. (2015, Sect. 3.4). BCsdb1 estimates mean values and variances of climatological beta distributions by 25-day running mean values of multi-year daily mean values and variances, respectively, and their upper bounds by 25-day running mean values of 25-day running maximum values of multi-year maximum values of daily mean rsds (solid red line in Fig. 1c). The idea behind this upper bound estimate is that 25-day running maximum values of multi-year maximum values of daily mean rsds resemble the multi-year monthly maximum values of daily mean rsds used by Ruane et al. (2015). Please note that using the same window length for the running maximum calculation and the additional smoothing ensures that the resulting upper bounds are always greater than or equal to the multi-year maximum values of daily mean rsds.

The BCsda1 method employs the climatology of daily mean shortwave insolation at the top of the atmosphere (rsdt; see ~~Sect.~~ Appendix B for how rsdt is calculated in this study) for the upper bound estimation. ~~The rsdt climatology at a given latitude is rescaled such that it sits just above the multi-year maximum values of~~ This is motivated by rsds being limited by rsdt in most locations and seasons, which suggests that the annual cycle of the upper bound of daily mean rsds ~~on all days with rsdt ≥ 50. On a given day of the year, the maximum of this rescaled rsdtvalue and the empirical multi-year maximum daily mean rsds is then used~~ has a similar shape as the climatology of daily mean rsdt. Therefore, method BCsda1 uses a rescaled daily mean rsdt climatology as the upper bound ~~of the beta distribution~~ climatology of daily mean rsds (solid blue line in Fig. 1c). The ~~reason for handling days with rsdt below and above~~ rescaling is done with the smallest possible factor that guarantees that the resulting upper bounds are greater than or equal to the multi-year maximum values of daily mean rsds on all days of the year with rsdt $\geq 50\,\mathrm{Wm^{-2}}$ ~~separately is that~~. An extension of this guarantee to days of the year with lower rsdt would inflate the rescaling factor because during dusk and dawn of polar night, rsds can exceed rsdt due to diffuse radiation coming in from lower latitudes. Therefore, on days of the year with rsdt $< 50\,\mathrm{Wm^{-2}}$, the maximum of the rescaled rsdt and the empirical multi-year maximum daily mean rsds is used as the upper rsds bound. Mean values and variances of the climatological beta distributions of the BCsda1 method are estimated by running mean values of multi-year daily mean values and variances, respectively. The window length used for these running mean calculations is 25 days by default. On days that are fewer than 13 days away from the beginning or end of polar night (as defined by daily mean rsdt going to zero), the window length is shortened to $2n+1$, where n is the number of days between the day in question and the beginning or end of polar night.

For longwave radiation, both the basic and the advanced daily bias correction methods use 25-day running mean values of multi-year daily mean values and variances to estimate climatological mean values and variances, respectively. The upper bounds used by BClda1 are not estimated by the often rather unsmooth 25-day running mean values of 25-day running maximum values of 24-year maximum values of daily mean rlds (~~rm25rx25yx24; solid light-blue~~ solid green line in Fig. 1a)

but by a suitably shifted and rescaled mean value climatology ~~: First, the mean value climatology curve is shifted and rescaled such that it best fits rm25rx25yx24 according to ordinary least squares. This fitted curve is then shifted once more such that the resulting upper bound climatology sits just above rm25rx25yx24~~ (solid blue line in Fig. 1a; formulas in Table 1).

Since the choice of the window length used for all ~~of~~ the running mean and maximum value calculations mentioned above is somewhat arbitrary, the window length dependence of the overall performances of the $\mathrm{BC}v\mathrm{da1}$ methods is investigated in ~~Sect.~~ Appendix D. Sensitivities are found to be very low for window lengths between 10 and 40 days.

3.1.2 Bias correction at the monthly time scale

In order to mimic a bias correction at the monthly time scale as ~~is~~ it was done by, e.g., Sheffield et al. (2006, Sect. 3.d.3), the $\mathrm{BC}v\mathrm{m}p1$ methods bias-correct 31-day running mean values and then rescale each daily value by the corrected-to-uncorrected ratio of the respective 31-day running mean value.

Mean values and variances of the climatological CDFs $F_{v\mathrm{m}p}^{\mathrm{E2OBS}}$ and $F_{v\mathrm{m}p}^{\mathrm{SRB}}$ of 31-day running mean values are simply estimated by 24-year (or 12-year for cross-validation) daily mean values and variances of 31-day running mean values, respectively, with February 29 values replaced by averages of February 28 and March 1 values.

Upper bounds of beta distributions are estimated by 31-day running mean values of the upper bounds of the corresponding ~~daily~~ CDFs $F_{v\mathrm{d}p}^{\mathrm{E2OBS}}$ and $F_{v\mathrm{d}p}^{\mathrm{SRB}}$ of daily mean radiation (cf. Fig. 1, Table 1) ~~as~~ because 31-day running mean values of multi-year maximum values of daily mean radiation are mathematically always greater than or equal to multi-year maximum values of 31-day running mean radiation. The resulting upper bounds are typically much larger than observed 24-year maximum monthly mean radiation (cf. Fig. 1d) because 31 consecutive days of daily mean radiation at the respective physical upper limit are very unlikely to occur in reality.

3.1.3 Disaggregation to the E2OBS grid

In principle, the disaggregation of aggregated and bias-corrected E2OBS data from the SRB to the E2OBS grid can be done in various ways. The simplest approach would arguably be a mere interpolation, which is disadvantageous since it ignores the sub-SRB-grid scale spatial variability present in the original E2OBS data. Probabilistic disaggregation methods, on the other hand, that are designed to retain that variability (cf. Sheffield et al., 2006, Sect. 3.b.1), are impractical if, as in the present case, the purpose of the disaggregation is the ~~construction~~ production and publication of a dataset, because all variants of the dataset that can potentially be generated by a probabilistic algorithm are, as long as all conceivable constraints have been incorporated in the algorithm, equally plausible candidates for the one dataset to be published. Therefore, not a probabilistic but the following deterministic disaggregation approach was used for the ~~construction~~ production of EWEMBI rlds and rsds and is adopted here for all $\mathrm{BC}vtp1$ methods.

First, E2OBS-grid scale upper bounds of daily mean radiation are estimated by bilinearly interpolated maximum values of the climatological upper bounds of SRB all-sky and clear-sky radiation, which ~~, in turn ,~~ in turn are estimated using the $\mathrm{BClda1}$ method for rlds and the $\mathrm{BCsda1}$ methods for rsds (cf. Table 1 and blue lines in Fig. 1a,c). The clear-sky radiation data are included in order to prevent the E2OBS-grid scale upper bounds from being much lower than the real physical limits of

daily mean radiation at that spatial scale, given that due to sub-SRB-grid scale spatial variability, upper radiation bounds at the E2OBS-grid scale may exceed those at the SRB-grid scale.

The original daily E2OBS data are then clamped between zero and these upper bounds, and the resulting values (or their distances to their upper bounds) are rescaled day by day and SRB-grid cell by SRB-grid cell such that their SRB-grid scale aggregates match the ~~respective~~ bias-corrected values. More precisely, if on a given day the SRB-grid scale aggregate of the (clamped) original values from the four E2OBS-grid cells contained in one SRB-grid cell is greater than the bias-corrected value of that day and SRB-grid cell, then the four values are all reduced by a common factor. Otherwise, the distances of the four values to their climatological upper bounds are reduced by a common factor.

3.2 Bias correction at the ~~E2OBS grid~~E2OBS-grid scale

3.2.1 The BC$vtp2$ methods

The disaggregation method introduced above corrects the original E2OBS values from the four E2OBS-grid cells contained in one SRB-grid cell as if they must all be too low ~~/high~~(high) if their area-weighted average is too low ~~/high~~(high). This implicit assumption is questionable since it rules out the possibility that the area-weighted average is too low because one of the four values is much too low while the others are slightly too high, to give just one example. A statistical manifestation of this problem is illustrated and discussed in Sect. 4.2.

The assumption does not need to be made if the bias correction is carried out directly at the E2OBS grid. With target distributions fixed at the SRB grid, ~~one possibility to define~~ target distributions at the E2OBS grid ~~is to require~~ can be defined such that the bias-corrected data ~~to (i)~~ have the SRB-grid scale target distributions and ~~(ii) retain~~ the sub-SRB-grid scale structure of the original E2OBS data. For parametric bias correction methods such as those introduced above, this can be achieved via suitable definitions of the parameters of the E2OBS-grid scale target distributions. Here, for every BC$vtp1$ method, a corresponding BC$vtp2$ method is defined to operate at the same temporal scale and to use the same source (at the E2OBS grid) and target (at the SRB grid) distribution type and parameter estimation technique (cf. Table 1). E2OBS-grid scale target climatologies of mean values, variances and (where necessary) upper bounds are defined as follows.

The mean value estimates of the original E2OBS data are shifted by a common offset per SRB-grid cell and day of the year to obtain the E2OBS-grid scale target mean values. The offsets are chosen such that the E2OBS-grid scale target mean values aggregated to the SRB grid match the corresponding SRB mean value estimates. E2OBS data bias-corrected using these E2OBS-grid scale target mean values have SRB grid-scale aggregates that match the SRB grid-scale target mean values because (i) the aggregation is a linear operation and (ii) the mean value of a linear combination of random variables is equal to the same linear combination of the mean values of these random variables.

~~The~~ To obtain the E2OBS-grid scale target variances, the variance estimates of the original E2OBS data are rescaled by a common (to all four E2OBS grid cells contained in one SRB grid cell) factor f_{ij} per day i of the year and SRB-grid cell j~~to obtain the E2OBS-grid scale target variances~~. For the derivation of the formula for f_{ij} let Y_{ijk} (and X_{ijk}) denote random variables representing bias-corrected (and original) E2OBS data from day i of the year and E2OBS-grid cells $k = 1, 2, 3, 4$

contained in SRB-grid cell j. Then the estimated variance of the SRB-grid scale aggregate of Y_{ijk} can be expanded to

$$\mathrm{Var}\left(\sum_{k=1}^{4} w_{jk} Y_{ijk}\right) = \sum_{k,l=1}^{4} w_{jk} w_{jl} \,\mathrm{Cov}(Y_{ijk}, Y_{ijl}) = \sum_{k,l=1}^{4} w_{jk} w_{jl} \,\mathrm{Cor}(Y_{ijk}, Y_{ijl}) \sqrt{\mathrm{Var}(Y_{ijk})\mathrm{Var}(Y_{ijl})}, \tag{1}$$

where w_{jk} is the area weight of E2OBS-grid cell jk with $\sum_{k=1}^{4} w_{jk} = 1$ for all j, $\mathrm{Cov}(Y_{ijk}, Y_{ijl})$ is the estimated covariance of Y_{ijk} and Y_{ijl}, $\mathrm{Cor}(Y_{ijk}, Y_{ijl})$ is the estimated Pearson correlation ~~between~~ of Y_{ijk} and Y_{ijl}, and $\mathrm{Var}(Y_{ijk})$ is the estimated variance of Y_{ijk}. A bias correction would be deemed successful if the left-hand side of Eq. (1) was equal to the estimated variance of Z_{ij}, the SRB data from day i of the year and grid cell j. On the right-hand side of Eq. (1), $f_{ij}\,\mathrm{Var}(X_{ijk})$ can be substituted for $\mathrm{Var}(Y_{ijk})$ by definition of the scaling factors, and $\mathrm{Cor}(Y_{ijk}, Y_{ijl})$ can be approximated by $\mathrm{Cor}(X_{ijk}, X_{ijl})$ since ~~parametric~~ quantile mapping preserves ~~trends~~ ranks and therefore rank correlations and therefore approximately Pearson correlations. The variance scaling factors f_{ij} for method $\mathrm{BC}vtp2$ are therefore calculated based on

$$\mathrm{Var}Z_{ij} = f_{ij} \sum_{k,l=1}^{4} w_{jk} w_{jl} \,\mathrm{Cor}(X_{ijk}, X_{ijl}) \sqrt{\mathrm{Var}(X_{ijk})\mathrm{Var}(X_{ijl})}, \tag{2}$$

where the variances are estimated using the respective $\mathrm{BC}vtp1$ approach (cf. Table 1), and the Pearson correlations are estimated ~~as~~ by inversely Fisher-transformed 25-day running mean values of Fisher-transformed 24-year daily Pearson correlations of daily (for $\mathrm{BC}vdp2$) or 31-day running mean (for $\mathrm{BC}vmp2$) radiation data. The Fisher transformations are invoked here in order to approximately account for correlation value-dependent sampling error intervals (Fisher, 1915, 1921).

The E2OBS-grid scale target upper bounds are calculated in the same way as the E2OBS-grid scale target mean values. This way, the latter rarely exceed the former. Where they do, the latter are reduced to 99% of the former. ~~This reduction is only necessary for some of the very low rsds values that occur under (near-) polar night conditions~~For longwave (shortwave) radiation, such reductions are necessary in four (11% of all) E2OBS grid cells, and there on an average of 15% (5%) of all days of the year.

~~In~~ Furthermore, in order to obtain realistic E2OBS-grid scale target beta distributions, ~~we further limit~~ the E2OBS-grid scale target variances calculated using Eq. (2) are limited to 40% of $\mu(b-\mu)$, where μ and b are the E2OBS-grid scale target mean values and upper bounds, respectively. This limit is imposed because (i) the variance σ^2 of a random variable taking values from within the interval $[a,b]$ can generally not be greater than $(\mu-a)(b-\mu)$ if μ is the random variable's mean value, (ii) if that random variable is beta-distributed and $\sigma^2 > (\mu-a)(b-\mu)/2$ then the probability density function is U-shaped (Wilks, 1995), which is considered unrealistic for climatological distributions of rlds and rsds, and (iii) $\sigma^2/(\mu(b-\mu))$ has an empirical upper limit of about 40% in the original E2OBS radiation data. The 40% condition is never met for longwave radiation whereas for shortwave radiation it is met in 14% of all E2OBS grid cells, and there on an average of 2% of all days of the year.

3.2.2 The $\mathrm{BC}vtp0$ methods

For the $\mathrm{BC}vtp0$ methods, daily SRB data are first bilinearly interpolated to the E2OBS grid. The E2OBS data are then bias-corrected directly at the E2OBS grid using the interpolated SRB data and transfer functions defined exactly as for the respective $\mathrm{BC}vtp1$ method.

4 Results

~~The~~ In the following, the bias correction methods introduced above are ~~assessed in a threefold way. First, original and bias-corrected E2OBS data are compared to SRB data~~ cross-validated at the SRB-grid scale ~~using a cross-validation approach.Secondly, they are compared to independent ground observations made at 54 BSRN stations. Thirdly,~~ (Sect. 4.1), and their disaggregation performance is assessed by comparing sub-SRB-grid scale spatial variability before and after bias correction ~~are compared in order to measure the disaggregation performance of all methods.~~

~~Data comparisons are done at the daily and monthly time scale in order to identify strengths and weaknesses of bias correction methods operating at either of these time scales. Metrics used to quantify statistical dissimilarity between E2OBS and SRB or BSRN data include differences between multi-year mean values, standard deviations, skewness and maximum values, root-mean-square deviations (RMSDs) between time series, and p-values of two-sample Kolmogorov-Smirnov (KS) test statistics for empirical CDF comparisons (see~~ (Sect. ~~C for details~~4.2).

4.1 Cross-validation at the SRB-grid scale

For the cross-validation against SRB data, 24 years worth of overlapping E2OBS and SRB data are divided into two 12-year samples of which the first one is used to calibrate and the second one to validate the method. ~~Switanek et al. (2017) have shown that if climatological distributions differ substantially between calibration and validation samples of either the observed (here SRB) or modelled (here E2OBS) data (such differences are hereafter denoted as calibration-validation~~ Common practice would be to use data from the first and second half of the 24-year period to define these samples. Yet due to climate change this definition may yield calibration and validation data samples that differ statistically. These differences in turn, which are essentially climate change signals~~or CVCCSs), then the remaining biases after quantile mapping trained on the calibration data sample and applied to the validation datasample are dominated by differences between observed and modelled CVCCSs. This implies that calibration and validation data samples should be made as statistically similar as possible if the~~ , may differ in extent between the E2OBS and SRB data. Switanek et al. (2017) have shown that such differences in climate change signals may then dominate cross-validation ~~is to only measure the~~ metrics and thereby distort the comparative validation of bias correction methods~~' imperfections. Hence~~. In order to minimise this climate change impact on cross-validation results, here, calibration and validation data samples are composed of data from every second and every other year or vice versa, respectively. The samples are accordingly labelled every1st and every2nd.

4.2 ~~Cross-validation against SRB data~~

~~In the following, cross-validation results are only shown and discussed for the BC$vtp0$ and~~ Please note that results for BC$vtp2$ are not shown or discussed in this section because BC$vtp1$ ~~methods since results for the corresponding BC$vtp1$~~ and BC$vtp2$ ~~are virtually identical . In order to (i) measure how the use of spatially interpolated SRB data for bias correction impacts~~ produce virtually identical data at the SRB-grid ~~scale biases, and (ii) assess the value of~~ scale.

4.1.1 BC$vtp0$ versus BC$vtp1$

The first question addressed here is how the bilinear spatial interpolation of SRB data to the E2OBS grid before bias correction with the BC$vtp0$ methods impacts the ~~extra complications involved in the parameter estimations of the advanced compared to the basic bias correction methods~~distribution of bias-corrected rlds and rsds values at the SRB-grid scale. To quantify these impacts, biases in multi-year daily mean values, standard deviations, ~~skewness~~ and maximum values remaining after bias correction with methods BC$v$da0 ~~, BC$v$da1 and BC$v$db1 are compared first.Then, bias correction methods operating at different temporal scales are compared with respect to their ability to adjust the interannual variability of monthly mean values . Lastly, the overall performance of all BC$vtp1$ methods is assessed via CDF comparisons at both the daily~~ and ~~monthly time scale.~~ BC$v$da1 are compared in the left and middle columns of Figs. 2 and 3.

~~Maps of biases in multi-year mean values, standard deviations , skewness~~ Since linear interpolation always yields values that are intermediate to the values at the interpolation knots it is expected that daily SRB data bilinearly interpolated to the E2OBS grid and then aggregated back up to the SRB grid will be more smooth overall both in space and time than the original SRB data. Manifestations of the increased smoothness in time are the more negative biases of standard deviations (Fig. 2) and maximum values ~~of daily mean rlds and rsds~~ (Fig. 3) remaining after bias correction ~~at the daily time scale are depicted in Figs. 2 and 3. Remaining mean value biases for BC$vdp1$ are small with medians~~ with BC$v$da0 than with BC$v$da1. Standard deviations after bias correction with BC$v$da0 in particular are negatively biased by more than 4% (median over calendar months ~~and~~ $\times$ validation data samples~~being within ± 1 at most locations. At low/high latitudes, BCsdb1 leaves smaller/larger mean value biases than BCsda1. In comparison to BC$v$da1, BC$v$da0 leaves greater mean value biases in particular over coastal and mountainous regions, where spatial gradients are large.~~

~~Medians of relative standard deviation biases~~) in most regions. In mountainous and therefore spatially heterogeneous regions, also multi-year monthly mean radiation is changed significantly by the interpolation, with median biases over calendar months $\times$ validation data samples remaining after bias correction with ~~BC$vdp1$ are mostly within $\pm 4\%$. Underestimations by more then 4% remain over large parts of subtropical Northern Hemisphere land. In most locations, BCldb1 leaves smaller rlds standard deviation biases than BClda1. Bias correction with~~ BC$v$da0 ~~yields systematically too low standard deviations in most locations, in particular for shortwave radiation.This is a result of the variance deflation the bilinear interpolation inflicts on the SRB data.~~ exceeding 2 $\mathrm{Wm^{-2}}$ in many such places (Fig. 2).

~~Large skewness biases with medians frequently exceeding $\pm 50\%$ remain~~

4.1.2 BC$vtax$ versus BC$vtbx$

Next is an assessment of how the treatment of the upper bound of the distributions estimated by the BC$vdp1$ methods impacts the distribution of bias-corrected rlds and rsds values at the SRB-grid scale. To quantify these impacts, biases in multi-year daily mean values, standard deviations, and maximum values remaining after bias correction with ~~any method. The median skewness of longwave radiation is mostly too low, in particular over the ocean and no matter if CDFs of beta or normal distributions are used in the transfer function. The median skewness of shortwave radiation is too low over most of the tropics~~

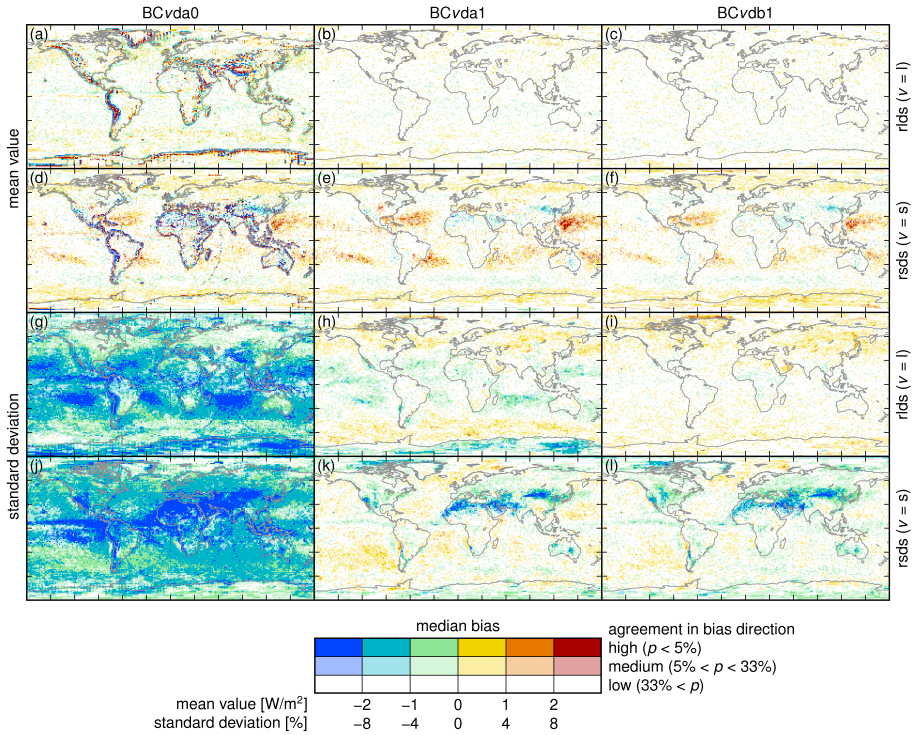

Figure 2. Biases relative to SRB in mean values (**a–f**) and standard deviations (**g–l**) of spatially aggregated (to the SRB grid) daily mean longwave (**a–c**, **g–i**) and shortwave (**d–f**, **j–l**) radiation after bias correction with methods BC$v$da0 (**left**), BC$v$da1 (**middle**) and BC$v$db1 (**right**). The biases are calculated individually for each calendar month (January to December) and calibration data sample (every1st, every2nd) pooling SRB and corrected E2OBS data from all years of the corresponding validation data sample (every2nd, every1st, respectively) and omitting shortwave radiation data from months with monthly mean rsdt less than 1 $\mathrm{Wm^{-2}}$ (cf. ~~Sect.~~ Appendix B and Fig. D1c). Depicted are median and agreement in direction (sign of bias) of these individual biases, represented by hue and saturation of a grid cell's colour, respectively. Categories of agreement in bias direction are defined based on one-sided p-values obtained from modelling underestimations and overestimations for individual calendar months and validation data samples as ~~possible~~ outcomes of independent fifty-fifty Bernoulli trials. More saturated colours indicate higher statistical significance of biases remaining after bias correction.

~~and high-latitude oceans and too high over most land masses and subtropical oceans. The biases of third- and higher-order moments of the distribution of daily mean radiation would arguably be better adjusted by non-parametric quantile mapping methods.~~ methods BC$v$da1 and BC$v$db1 are compared in the middle and right columns of Figs. 2 and 3.

~~Medians of remaining biases in~~ For longwave radiation, the basic method BCldb1 assumes normally distributed values and therefore does not account for any upper physical limit of rlds whereas the advanced method BClda1 assumes the existence of such a limit and estimates it empirically. Figure 3 shows that the advanced method generally yields a better correction of 12-year maximum values~~are mostly within ±10. Compared to BC$v$da1, these biases are shifted to more negative values for BC$v$da0.This is related to the negative standard deviation biases that remain after bias correction with BC$v$da0. For rlds, the~~

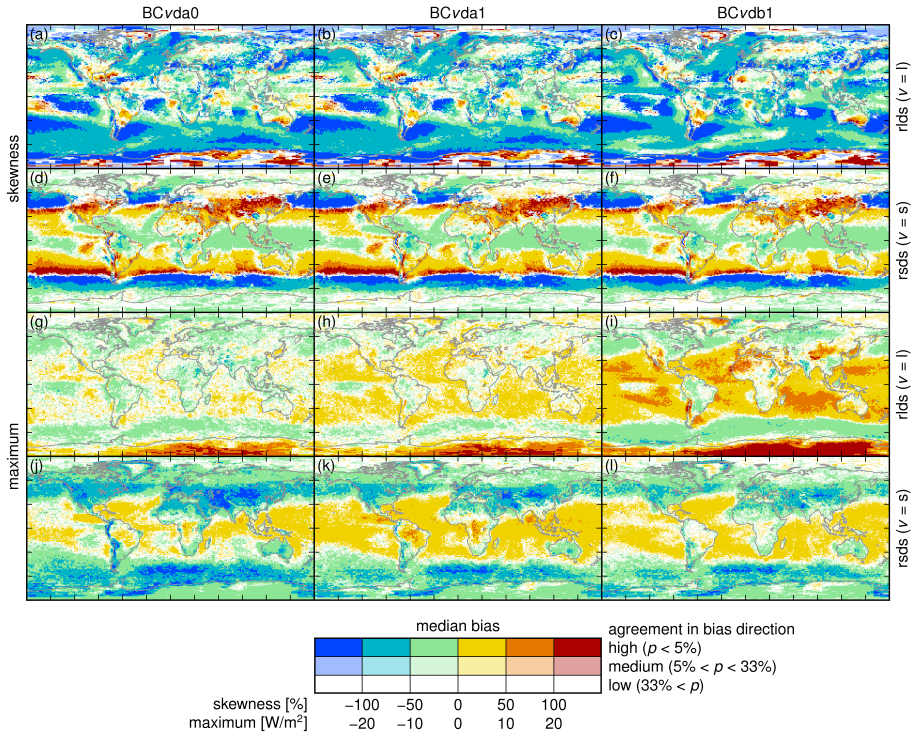

Figure 3. Same as Fig. 2 but for biases in skewness (**a–f**) and 12-year maximum values (**g–l**).

~~use of beta instead of normal distributionsclearly reduces the remaining maximum value biases. For rsds, the basic estimates of upper radiation bounds yield a slightly greater reduction of maximum value biases than the advanced ones.~~. In contrast, standard deviations are slightly better corrected by the basic method and mean values are equally well corrected by both methods (Fig. 2).

For shortwave radiation, both the basic and the advanced method empirically estimate upper physical limits of rsds and take these into account in the form of upper bounds of beta distributions. The limit estimates are based on downwelling shortwave radiation at the surface and at the top of the atmosphere for BCsda1, and on rsds only for BCsdb1. Figure 3 shows that the basic method generally yields a better correction of 12-year maximum values. Also standard deviations and mean values are slightly better corrected by BCsda1 than by BCsdb1 (Fig. 2).

4.1.3 BCvdpx versus BCvmpx

~~Since multi-year mean values of monthly mean values are identical to multi-year mean values of the underlying daily values, bias correction at the~~ Next is a comparative cross-validation of methods BCvdpx and BCvmpx operating at the daily and monthly time scale, respectively. The cross-validation itself is also done at the daily and monthly time scale ~~adjusts multi-year mean values of daily rlds and rsds similarly well as~~ based on statistics of daily and monthly mean radiation, respectively. A

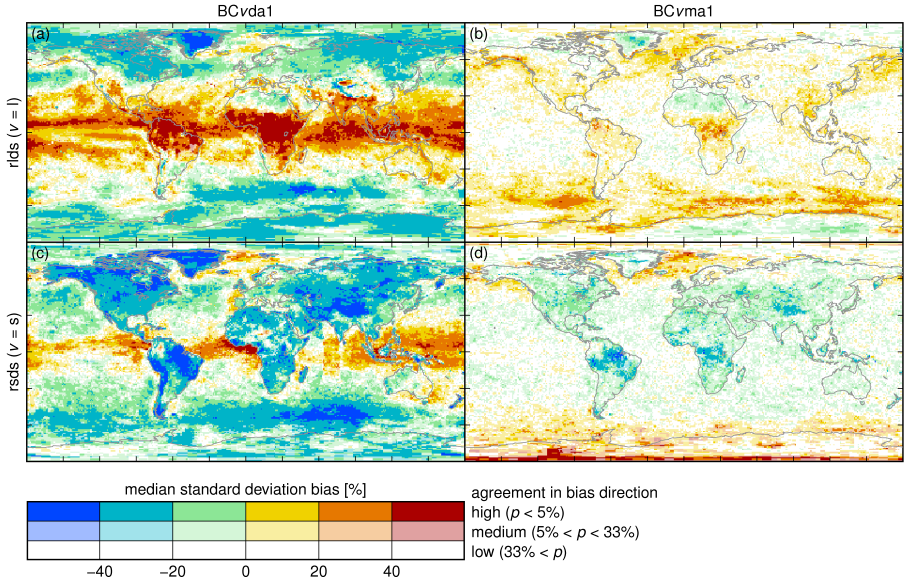

Figure 4. Same as Fig. 2 but for relative biases in interannual standard deviations of monthly mean radiation remaining after bias correction with methods BC$v$da1 (**left**) and BC$v$ma1 (**right**).

joint assessment of these cross-validations shall reveal whether bias correction at the daily ~~time scale (cf. Fig. 2a–f). However, the~~ or monthly time scale is better overall.

By design, the BCvdpx and BCvmpx methods ~~leave larger and spatially less homogeneous biases of~~ are equally good at correcting multi-year ~~standard deviations and maximum~~ mean values of daily mean radiation~~than the~~. However, both day-to-day and year-to-year variability are expected to be differently well corrected by the methods operating at different time scales. Since day-to-day variability is (not) explicitly adjusted by the methods operating at the daily (monthly) time scale the BCvdpx ~~methods, with medians over calendar months and validation data samples being mostly within ±20% and ±20 , respectively. In general, bias correction at the monthly~~ methods are expected to perform better at the daily time scale ~~is expected to leave smaller biases at the monthly time scale~~ than ~~bias correction~~ the BCvmpx methods. The year-to-year variability, on the other hand, is explicitly corrected by the BCvmpx methods and it is not by the BCvdpx methods because daily data from different years are pooled before quantile mapping is carried out at the daily time scale. ~~This is exemplified in Fig. 4, where median biases of~~ Consequently, biases in interannual standard deviations of monthly mean ~~rlds and rsds are shown to be mostly within/beyond ±20%~~ radiation are much larger after bias correction with ~~BC$v$ma1/~~BC$v$da1 ~~.~~ than with BC$v$ma1 (Fig. 4), and the BCvmpx methods are generally expected to perform better at the monthly time scale than the BCvdpx methods.

In order to assess whether bias correction at the daily or monthly time scale is more effective overall, a performance measure is needed that is comparable across time scales. Common performance measures of distribution adjustments at individual time scales are the two-sample Kolmogorov-Smirnov ~~test statistics of the respective E2OBS and SRB data before (black) and after (colours) bias correction (cf. Sect. C; greater~~ (KS) and Kuiper's two-sample test statistic. While Kuiper's test is equally

sensitive to CDF differences at all quantiles, the KS test is more sensitive at the median than in the tails. A straightforward comparison of these test statistics across time scales is not very meaningful because sample sizes at the daily and monthly time scale differ by a factor of thirty, which implies that the same value of a test statistic has different statistical significance at the daily and monthly time scale. A better comparability can be achieved by comparing the test statistic's p~~-values indicate stronger agreement of~~ -value, which represents the statistical significance of CDF differences. In the present cross-validation, the CDFs compared are based on bias-corrected E2OBS and ~~SRB distributions). The~~ the corresponding SRB data, and a higher p-value indicates more similar CDFs and therefore a better bias correction. For details of the calculation of p-values ~~are determined individually for each grid cell, calendar month and calibration data sample, with all corresponding values pooled into one distribution and omitting shortwave radiation data from months with average rsdt less than 1 . The horizontal lines~~ of ~~each box-whisker plot represent~~ the ~~90th, 75th, 50th, 25th and 10th (from top to bottom) grid-cell area-weighted percentile of the natural logarithms of these p-values over calibration data sample (1sthalf, 2ndhalf), latitude and longitude~~two-sample KS and Kuiper's two-sample test statistic see Appendix C. ~~The grey horizontal line marks the $p = 10\%$ significance level. Compared to $\mathrm{BC}vtp1$, p-values produced by $\mathrm{BC}vtp0$ are slightly lower but qualitatively similar.~~

~~The overallperformance of the $\mathrm{BC}vtp1$ methods is examined next. As delineated above, it is measured by similarities of empirical CDFs of (spatially aggregated) E2OBS and SRB data before and after bias correction, quantified by~~

Global distributions of p-values of two-sample ~~KS test statistics (cf. Sect. C; greater p-values indicate stronger agreement of E2OBS and SRB distributions). For all radiation types, validation time scales, calendar months and $\mathrm{BC}vtp1$ methods, distributions of these p-values over calibration data sample, latitude and longitude are depicted as box-whisker plots~~ test statistics for seasonal distributions of daily and monthly mean rlds and rsds are shown in Fig. 5 ~~.~~

~~In all panels of Fig. 5, the~~ for the KS test and Fig. 6 for Kuiper's test. In accordance with expectations, both tests indicate that CDFs are generally better adjusted by $\mathrm{BC}vdpx$ than by $\mathrm{BC}vmpx$ at the daily time scale and vice versa at the monthly time scale. Yet performance differences between $\mathrm{BC}vdpx$ and $\mathrm{BC}vmpx$ are clearly more significant at the daily than at the monthly time scale. This suggests that bias-correcting at the daily instead of at the monthly time scale yields bias decrements at the daily time scale that exceed bias increments at the monthly time scale. Therefore, bias correction at the daily time scale is deemed more effective overall then bias correction at the monthly time scale.

To elaborate this further, the $p = 10\%$ significance level is marked ~~with~~ by a grey horizontal line ~~.~~ in all panels of Figs. 5 and 6 and is to be compared with the 10th percentiles of the global distributions of p-values of the two-sample test statistics. Any coincidence of ~~a p-value distributions'~~ such a 10th percentile with the 10% significance level suggests that ~~that~~ the corresponding p-value distribution is in agreement with the null hypothesis of the ~~KS test, which~~ respective test. Since the null hypothesis of both tests is that the ~~compared samples were drawn~~ samples compared are from the same underlying distribution, ~~and this indicates that the respective bias correction method did its job~~such a coincidence suggests that the bias correction which produced one of the samples compared worked perfectly within the limits of sampling uncertainty. Similarly, 10th percentiles of p-value distributions ~~below/above~~ above (below) the 10% significance level ~~indicate undercorrections/overcorrections .~~

~~As expected by design, the $\mathrm{BC}vdp1$ methods outperform the $\mathrm{BC}vmp1$ methods at the daily~~ suggest overcorrections (undercorrections) in terms of sampling uncertainty. In that sense, the $\mathrm{BC}vtpx$ methods are generally overcorrecting at the monthly time scale

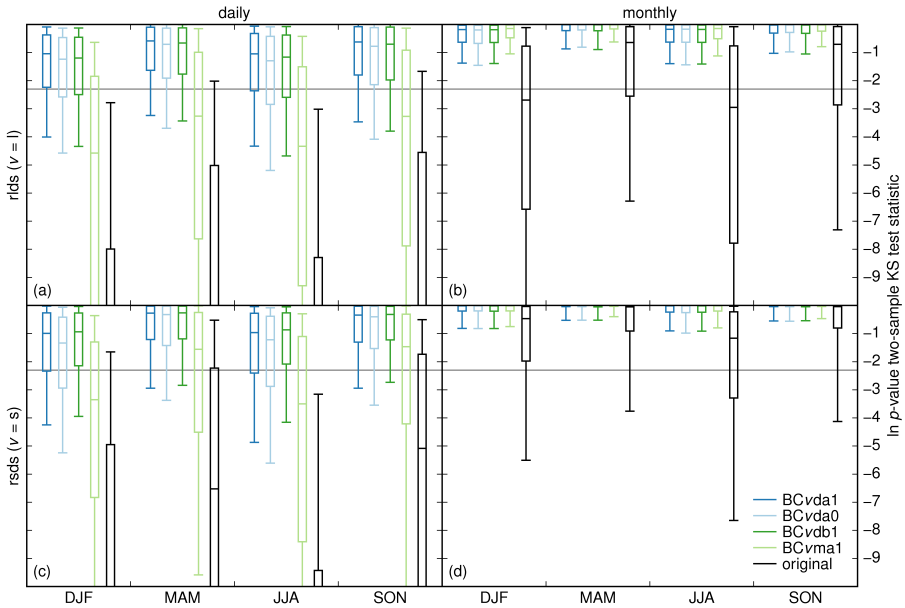

Figure 5. Overall performance of bias correction methods ~~BC$vtp1$~~ BC$v$da1, BC$v$da0, BC$v$db1, and BC$v$ma1 for longwave (**top**) and shortwave (**bottom**) radiation at the daily (**left**) and monthly (**right**) time scale as quantified by p-values of two-sample Kolmogorov-Smirnov test statistics of the respective E2OBS and SRB data before (black) and after (colours) bias correction (cf. ~~Sect.~~ Appendix C; greater p-values indicate stronger agreement of E2OBS and SRB distributions). The p-values are determined individually for each grid cell, ~~calendar month~~ season, and calibration data sample, with all corresponding values pooled into one distribution and omitting shortwave radiation data from months with average rsdt less than $1\,\mathrm{Wm}^{-2}$. The horizontal lines of each box-whisker plot represent the 90th, 75th, 50th, 25th, and 10th (from top to bottom) grid-cell area-weighted percentile of the natural logarithms of these p-values over calibration data sample (1sthalf, 2ndhalf), latitude and longitude. The grey horizontal line marks the $p = 10\%$ significance level. ~~Compared to BC$vtp1$, p-values produced by BC$vtp0$ are slightly lower but qualitatively similar.~~

and ~~the latter outperform the former at the monthly time scale. Yet performance gaps are much larger at the daily than at the monthly~~ undercorrecting at the daily time scale. ~~The small performance gaps at the monthly time scale demonstrate that even though a direct bias correction of monthly mean values adjusts their distribution more precisely than a correction of the distribution of the underlying daily values (cf.Fig. 4), the statistical significance of this adjustment is low for sample sizes as small as in this cross-validation study. Rather, The p-value distributions depicted in Fig. 5b,d suggest that if sampling errors are taken into account then the BC$vdp1$ methods correct the distributions of monthly mean values almost as well as the BC$vmp1$ methods . In fact, both types of methods tend to overcorrect them.~~

~~In contrast, distributions of daily E2OBS data are undercorrected across the board. For BC$vdp1$, this is linked to an insufficient adjustment of third- and higher-order moments of the distributions of daily mean radiation(cf. Fig. 3). Throughout the year, BCldb1 performs slightly worse than BClda1 while BCsdb1 performs slightly better than BCsda1. The findings around Figs. 2 and 3 suggest that these differences in overall performance can be explained by how well the respective methods~~

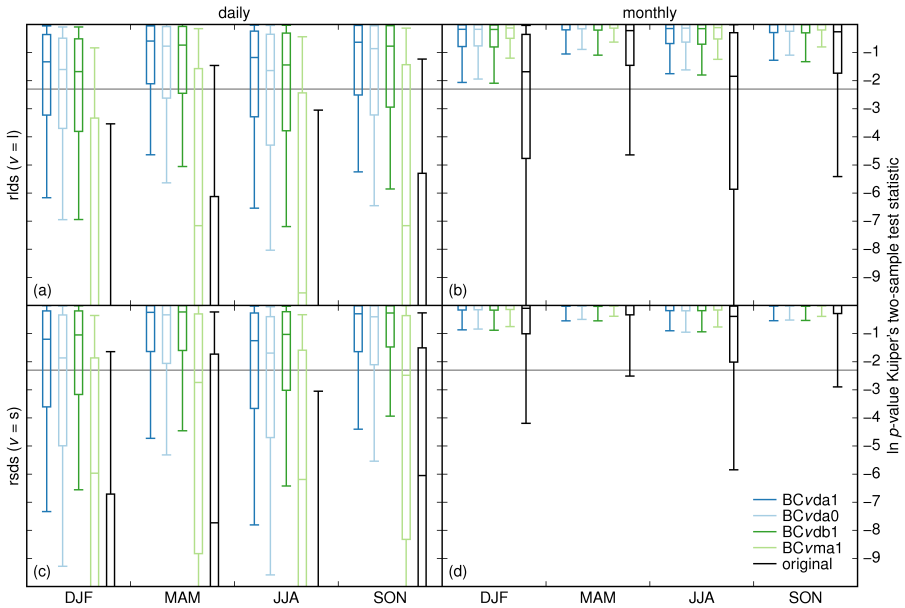

Figure 6. Same as Fig. 5 but based on p-values of Kuiper's two-sample test statistic.

~~correct the upper tail of the rlds and rsds distributions. Finally, it is worth noting that rlds biases do not exhibit any pronounced seasonality whereas rsds biases are particularly large in the solstice months of June and December,~~

The KS and Kuiper's test statistics also confirm the finding of Sect. 4.1.2 that at the daily time scale, the BC$v$da1 methods outperform the BC$v$db1 methods for longwave radiation and vice versa for shortwave radiation. This holds true for all seasons and irrespective of CDF differences being generally greater in summer and winter (DJF and JJA) than in the transition seasons (MAM and SON) both before and after bias correction~~with any method.~~ . Moreover, the test statistics find both BC$v$da1 and BC$v$db1 to outperform BC$v$da0 at the daily time scale, which is in line with the finding of Sect. 4.1.1 that the BC$v$da0 methods deflate day-to-day variability.

The fact that all BCvdp1 methods are undercorrecting at the daily time scale demonstrates the imperfections of these parametric quantile mapping methods. The remaining CDF differences must be linked to imperfect bias corrections of moments of higher than second order since multi-year mean values and standard deviations are well adjusted by design. To illustrate this, relative skewness biases remaining after bias correction with BCvdp1 are shown to exceed 50% (median over calendar months $\times$ validation data samples) in many regions (Fig. 3). Another manifestation of the imperfections are remaining biases in the tails of the distribution of daily mean rlds and rsds. These must be larger than the remaining median biases because p-values of Kuiper's test statistics for these distributions are generally larger than those of the corresponding KS test statistics.

4.2 ~~Validation against BSRN data~~

~~Rankings of original (black) and bias-corrected (colours) E2OBS data according to their similarity in distribution to the corresponding BSRN data from 54 stations (Table 1) and the 1992–2014 time period for daily (**left**) and monthly (**right**) mean longwave (**top**) and shortwave (**bottom**) radiation. Similarity in distribution is quantified by p-values of two-sample Kolmogorov-Smirnov test statistics (cf. Sect. C). Higher ranks indicate greater p-values and thus greater similarity in distribution. The map shows the highest ranking E2OBS dataset per station and calendar month (see legend). Rankings for rsds are not computed for months with average rsdt less than 1 . Rank distributions over stations and months are shown in the inset at the lower left of each panel. The percentages at the upper left of each panel display in how many cases (stations, calendar months) E2OBS data bias-corrected with a certain method (colour) outrank the corresponding original E2OBS data. Very similar results are obtained for the corresponding basic bias correction methods, with outranking percentages deviating by $\pm 1\%$ at most.~~

~~Same as Fig. 7 but for biases in multi-year daily standard deviations (**left**) and multi-year monthly mean values (**right**), with higher ranks corresponding to lower absolute values of these biases.~~

~~Rankings of original and bias-corrected E2OBS data according to their similarity in distribution to the corresponding BSRN data for daily and monthly mean longwave and shortwave radiation are depicted in Fig. 7. The distribution of rankings over BSRN stations and calendar months suggest that in most cases, bias correction of E2OBS using SRB data is beneficial either with any method or not at all, depending on whether the SRB or the original E2OBS data are less biased relative to the BSRN ground truth.~~

~~More often than with BC$ltp1$ or BC$ltp2$, the bias correction with BC$ltp0$ reduces rlds biases relative to BSRN. Smaller and opposite differences are found between the BCstpx methods. At the monthly time scale, the differences in rlds distribution similarity are mainly determined by long-term mean value biases (cf. Figs. 7b and 8b). This suggests that more often than not, the bilinear interpolation included in BC$vtp0$ yields more realistic long-term mean rlds values at the E2OBS grid than the disaggregation methods of BC$vtp1$ and BC$vtp2$. Presumably, this is due to an elevation correction implicitly carried out along with the interpolation.~~

~~At the daily time scale , standard deviation biases explain most of the method dependencies of distribution similarities between BSRN and bias-corrected E2OBS data (cf. Figs. 7a,c and 8a,c). For rlds, compared to BClmpx, BCldpx leaves lower/higher standard deviation biases and yields higher/lower distribution similarities over the tropics/extratropics. For rsds, BCsdpx leaves lower standard deviation biases and yields higher distribution similarities everywhere, and differences between standard deviation biases generated by BCsd$p0$, BCsd$p1$ and BCsd$p2$ are in line with cross-validation results (Fig. 2).~~

~~Changes in RMSDs of E2OBS from BSRN daily (**left**) and monthly (**right**) mean longwave (**top**) and shortwave (**bottom**) radiation at 54 stations (Table 1) after bias correction with methods BC$vtax$. At each station, RMSDs are computed over the whole time series of available BSRN data, only omitting shortwave radiation data from months with average rsdt less than 1 . Listed in grey are the stations where bias correction on average decreases (left list; the more the lower) or increases (right list; the more the higher) RMSDs of monthly mean radiation by more than 10 . Annotated in black are stations where RMSD changes (**b**) for BC$lta0$ differ by more than 5 from those obtained with any other method, (**c**) spread a range larger than 10 over all bias correction methods. Numbers given in each panel are station mean RMSD changes per bias correction method (colour;~~

~~) over all high-latitude (beyond 66or 66; first row), all mid-latitude (66–33and 33–66; second row), all low-latitude (33–33; third row) and all (fourth row) stations. These numbers change by at most ±0.2 for the respective basic bias correction method.~~

~~Bias correction-induced changes in RMSDs of E2OBS from BSRN time series of daily and monthly mean longwave and shortwave radiation are shown in Fig. 9. At both time scales and for both radiation types, the between-station variance of the RMSD changes is larger than their within-station variance, which again suggests that biases relative to BSRN after bias correction using SRB data depend more on the corresponding SRB data biases than on the method used for the bias correction.~~

~~However, there are two notable exceptions from this rule. First, more often than not, rlds RMSDs are systematically lower after bias correction with BC$ltp0$ than with BC$ltp1$ or BC$ltp2$. This is particularly well visible at the monthly time scale. As conjectured above, this might be the result of an elevation correction of long-term rlds mean values implicitly done by BC$ltp0$ along with its bilinear interpolation of SRB data to the E2OBS grid.~~

~~Secondly, at eight stations (BER, COC, ISH, IZA, KWA, MAN, MNM, NAU; cf. Table 1), daily rsds RMSDs after bias correction with different methods spread over a range wider than 10 (in six cases even 20). These stations are all located on islands that are smaller than one SRB-grid cell but large enough to be resolved by the original E2OBS data, which is to say that the E2OBS climatologies at the corresponding 0.5grid cells stand out against those at the neighbouring 0.5grid cells. Bias correction results at these stations can therefore be expected to depend on how a given method modifies the sub-SRB-grid scale spatial variability of the original E2OBS data. It turns out that, at all of these stations except IZA, daily rsds RMSDs are smaller after bias correction with BC$stp1$ than with BC$stp0$ and BC$stp2$~~

~~On average over the respective stations, rlds RMSDs are reduced by all bias correction methods at all latitudes. In contrast, bias correction results are more heterogeneous for rsds. At low latitudes (33–33), bias correction has a neutral average effect, at middle latitudes (66–33and 33–66) it reduces average rsds RMSDs, and at high latitudes (beyond 66or 66) it strongly increases them. Five out of the six stations with the greatest rsds RMSD increases are high-latitude stations (ALE, BAR, DOM, GVN, SYO; cf. Table 1). believe that the difficult cloud characterisation over surfaces that are frequently covered by snow, ice or water is the primary reason for large SRB shortwave radiation biases at such polar sites.~~

4.2 ~~Disaggregation~~ Spatial disaggregation and sub-SRB-grid scale spatial variability

As outlined in Sect. 3.2.1, the ~~BC$vtp1$ disaggregation method~~ BC$vtp1$ approach to the disaggregation of bias-corrected daily mean rlds and rsds values from the SRB- to the E2OBS-grid scale is based on the implicit assumption that the original E2OBS values of daily mean radiation onto the four E2OBS-grid cells contained in one SRB-grid cell must all be too low ~~/high~~(high) if their area-weighted average is too low ~~/high~~(high). The four original values are then all increased ~~/decreasedby the disaggregation~~ (decreased) by the BC$vtp1$ method. In order to account for their upper ~~/lower~~(lower) physical bounds, the increases ~~/decreases~~(decreases) are done by a common scaling factor applied to the distances to ~~the respective upper/lower~~ these bounds. This leads to a reduction of the differences between the four values (necessarily if the four bounds are equal, in most cases if they are similar), i.e., to a deflation of sub-SRB-grid scale spatial variability.

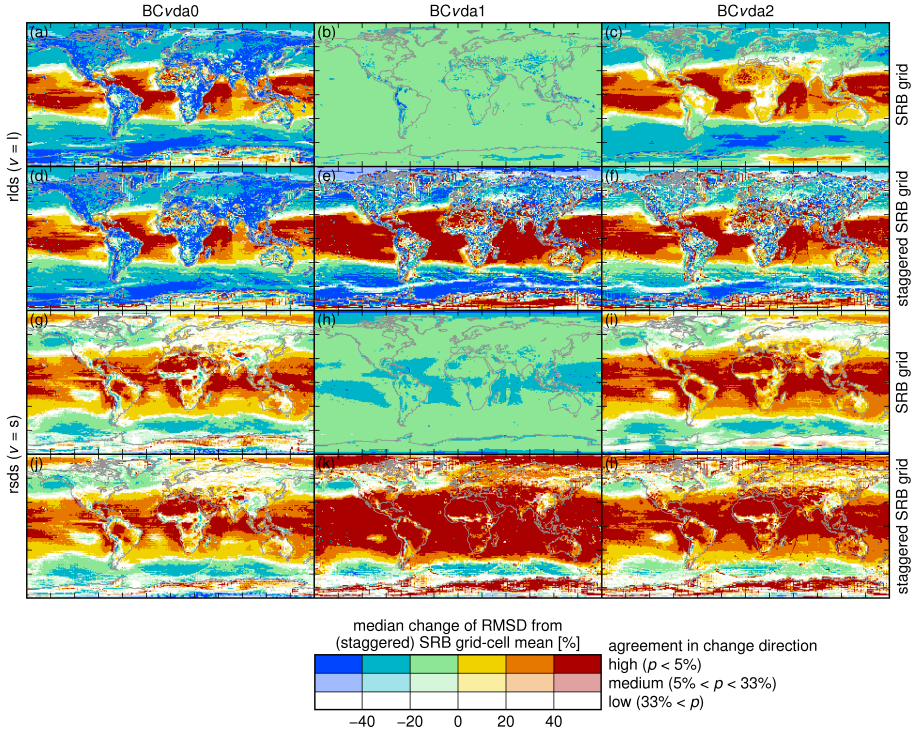

Figure 7. Relative change ~~after~~ by bias correction with methods BC$v$da0 (**left**), BC$v$da1 (**middle**), and BC$v$da2 (**right**) of the RMSD of daily ~~0.5~~mean E2OBS-grid scale longwave (**a–f**) and shortwave (**g–l**) radiation from the ~~respective 1grid-cell mean~~ aggregated SRB-grid scale values based on 1° grid cells of the SRB grid (**a–c**, **g–i**) and the staggered SRB grid (**d–f**, **j–l**; see text). For every 1° grid cell and calendar month, the RMSDs are calculated using original or bias-corrected E2OBS data from the four ~~associated~~ 0.5° grid cells ~~and~~ contained in the 1° grid cell, pooling data from the entire 12/1983–11/2007 time period ~~,~~ and omitting shortwave radiation data from months with average rsdt less than 1 $\mathrm{Wm^{-2}}$~~(cf. Sect. B and Fig. D1c).~~ Depicted are median and agreement in direction of ~~individual~~ monthly RMSD changes ~~after~~ by bias correction (~~cf.~~ same colouring scheme as in Fig. 2). Very similar results are obtained for the corresponding basic bias correction methods.

In order to illustrate and ~~measure~~ quantify the extent of this ~~deflation,~~ variability deflation and compare the BCvtp0, BCvtp1, and BCvtp2 methods in terms of their impact on sub-SRB-grid scale spatial variability, the RMSD of the four ~~E2OBS~~ E2OBS-grid scale values of daily mean radiation per SRB-grid cell from their area-weighted average is calculated ~~for every time step~~ over all days of a given calendar month both before and after bias correction with ~~methods BC$v$da1. The median~~ either method. Median relative bias correction-induced ~~deflation~~ changes of these RMSDs ~~over all calendar monthsis found to exceed~~ are depicted in Fig. 7 and demonstrate that BC$v$da1 indeed generally deflates them, in some regions by more than 20~~% in some locations~~ % (median over calendar months) for both longwave and shortwave radiation~~(Fig. 7b, h)~~. In contrast, BCvtp0 and BCvtp2 deflate or inflate them depending on variable and region.

In an analogous manner, such RMSDs can be computed based on data from the four E2OBS-grid cells contained in ~~each~~ one staggered SRB-grid cell, where the staggered SRB grid is a regular $1.0° \times 1.0°$ latitude-longitude grid shifted by 0.5° latitude and 0.5° longitude relative to the SRB grid, i.e., every staggered SRB-grid cell contains E2OBS-grid cells contained in four different SRB-grid cells. ~~Bias correction with methods BC$v$da1 has a strong impact on these RMSDs (~~Median relative bias correction-induced changes of these RMSDs are also depicted in Fig. 7~~e, k), with increases/decreases found mostly over tropical and polar regions/middle latitudes. Most importantly, the RMSD change patterns at the~~ . Ideally, bias correction-induced changes of RMSDs from SRB and staggered SRB ~~grid are very different from those at the~~ grid-cell mean values would be equal. It would then be impossible to tell from their comparison whether the bias correction's target distributions were defined on the SRB or on the staggered SRB grid. ~~This is considered to be an artefact caused by the BC$vtp1$ disaggregation method.~~

~~E2OBS data bias-corrected with BC$v$da2 do not suffer from the deflation of sub-SRB-grid scale spatial variability that results from~~ The BCvd$p1$ methods do not fulfil this criterion as they deflate RMSDs from SRB-grid cell mean values everywhere while inflating RMSDs from staggered SRB-grid cell mean values in many regions, in particular over the tropical oceans. The criterion is much better fulfilled by the BCvd$p2$ and BCvd$p0$ methods. The RMSDs are generally greater after bias correction with ~~BC$v$da1 (Fig. 7c, i). Moreover, the RMSD change patterns at the SRB and the staggered SRB grid are much more similar after bias-correction with BC$v$da2 than with BC$v$da1, except for rlds over land, where RMSDs are more noisy and smaller on average at the staggered SRB grid (Fig. 7f,l).~~

~~Bias correction with BC$vtp0$ yields virtually identical RMSD change patterns at the SRB and the staggered SRB grid. For rsds, these patterns are very similar to those obtained with BC$vtp2$. In contrast, rlds RMSDs over land are reduced much more by BC$vtp0$ than by BC$vtp2$. As a consequence, the~~ BCvd$p2$ than with BCvd$p0$, i.e., BCvd$p2$ produces data with greater sub-SRB-grid scale spatial variability than BCvd$p0$. This difference is most visible for longwave radiation, for which BCvd$p0$ produces a stark land-sea contrast of ~~rlds RMSD changes is much larger for BC$vtp0$ than for BC$vtp2$, in particular over the tropics. The~~ RMSD changes with strong RMSD reductions over land whereas BCvd$p0$ does so to a much lesser extent. This strong deflation of ~~rlds~~ sub-SRB-grid scale spatial variability ~~produced~~ by BC$vtp0$ ~~over land is considered~~ is believed to be another artefact caused by the bilinear interpolation of SRB data to the E2OBS grid. ~~The magnitude of the deflation of more than 40% in most cases cannot be explained by the associated deflation of SRB-grid scale temporal variability, which in most cases does not exceed 8% (Fig. 2g,j). Presumably, it is mainly due to a deflation of the sub-SRB-grid scale spatial variability of long-term mean rlds values caused by the bilinear interpolation.~~

5 Summary and conclusions

This article introduces various parametric quantile mapping methods for the bias correction of E2OBS daily mean surface downwelling longwave and shortwave radiation using ~~SRB satellite estimates. Bias correction results are cross-validated as well as validated using independent BSRN ground observations.~~ the corresponding SRB data. The quantile mapping methods

differ in (i) the time scale at which they operate, (ii) if and how they take physical upper radiation bounds into account, and (iii) how they handle the spatial resolution gap between E2OBS and SRB.

~~As expected,~~ A cross-validation ~~results suggest~~ at the SRB-grid scale demonstrates that statistics of daily mean radiation are ~~mostly~~ better corrected by methods operating at the daily time scale than by ~~those~~ methods operating at the monthly time scale, and vice versa for statistics of monthly mean radiation. ~~However, compared to BSRN observations, daily mean longwave radiation is mostly better corrected by the methods operating at the monthly time scale because the methods operating at the daily time scale adulterate the day-to-day variability of the original data. Given the composition of the E2OBS data, this suggests that the day-to-day variability of ERA-Interim longwave radiation is mostly more realistic than the corresponding SRB estimates.~~

~~While the methods operating~~ Since these performance differences are statistically more significant at the daily than at the monthly time scale~~are best at adjusting the interannual variability of monthly mean values, our cross-validation results show that for calibration and validation sample sizes of only 12 years each, the methods operating~~ , overall, bias correction at the daily time scale ~~perform almost as well if sampling errors are taken into account. In that case, the methods operating~~ is deemed more effective then bias correction at the monthly time scale~~are in fact overcorrecting. This result should be seen as an incentive to develop bias correction methods that take sampling errors into account.~~ .

~~Methods that do and that do not take~~ The cross-validation further suggests that it is generally worthwhile to explicitly take physical upper radiation bounds into account during quantile mapping~~are applied to daily mean longwave radiation. It is found that multi-year monthly maximum values as well as the shape of the whole distribution is better adjusted by methods that respect the estimated upper bounds.~~ . For shortwave radiation, different approaches to ~~estimating the upper bounds~~ their estimation are tested. A simple ~~method based on~~ approach using running maximum values is found to ~~perform better in cross-validation than~~ outperform a more complicated one ~~that uses rescaled~~ based on daily mean insolation at the top of the atmosphere ~~. Arguably, that is because~~ (rsdt). This must be due to other factors besides rsdt that influence the upper physical bounds ~~to downwelling shortwave radiation at the surface (rsds) are determined by downwelling shortwave radiation at the top of the atmosphere (rsdt) as well as by other factors such as atmospheric humidity. In fact, the~~ of rsds. Atmospheric humidity is an example for such a factor: The highest rsds values usually occur under clear-sky conditions and they are the higher the drier the atmosphere. Atmospheric humidity ~~, in turn ,~~ in turn is limited by the water vapour holding capacity of the atmosphere, which is controlled by atmospheric temperature. ~~Since the~~ The climatology of atmospheric temperature lags that of rsdt. Hence, the climatology of ~~upper bounds to rsds can also~~ the upper physical bounds of rsds can be expected to deviate from ~~any rescaled~~ the rsdt climatology.

The ~~most simple approach tested here to bridging the spatial resolution gap between E2OBS and SRB data is to bilinearly interpolate the more coarsely resolved SRB data to the E2OBS grid and to use these interpolated data for a bias correction at~~ cross-validation also reveals to what extent the bilinear spatial interpolation of SRB data to the E2OBS ~~grid. Methods operating this way are found to erroneously deflate both the temporal and the spatial variability of the original E2OBS data. On the other hand, relative to BSRN observations, the interpolation is found to more often than not benefit the multi-year monthly mean longwave radiation. This positive effect of the interpolation is interpreted as the result of an implicit elevation adjustment from~~

~~the SRB to the E2OBS grid. This outcome encourages elevation adjustments preceding future bias corrections of longwave radiation using, e. g., the Stackhouse Jr. et al. (2011) formula or the Cosgrove et al. (2003) method~~grid prior to bias correction with the BC$vtp0$ methods deflates day-to-day variability. This variability deflation has a greater effect on bias correction performance than a change of if and how physical upper radiation bounds are taken into account during quantile mapping, but a much smaller effect than a change of the time scale at which the quantile mapping is carried out.

Lastly, the cross-validation at the daily time scale shows that none of the quantile mapping methods tested here is perfect, concerning in particular the adjustment of distribution tails and moments of higher than second order. This indicates that the true distribution of rlds and rsds is not always exactly normal or beta, as assumed by the parametric quantile mapping methods tested here. Potentially, non-parametric quantile mapping methods (that do not rely on such assumptions) could yield better cross-validation results as long as overfitting is avoided (e.g., Gudmundsson et al., 2012). However, an introduction of and comparison to such methods is beyond the scope of this article.

~~The second approach used here aggregates the original~~ To bridge the spatial resolution gap between E2OBS ~~data to the SRBgrid, where the bias correction is done, and disaggregates these aggregated~~ and SRB, the methods used for the production of EWEMBI rlds and rsds deterministically disaggregate the E2OBS data previously aggregated to and bias-corrected ~~data back to the E2OBS grid. The deterministic disaggregation~~ at the SRB grid. It is shown that the method used for that ~~purpose is found to deflate~~ disaggregation introduces artefacts in the sub-SRB-grid scale spatial variability~~of the original data. Yet it also has its merits, where sub-SRB-grid scale spatial gradients in radiation statistics are very large, such as over islands covering just one E2OBS grid cell. There, the aggregation-correction-disaggregation of rsds produces substantially lower root-mean-square deviations from daily mean BSRN values than the other approaches.~~

~~The third approach introduced here corrects biases~~ , which can be overcome by applying quantile mapping directly at the E2OBS grid using either bilinearly interpolated SRB data or target distribution parameters that are based on the more coarsely resolved SRB data as well as on sub-SRB-grid scale spatial variability present in the original E2OBS data~~to the end of adjusting~~ . This latter approach yields both good cross-validation results at the SRB-grid scale ~~biases while preserving~~ and suitable adjustments of the sub-SRB-grid scale spatial variability. ~~The latter objective is achieved here by preserving sub-SRB-grid scale ratios between climatological standard deviations and offsets between climatological mean values and upper bounds. Potentially more suitable non-linear relationships might be tested in future studies. By design, the third approach precludes any of the variability deflations caused by the first and second approach.~~

~~The cross-validation reveal that substantial skewness biases remain after~~

The best methods identified here are therefore BClda2 for rlds and BCsdb2 for rsds. In comparison to BClda1 and BCsda1 used for the production of EWEMBI rlds and rsds, bias correction with ~~any of the parametric quantile mapping methods introduced here, as these do not explicitly adjusted third- andhigher-order moments. Better results might be obtained using non-parametric quantile mapping methods~~these methods yields more natural sub-SRB-grid scale spatial variability and, in the case of rsds, slightly better cross-validation results at the SRB-grid scale.

~~Deviations of bias-corrected E2OBS data from BSRN observations turn out to be dominated by the corresponding SRB data biases. This exemplifies that bias correctionusually does not actually correct biases but merely adjusts them to those of another~~

~~dataset, which is why some colleagues prefer the term *bias adjustment* over *bias correction*. In the example studied here, this is most painfully visible at polar BSRN stations, where in most cases bias correction using SRB estimates substantially increases E2OBS shortwave radiation biases relative to the BSRN ground truth. Yet apart from these cases, the validation against BSRN observations suggests that, overall, bias correction of E2OBS radiation using SRB data has a slightly positive effect on longwave radiation and a neutral effect on shortwave radiation. For the EWEMBI dataset (Lange, 2016), E2OBS longwave radiation was adjusted to the SRB 3.1 primary-algorithm product using the BClda1 method, and E2OBS shortwave radiation was adjusted to the SRB 3.0 primary-algorithm product using the BCsda1 method. In that application, the full 24 years worth of SRB data and the same 24 years worth of E2OBS data aggregated to the SRB grid were used to derive the transfer function parameters. The present study identifies shortcomings of the BClda1 and BCsda1 methods and tests modifications of these methods as remedies. In terms of cross-validation results and variability deflation issues, the best methods tested here are BClda2 and BCsdb2, whereas biases relative to BSRN observations are most effectively reduced by BClmb0 and BCsdb1.~~

Data availability. The EWEMBI dataset is publicly available via https://doi.org/10.5880/pik.2016.004.

Appendix A: Quantile mapping and statistical downscaling

Quantile mapping is used to adjust the distribution of values from a data sample. In the context of bias correction, the distribution to be adjusted – the source distribution – is believed or known to be more biased than the distribution the source distribution is adjusted to – the target distribution. In practise, source and target distributions are empirically estimated from the respective samples, in the present case of E2OBS and SRB radiation data, in the form of cumulative distribution functions (CDFs) F^{E2OBS} and F^{SRB}, respectively. Quantile mapping is then defined by

$$x \mapsto F^{\mathrm{SRB}\,-1}(F^{\mathrm{E2OBS}}(x)), \tag{A1}$$

where $F^{\mathrm{SRB}\,-1}(F^{\mathrm{E2OBS}}(\,\cdot\,))$ is called the transfer function.

Quantile mapping is called parametric if the CDFs are assumed to take certain functional forms. Their estimation then reduces to the estimation of the parameters and these functions. Otherwise, quantile mapping is called non-parametric and CDFs are estimated by estimating selected quantiles, between and beyond which quantiles are interpolated and extrapolated, respectively (e.g., Gudmundsson et al., 2012).

In the present study, source and target distributions are assumed to be normal or beta distributions. Mean values and variances of normal distributions are estimated by running mean values of multi-year daily sample mean values and variances. Lower and upper bounds of beta distributions are set to zero and estimated by physical upper limits of daily mean radiation, respectively. Shape parameters of beta distributions are estimated with the method of moments (Wilks, 1995) using running mean values of multi-year daily sample mean values and variances.

Bias correction includes a spatial disaggregation or downscaling step if the data behind source and target distributions have different spatial resolution, as in the present case, or represent area mean values and point values, as in the case of quantile mapping between gridded and station data. If the data behind the target distribution have higher resolution/represent finer spatial scales than the data behind the source distribution, then quantile mapping may lead to both temporal and spatial variability inflation (Maraun, 2013). For the reverse case, the present study shows how quantile mapping may lead to both temporal and spatial variability deflation. Maraun (2013) suggests to solve the inflation issue with stochastic downscaling. It is shown here that the deflation issue of the reverse case can also be overcome with deterministic downscaling at the transfer function level.

Appendix B: Daily mean insolation at the top of the atmosphere

Over the course of a year, the total solar irradiance, S, varies according to $S = S_0(1 + e\cos(\Theta))^2$, where $S_0 = 1360.8\,\mathrm{Wm}^{-2}$ is the solar constant (Kopp and Lean, 2011), $e = 0.0167086$ is the Earth's current orbital eccentricity and Θ is the angle to the Earth's position from its perihelion, as seen from the Sun. If the orbital angular velocity of the Earth is approximated to vary sinusoidally in time then the total solar irradiance on day n after January 1 of the first year of a four-year cycle including one leap year is approximately given by

$$S = S_0\left(1 + e\cos\left(2\pi\frac{n-2}{365.25} + 2e\sin\left(2\pi\frac{n-2}{365.25}\right)\right)\right)^2, \tag{B1}$$

since S is at its maximum when the Earth is at its perihelion, which on average occurs on January 3.

The daily mean insolation at the top of the atmosphere, rsdt, at some fixed geolocation depends on the location's latitude, ϕ, and on the declination of the Sun, δ, which varies over the course of a year. On day n after January 1 of the first year of a four-year cycle including one leap year, the declination of the sun is approximately given by

$$\sin\delta = \cos\left(2\pi\frac{n+10}{365.25} + 2e\sin\left(2\pi\frac{n-2}{365.25}\right)\right)\sin\delta_{\min}, \tag{B2}$$

since δ is at its minimum value $\delta_{\min} = -23.4392811°$ at the December solstice, which on average occurs on December 22. Latitude and declination of the Sun determine the hour angle at sunrise, h, according to

$$\cos h = \min\{1, \max\{-1, -\tan\phi\tan\delta\}\}. \tag{B3}$$

The daily mean insolation at the top of the atmosphere at latitude ϕ on day n is then given by

$$\mathrm{rsdt} = \frac{S}{\pi}(h\sin\phi\sin\delta + \sin h\cos\phi\cos\delta). \tag{B4}$$

For a given latitude, the rsdt climatology used to estimate the upper bounds of the climatological beta distribution of rsds in the $\mathrm{BCsda}x$ methods is derived using Eqs. (B1)–(B4) to compute rsdt over a four-year cycle including one leap year and then averaging calendar day values over the four cases of leap year occurrence in the four-year cycle.

Appendix C: Two-sample Kolmogorov-Smirnov test and Kuiper's two-sample test

The overall effectivity of the bias correction methods introduced in this study is measured by similarities of empirical CDFs of SRB and E2OBS data before and after bias correction using the two-sample Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933; Smirnov, and Kuiper's two-sample test (Kuiper, 1962; Stephens, 1965). Let F_1 be the empirical CDF of uncorrected or corrected daily or monthly mean longwave or shortwave E2OBS data for one particular grid cell, calendar month and validation data sample, with all corresponding values pooled into one distribution, and let F_2 be the empirical CDF of the corresponding SRB ~~or BSRN~~ data. Then the two-sample KS test statistic, D, and Kuiper's two-sample test statistic, V, of these CDFs ~~is given by $D = \sup_r |F_1(r) - F_2(r)|$.~~ are given by

$$D = \sup_r |F_1(r) - F_2(r)|\,, \tag{C1}$$

$$V = \sup_r (F_1(r) - F_2(r)) + \sup_r (F_2(r) - F_1(r))\,. \tag{C2}$$

The null hypothesis of both the KS test and Kuiper's test is that the two data samples whose empirical CDFs are compared have the same underlying distribution. According to Vetterling et al. (1992, Sect. 14.3), the probability p of incorrectly rejecting this null hypothesis can be approximated by

$$p = 1 - F\left(\left[\sqrt{n} + 0.12 + 0.11/\sqrt{n}\right] D\right),$$

$$p = 1 - F\left(\left[\sqrt{n} + 0.12 + 0.11/\sqrt{n}\right] D\right) \quad \text{and} \tag{C3}$$

$$p = 1 - G\left(\left[\sqrt{n} + 0.155 + 0.24/\sqrt{n}\right] V\right) \tag{C4}$$

for the KS test and Kuiper's test, respectively, where F ~~is the CDF of the Kolmogorov distribution,~~ and G are the CDFs of the asymptotic distributions of $\sqrt{n}D$ and $\sqrt{n}V$, respectively, $n = n_1 n_2/(n_1+n_2)$ is the effective sample size, and n_1 and n_2 are the sizes of the samples behind F_1 and F_2, respectively. This approximation of the true p-value is not only asymptotically accurate but already quite good for $n \geq 4$ ~~(cf. von Mises, 1964; Vetterling et al., 1992)~~(cf. Stephens, 1970; Vetterling et al., 1992).

In order to adjust these p-values for potential autocorrelations in the samples compared here, which are in fact time series, n_1 and n_2 in the formula for n are replaced by $n_1(1-\rho_1)$ and $n_2(1-\rho_2)$, respectively, as proposed by Xu (2013), where the autoregression coefficients ρ_1 and ρ_2 of first-order autoregressive processes fitted to the time series are estimated by the respective sample autocorrelation at lag one.

Appendix D: Window length for running mean and maximum calculations

The climatologies of mean values, variances, and upper bounds of daily mean radiation estimated by the $\mathrm{BC}vdpx$ methods are based on running mean values of empirical multi-year daily mean values, variances and running maximum values, respectively.

A common window length of 25 days is used for these running mean and maximum value calculations (cf. Table 1). An obvious question is how sensitive the bias correction results are to the choice of this window length.

The question is addressed here via variants of the $\mathrm{BC}v\mathrm{da1}$ methods that use uneven window lengths between 10 and 40 days for their running mean and maximum value calculations and are otherwise identical to the $\mathrm{BC}v\mathrm{da1}$ method introduced in Sect. 3.1.1. The performance of these $\mathrm{BC}v\mathrm{da1}$ variants is then quantified by p-values of two-sample KS statistics of bias-corrected E2OBS data cross-validated against SRB data (cf. Sect. ~~4 and Sect.~~ 4.1 and Appendix C). The window lengths that maximise these p-values vary considerably with location, calendar month and calibration data sample (Fig. D1). The reason for this high variability is illustrated in Fig. D2, where the overall performance of the $\mathrm{BC}v\mathrm{da1}$ variants, quantified by p-values of two-sample KS statistics aggregated over time (calendar months) and space (grid cells), is shown to only weakly depend on the chosen window length.

The optimal window length is thus highly uncertain. For longwave ~~/shortwave~~(shortwave) radiation, the overall performance of the $\mathrm{BC}v\mathrm{da1}$ variants is slightly higher for window lengths from the upper ~~/lower~~(lower) end of the investigated range (Fig. D2). For practical matters, one can apply the methods using any window length between 10 and 40 days and expect similarly well adjusted radiation biases. The choice of 25-day running windows made here for both longwave and shortwave radiation ensures a close-to-optimal performance of the $\mathrm{BC}v\mathrm{da1}$ methods for both variables.

Competing interests. The author declares that no competing interests are present.

Acknowledgements. The author is grateful to Katja Frieler, Jan Volkholz and Alex Cannon for various helpful discussions at different stages of this work, to ~~Hannes Müller-Schmied for suggesting the validation against BSRN data, to Gert König-Langlo and Amelie Driemel for their help during BSRN data acquisition and processing, to~~ Paul W. Stackhouse Jr. for his guidance with SRB data products and the provision of SRB elevation data, ~~and~~ to Graham Weedon and Emanuel Dutra for their guidance during the initial stage of assembling the EWEMBI dataset, and to the two anonymous referees who provided highly valuable comments to the discussion paper version of this manuscript. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641816 Coordinated Research in Earth Systems and Climate: Experiments, kNowledge, Dissemination and Outreach (CRESCENDO).

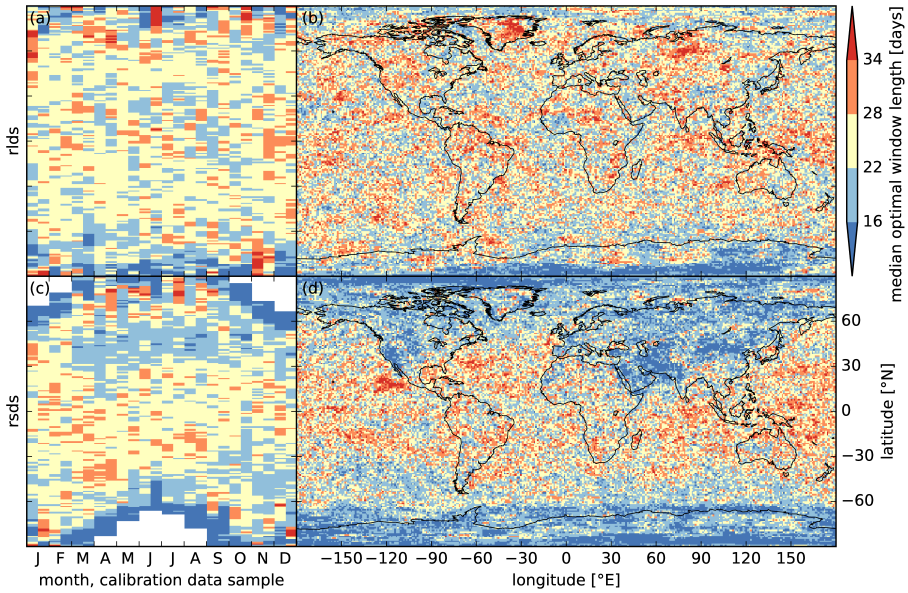

Figure D1. Optimal window length for running mean and maximum calculations that precede the estimation of parameters of the climatological distributions of longwave ($v = \mathrm{l}$; **top**) and shortwave ($v = \mathrm{s}$; **bottom**) radiation that are used for bias correction with $\mathrm{BC}v\mathrm{da1}$ (cf. Table 1). Window lengths are varied between 10 and 40 days. Optimal window lengths maximise the p-value of the two-sample KS statistic of bias-corrected E2OBS data cross-validated against SRB data (cf. Sect. 4 and ~~Sect.~~ Appendix C) and are determined individually for every grid cell, calendar month (with all corresponding values pooled into one distribution) and calibration data sample (every1st, every2nd). Zonal medians of optimal window lengths for each month and calibration data sample are shown in panels (**a**) and (**c**). Results are masked in (**c**) where and when the monthly mean rsdt (Eqs. (B1)–(B4)) is less than $1\,\mathrm{W\,m^{-2}}$. Panels (**b**) and (**d**) show medians of optimal window lengths over months and calibration data samples.

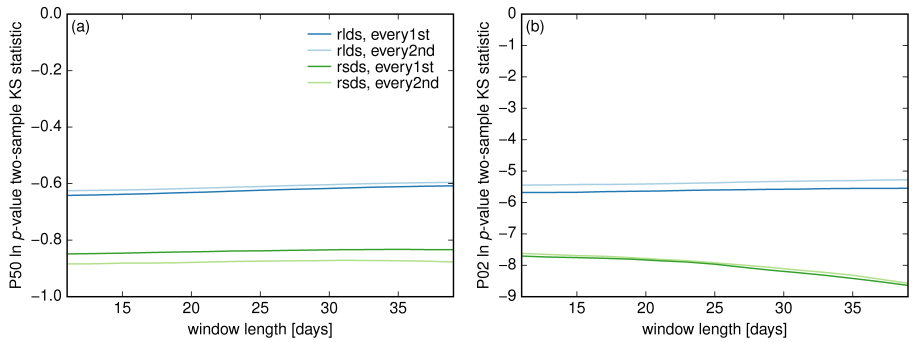

Figure D2. Dependence of two-sample KS statistic p-values on window length for different radiation types and calibration data samples (see text and Fig. D1). Plotted are the grid-cell area-weighted 50th (~~left~~**a**) and 2nd (~~right~~**b**) percentiles of the natural logarithms of the p-values over months, latitudes and longitudes.

References

Calton, B., Schellekens, J., and Martinez-de la Torre, A.: Water Resource Reanalysis v1: Data Access and Model Verification Results, https://doi.org/10.5281/zenodo.57760, 2016.

Cannon, A. J.: Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables, Climate Dynamics, pp. 1–19, https://doi.org/10.1007/s00382-017-3580-6, 2017.

Chang, J., Ciais, P., Wang, X., Piao, S., Asrar, G., Betts, R., Chevallier, F., Dury, M., François, L., Frieler, K., Ros, A. G. C., Henrot, A.-J., Hickler, T., Ito, A., Morfopoulos, C., Munhoven, G., Nishina, K., Ostberg, S., Pan, S., Peng, S., Rafique, R., Reyer, C., Rödenbeck, C., Schaphoff, S., Steinkamp, J., Tian, H., Viovy, N., Yang, J., Zeng, N., and Zhao, F.: Benchmarking carbon fluxes of the ISIMIP2a biome models, Environmental Research Letters, 12, 045 002, https://doi.org/10.1088/1748-9326/aa63fa, 2017.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Marshall, C., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D., and Meng, J.: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD003118, 8842, 2003.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Dutra, E.: Report on the current state-of-the-art Water Resources Reanalysis, Earth2observe deliverable no. d.5.1, http://earth2observe.eu/files/Public%20Deliverables, 2015.

Fisher, R. A.: Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population, Biometrika, 10, 507–521, http://www.jstor.org/stable/2331838, 1915.

Fisher, R. A.: On the "probable error" of a coefficient of correlation deduced from a small sample, Metron, 1, 3–32, http://hdl.handle.net/2440/15169, 1921.

Frieler, K., Betts, R., Burke, E., Ciais, P., Denvil, S., Deryng, D., Ebi, K., Eddy, T., Emanuel, K., Elliott, J., Galbraith, E., Gosling, S. N., Halladay, K., Hattermann, F., Hickler, T., Hinkel, J., Huber, V., Jones, C., Krysanova, V., Lange, S., Lotze, H. K., Lotze-Campen, H., Mengel, M., Mouratiadou, I., Müller Schmied, H., Ostberg, S., Piontek, F., Popp, A., Reyer, C. P. O., Schewe, J., Stevanovic, M., Suzuki, T., Thonicke, K., Tian, H., Tittensor, D. P., Vautard, R., van Vliet, M., Warszawski, L., and Zhao, F.: Assessing the impacts of 1.5°C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), Geoscientific Model Development Discussions, 2016, 1–59, https://doi.org/10.5194/gmd-2016-229, 2016.

Frieler, K., Lange, S., Piontek, F., Reyer, C. P. O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T. D., Elliott, J., Galbraith, E., Gosling, S. N., Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, H., Mouratiadou, I., Pierson, D., Tittensor, D. P., Vautard, R., van Vliet, M., Biber, M. F., Betts, R. A., Bodirsky, B. L., Deryng, D., Frolking, S., Jones, C. D., Lotze, H. K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and

Yamagata, Y.: Assessing the impacts of 1.5°C global warming – simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), Geoscientific Model Development, 10, 4321–4345, https://doi.org/10.5194/gmd-10-4321-2017, 2017.

Garratt, J. R.: Incoming Shortwave Fluxes at the Surface—A Comparison of GCM Results with Observations, Journal of Climate, 7, 72–80, https://doi.org/10.1175/1520-0442(1994)007<0072:ISFATS>2.0.CO;2, 1994.

Gennaretti, F., Sangelantoni, L., and Grenier, P.: Toward daily climate scenarios for Canadian Arctic coastal zones with more realistic temperature-precipitation interdependence, Journal of Geophysical Research: Atmospheres, 120, 11,862–11,877, https://doi.org/10.1002/2015JD023890, 2015.

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

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset, International Journal of Climatology, https://doi.org/10.1002/joc.3711, 2013.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., and Piontek, F.: A trend-preserving bias correction – the ISI-MIP approach, Earth System Dynamics, 4, 219–236, https://doi.org/10.5194/esd-4-219-2013, 2013.

Iizumi, T., Takikawa, H., Hirabayashi, Y., Hanasaki, N., and Nishimori, M.: Contributions of different bias-correction methods and reference meteorological forcing data sets to uncertainty in projected temperature and precipitation extremes, Journal of Geophysical Research: Atmospheres, https://doi.org/10.1002/2017JD026613, 2017JD026613, 2017.

Ito, A., Nishina, K., Reyer, C. P. O., François, L., Henrot, A.-J., Munhoven, G., Jacquemin, I., Tian, H., Yang, J., Pan, S., Morfopoulos, C., Betts, R., Hickler, T., Steinkamp, J., Ostberg, S., Schaphoff, S., Ciais, P., Chang, J., Rafique, R., Zeng, N., and Zhao, F.: Photosynthetic productivity and its efficiencies in ISIMIP2a biome models: benchmarking for impact assessment studies, Environmental Research Letters, 12, 085 001, https://doi.org/10.1088/1748-9326/aa7a19, 2017.

Jones, P. W.: First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates, Monthly Weather Review, 127, 2204–2210, https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, 77, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Kiehl, J. T. and Trenberth, K. E.: Earth's Annual Global Mean Energy Budget, Bulletin of the American Meteorological Society, 78, 197–208, https://doi.org/10.1175/1520-0477(1997)078<0197:EAGMEB>2.0.CO;2, 1997.

Kistler, R., Collins, W., Saha, S., White, G., Woollen, J., Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., and Fiorino, M.: The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation, Bulletin of the American Meteorological Society, 82, 247–267, https://doi.org/10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2, 2001.

Kolmogorov, A.: Sulla determinazione empirica di una leggi di distribuzione, Giornale dell' Istituto Italiano degli Attuari, 4, 83–91, 1933.

König-Langlo, G., Sieger, R., Schmithüsen, H., Bücker, A., Richter, F., and Dutton, E.: The Baseline Surface Radiation Network and its World Radiation Monitoring Centre at the Alfred Wegener Institute, WCRP Report 24/2013, GCOS 174, Geneva, Switzerland, http://www.wmo.int/pages/prog/gcos/Publications/gcos-174.pdf, 2013.

Kopp, G. and Lean, J. L.: A new, lower value of total solar irradiance: Evidence and climate significance, Geophysical Research Letters, 38, https://doi.org/10.1029/2010GL045777, l01706, 2011.

Krysanova, V. and Hattermann, F. F.: Intercomparison of climate change impacts in 12 large river basins: overview of methods and summary of results, Climatic Change, 141, 363–379, https://doi.org/10.1007/s10584-017-1919-y, 2017.

Kuiper, N. H.: Tests concerning random points on a circle, in: Koninklijke Nederlandse Akademie van Wetenschappen, vol. 63 of *A*, pp. 38–47, 1962.

Lange, S.: EartH2Observe, WFDEI and ERA-Interim data Merged and Bias-corrected for ISIMIP (EWEMBI), https://doi.org/10.5880/pik.2016.004, 2016.

Ma, Q., Wang, K., and Wild, M.: Evaluations of atmospheric downward longwave radiation from 44 coupled general circulation models of CMIP5, Journal of Geophysical Research: Atmospheres, 119, 4486–4497, https://doi.org/10.1002/2013JD021427, 2013JD021427, 2014.

Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, Journal of Climate, 26, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1, 2013.

Müller Schmied, H., Müller, R., Sanchez-Lorenzo, A., Ahrens, B., and Wild, M.: Evaluation of Radiation Components in a Global Freshwater Model with Station-Based Observations, Water, 8, https://doi.org/10.3390/w8100450, 2016.

Roesch, A., Wild, M., Ohmura, A., Dutton, E. G., Long, C. N., and Zhang, T.: Assessment of BSRN radiation records for the computation of monthly means, Atmospheric Measurement Techniques, 4, 339–354, https://doi.org/10.5194/amt-4-339-2011, 2011.

Ruane, A. C., Goldberg, R., and Chryssanthacopoulos, J.: Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation, Agricultural and Forest Meteorology, 200, 233–248, https://doi.org/10.1016/j.agrformet.2014.09.016, 2015.

Rust, H. W., Kruschke, T., Dobler, A., Fischer, M., and Ulbrich, U.: Discontinuous Daily Temperatures in the WATCH Forcing Datasets, Journal of Hydrometeorology, 16, 465–472, https://doi.org/10.1175/JHM-D-14-0123.1, 2015.

Schild, P.: Macro-enabled Excel spreadsheet to calculate hourly-averages from BSRN .dat files, http://hdl.handle.net/10013/epic.48977.d002, 2016.

Sheffield, J., Goteti, G., and Wood, E. F.: Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling, Journal of Climate, 19, 3088–3111, https://doi.org/10.1175/JCLI3790.1, 2006.

Smirnov, N.: Table for Estimating the Goodness of Fit of Empirical Distributions, The Annals of Mathematical Statistics, 19, 279–281, https://doi.org/10.1214/aoms/1177730256, 1948.

Stackhouse Jr., P. W., Gupta, S. K., Cox, S. J., Mikovitz, C., Zhang, T., and Hinkelman, L. M.: The NASA/GEWEX surface radiation budget release 3.0: 24.5-year dataset, Gewex news, 21(1):10–12, http://www.gewex.org/resources/gewex-news/, 2011.

Stephens, M. A.: The Goodness-Of-Fit Statistic V_n: Distribution and Significance Points, Biometrika, 52, 309–321, https://doi.org/10.2307/2333685, 1965.

Stephens, M. A.: Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables, Journal of the Royal Statistical Society. Series B (Methodological), 32, 115–122, http://www.jstor.org/stable/2984408, 1970.

Switanek, M. B., Troch, P. A., Castro, C. L., Leuprecht, A., Chang, H.-I., Mukherjee, R., and Demaria, E. M. C.: Scaled distribution mapping: a bias correction method that preserves raw climate model projected changes, Hydrology and Earth System Sciences, 21, 2649–2666, https://doi.org/10.5194/hess-21-2649-2017, 2017.

Trenberth, K. E., Fasullo, J. T., and Kiehl, J.: Earth's Global Energy Budget, Bulletin of the American Meteorological Society, 90, 311–323, https://doi.org/10.1175/2008BAMS2634.1, 2009.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L.

V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis, Quarterly Journal of the Royal Meteorological Society, 131, 2961–3012, https://doi.org/10.1256/qj.04.176, 2005.

Veldkamp, T. I. E., Wada, Y., Aerts, J. C. J. H., Döll, P., Gosling, S. N., Liu, J., Masaki, Y., Oki, T., Ostberg, S., Pokhrel, Y., Satoh, Y., Kim, H., and Ward, P. J.: Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century, 8, 15 697–, https://doi.org/10.1038/ncomms15697, 2017.

Vetterling, W. T., Press, W. H., Teukolsky, S. A., and Flannery, B. P.: Numerical Recipes in C (The Art of Scientific Computing), Cambridge University Press, 2nd edn., 1992.

von Mises, R.: Mathematical Theory of Probability and Statistics, Academic Press, New York, 1964.

Weedon, G. P., Gomes, S., Viterbo, P., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M.: The WATCH forcing data 1958–2001: A meteorological forcing dataset for land surface and hydrological models, Technical report no. 22, http://www.eu-watch.org/publications/technical-reports, 2010.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M.: Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century, Journal of Hydrometeorology, 12, 823–848, https://doi.org/10.1175/2011JHM1369.1, 2011.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resources Research, 50, 7505–7514, https://doi.org/10.1002/2014WR015638, 2014.

Wild, M., Folini, D., Schär, C., Loeb, N., Dutton, E. G., and König-Langlo, G.: The global energy balance from a surface perspective, Climate Dynamics, 40, 3107–3134, https://doi.org/10.1007/s00382-012-1569-8, 2013.

Wild, M., Folini, D., Hakuba, M. Z., Schär, C., Seneviratne, S. I., Kato, S., Rutan, D. A., Ammann, C., Wood, E. F., and König-Langlo, G.: The energy balance over land and oceans: an assessment based on direct observations and CMIP5 climate models, Climate Dynamics, 44, 3393–3429, https://doi.org/10.1007/s00382-014-2430-z, 2015.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego, CA, 1995.

Xu, X.: Methods in Hypothesis Testing, Markov Chain Monte Carlo and Neuroimaging Data Analysis, Ph.D. thesis, Harvard University, http://nrs.harvard.edu/urn-3:HUL.InstRepos:11108711, 2013.

Zhang, T., Stackhouse, P. W., Gupta, S. K., Cox, S. J., and Mikovitz, J. C.: The validation of the GEWEX SRB surface long-wave flux data products using BSRN measurements, Journal of Quantitative Spectroscopy and Radiative Transfer, 150, 134–147, https://doi.org/10.1016/j.jqsrt.2014.07.013, topical issue on optical particle characterization and remote sensing of the atmosphere: Part I, 2015.

Zhao, M. and Dirmeyer, P. A.: Production and analysis of GSWP-2 near-surface meteorology data sets, vol. 159, Center for Ocean-Land-Atmosphere Studies Calverton, http://ww.w.monsoondata.org/gswp/gswp2data.pdf, 2003.