Best practice for the calibration of seasonal forecasts of climate indices

Jonas Bhend, Mark Liniger

July 29, 2016

# Abstract

Climate information indices (CIIs) represent a simple and easy-to-implement way to describe impact-relevant quantities derived from basic meteorological variables such as rainfall or temperature. CIIs are usually aggregated quantities derived from daily time series of said underlying variables. The computation of CIIs often involves non-linear transfer functions (e.g. based on thresholds). This poses challenges in a seasonal forecasting context where such indices have to be derived from daily time series of forecasting systems with sometimes considerable systematic errors. Here we analyze how various calibration and recalibration methods applied to the underlying meteorological variable affect forecast skill of CII forecasts. In addition, we also recalibrate the seasonal CII forecasts and contrast the direct approach (i.e. daily (re)calibration only) and the two step approach (i.e. daily calibration and recalibration of seasonal quantities).

# Introduction

Seasonal forecasting is an emerging research area and its predictions are becoming increasingly used in decision processes. In order for seasonal forecasts to be useful, these have to be skillful, reliable and relevant to users. One way to increase the relevance of seasonal forecasts is to derive impact-relevant quantities from the basic meteorological outputs of seasonal forecasting systems. Examples of such climate information indices (CIIs) include the number of frost days, heating or growing degree days, days with snowfall, or the apparent temperature (a combination of humidity and air temperature related to human comfort). CIIs are usually derived from daily time series and the computation often involves non-linear transfer functions and absolute thresholds. Therefore, the underlying meteorological variables from the forecasting system have to be calibrated, i.e. systematic biases of the forecasting system with respect to the verifying observations have to be removed before CIIs are computed. Without calibration, the base rate or average frequency of occurrence of CIIs may be much too low or high compared to observations.

In addition to being unbiased, we also want probabilistic forecasts to be reliable. That is, forecasted probabilities should reflect observed frequencies on average. To achieve this, forecasts may be recalibrated. In contrast to calibration, recalibration also adjusts the intra-ensemble spread (the forecast probabilities) to ensure reliability of the forecasts.

Here we analyze the skill of seasonal CII forecasts derived from calibrated daily time series. We contrast calibration schemes on the daily input data (here temperature) with increasing complexity and also analyze the effect of an additional recalibration of the seasonally aggregated CII forecasts. In order to better structure the discussion and in order to avoid the pitfall that often CIIs are only relevant in some parts of the world (e.g. frost days), we analyze synthetic generalized CIIs that are not computed with respect to an absolute and spatially uniform threshold, but with respect to a percentile of the long-term observed climatology.

# Forecast data and verifying observations

We use long-range forecasts for daily temperature from the European Centre for Medium-range Weather Forecasting???s (ECMWF) System 4 seasonal forecasting model (Molteni et al. 2011). The ECMWF System 4 is a fully coupled ocean-atmosphere dynamical forecasting system. The atmospheric component of ECMWF System 4 corresponds to an earlier version of ECMWF???s operational medium-range weather forecasting model. The atmospheric model is a spectral model run at T255 horizontal resolution (approximately 0.7 degree) with 91 vertical levels. The ocean model (NEMO) is run at 1 degree horizontal resolution and 42 vertical levels with a refinement in resolution in equatorial regions.

The real-time forecasts of System 4 are issued every month for the coming 215 days (approximately 7 months) and consist of 51 members initialized at the beginning of the month. Historical re-forecasts (hindcasts) using System 4 are available since 1981. We analyze these hindcasts for the boreal winter (DJF) and summer (JJA) seasons following the initialization dates in November and May respectively.

Indices derived from hindcasts from 1981 to 2013/14 are compared to indices derived from daily temperature time series of the ERA Interim reanalysis (Dee et al. 2011). In contrast to station based observations, the ERA Interim suffers less from representativity and homogeneity errors. On the other hand, systematic model errors of the ERA Interim system will affect the representation of temperature in the reanalysis.

Previous to the analysis, we have interpolated the daily seasonal forecast and reanalysis data on a common 2x2 degree latitude longitude grid.

# Climate information indices

We compute a set of simple climate information indices (CIIs) derived from daily temperature time series. These CIIs include mean seasonal temperature (tas), seasonal mean inter-diurnal temperature variability (ITV), and a set of seasonally accumulated departures above (below) a given threshold. ITV and the threshold exceedance indices are shortly introduced in the following.

The inter-diurnal temperature variability (ITV) is computed following Cattiaux et al. (2015) where ITV is defined as the mean absolute daily temperature difference. Threshold exceedances are computed relative to percentiles of the daily temperature distribution from the ERA Interim reanalysis. The percentiles are computed using a 31-day moving window centered about the day of year of interest. Based on the observed percentile threshold, we compute accumulated threshold exceedances for the verifying observations and the (calibrated) daily forecast time series. In boreal summer (JJA) we compute accumulated exceedances of the 90th, 95th and 98th daily percentiles (denoted as DD10, DD05 and DD02). In contrast, we compute accumulated temperature deviations below the 10th, 5th and 2nd percentile in boreal winter (DJF, also denoted as DD10, DD05, and DD02). These accumulated temperature deviations above (below) a pre-specified threshold are expected to behave very similarly to the traditional degree days that are widely used in the impact community (Christenson, Manz, and Gyalistras 2006 Schlenker, Hanemann, and Fisher (2007)). Forecasts of degree days based on absolute thresholds, however, are trivial or difficult to interpret in regions where the threshold is never or only very rarely exceeded. Our indices of accumulated threshold exceedances, on the other hand, are relevant and applicable globally due to their definition relative to the local climatology.

# Calibration methods

We compare variants of a regression-based calibration approach with a quantile mapping approach. While the regression approach allows us to explicitly recalibrate the ensemble spread to produce reliable predictions. While the ensemble spread may be affected by quantile mapping, on the other hand, explicit recalibration of ensemble predictions is not possible with quantile mapping.

## Quantile mapping

In addition to the regression-based calibration and recalibration approaches, we also use quantile mapping (Panofsky and Brier 1968) to calibrate the daily temperature series. We apply the quantile correction using a 31-day window centred on the day of interest to account for the seasonal cycle and lead-time dependence of model errors as in (**???**)??l2011 and Wilcke, Mendlik, and Gobiet (2013).

## Calibration in a linear regression framework

Under a linear regression framework, we perform the calibration of probabilistic ensemble forecasts in two steps: in the first step, the signal calibration, the mean squared error of the deterministic forecast given by the ensemble mean is minimized and in the second step, the recalibration, the ensemble spread is adjusted in order to produce reliable forecasts. Given re-forecasts with m ensemble members for lead times where , , denote the lead time, forecast, and ensemble member respectively, and given verifying observations , our linear model for calibration can be expressed as follows.

is the calibrated forecast for lead time , forecast , and ensemble member . The bar, tilde, and ticks denote the lead time dependent climatology, the ensemble mean, and anomalies respectively, such that

is the lead time dependent observed climatology,

are the ensemble mean anomalies from the model climatology at lead time , are anomalies of the forecast index or forecast year, and

are the ensemble anomalies from the ensemble mean at lead time and for forecast . The lead time dependent parameters , , and have to be estimated.

This linear model framework for forecast calibration allows us to explore a variety of calibration methods with increasing complexity. The variants of the calibration methods analyzed in this study are summarized in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name |  |  |  | Description |
| debias | 1 | 0 | 1 | Mean de-biasing |
| debiasRecal | 1 | 0 |  | ... with rescaling of ensemble spread |
| trend | 1 |  | 1 | Bias with linear time trend |
| trendRecal | 1 |  |  | ... with rescaling of ensemble spread |
| conditional |  | 0 | 1 | Bias conditional on ensemble mean forecast |
| conditionalRecal |  | 0 |  | ... with rescaling of ensemble spread |
| all |  |  | 1 | Bias conditional on ensemble mean forecast and with linear time trend |
| allRecal |  |  |  | ... with rescaling of ensemble spread |

Calibration methods used in this study. Parameters with a hat are estimated (see text).

### Signal calibration

Following Mahlstein et al. (2015) we apply a local linear regression smoothing on both the lead time dependent observation and model climatologies ( and respectively).

The parameters for the signal calibration, i.e. and , are estimated using ordinary least squares. Thereby, the root mean squared error of the deterministic forecast given by the ensemble mean is minimized. As the sample of re-forecast observation pairs is fairly limited and the estimates for the parameters are affected by considerable sampling uncertainty, we do not estimate the parameters for each lead time separately. Instead we estimate the parameters jointly for all lead times and allow the parameters to vary smoothly across lead times. This is implemented through interaction terms with a third-order polynomial on the lead time and also with interactions with a decaying exponential with e-folding time of 5 days to represent the initial shock early in the simulation. The latter interaction does not affect the calibration during the lead months 2-4 which are analyzed here, but helps to improve the fit early in the forecast.

The underlying regression assumption that residuals are identically and independently distributed is clearly violated due to auto-correlation and heteroscedasticity across lead times. Also the ensemble mean forecast anomalies are not completely noise free due to the limited ensemble size leading to a negative bias in the regression estimates (Allen and Stott 2003).

Using maximum likelihood estimation on the above regression equations is equivalent to minimizing the ignorance score on the training data as shown by Gneiting et al. (2005). In their study, Gneiting *et al.* find that fitting regression-based calibration methods minimizing the continuous ranked probability score produce sharper forecasts that are better calibrated. Such an approach, however, requires us to find the coefficients through numerical optimization rather than the analytic relationship available for maximum likelihood estimation. In the interest of computational tractability, we use maximum likelihood estimation despite its known limitations.

### Daily recalibration

In addition to signal calibration, we also adjust the ensemble spread to produce reliable forecasts. A necessary condition for reliability is given if the mean ensemble standard deviation (the ensemble spread) equals the root mean squared difference between the ensemble mean forecast and the verifying observations (A. P. Weigel, Liniger, and Appenzeller 2009).

We estimate the mean ensemble spread and the standard deviation of the observed residuals for each lead time separately. To minimize sampling uncertainty and spurious corrections, we then smooth the observed residual standard deviation and the ensemble spread using local linear regression. We fit the local linear regression on the log-transformed standard deviation and lead times to guarantee positive definite standard deviations and to allow for more strongly varying estimates early in the forecast and more smoothly changing standard deviations later on. The coefficient for ensemble spread inflation (shrinkage), , is then estimated using the smoothed standard deviations.

As shown in (1), the recalibration does not affect the spread-skill relationship, i.e. the ensemble spread is simply rescaled but sharp forecasts will remain sharper than dispersive forecasts. This is different to the ensemble modelling output statistics (EMOS) approach as proposed by Gneiting et al. (2005), with which the spread-skill relationship of the forecast can be altered as part of the calibration.

## Climate conserving recalibration on seasonal indices

In addition, we also analyze the effect of an additional recalibration of the seasonally aggregated CII forecasts. We use the climate conserving recalibration (CCR) as proposed by A. P. Weigel, Liniger, and Appenzeller (2009). The CCR can be expressed in the framework of linear regression introduced above as well (see appendix). In contrast to the daily calibration methods, we treat lead-time dependency implicitly by applying the CCR for each lead time separately. The CCR for forecast and ensemble member can thus be formulated as follows:

The regression coefficients and are given by

with denoting the correlation between the observations o and the ensemble mean forecast , is the observed standard deviation, is the standard deviation of the ensemble mean forecast, and is the mean ensemble standard deviation (mean ensemble spread).

As noted by Siegert (2015), such regression-corrected forecasts are systematically overconfident for out of sample re-calibration due to sampling errors causing a negative bias in forecast signal variance. Therefore, we adjust the CCR spread calibration to take into account the effect of uncertainty in the signal calibration (parameter ??) on the spread calibration (??). Following Siegert (2015), the adjusted parameter for forecast at lead time is expressed as follows:

The additional spread inflation to account for uncertainties in signal calibration is applied both for the recalibration of daily and seasonal average quantities.

# Verification

## Verification metrics

To validate the bias corrected CII forecasts, we use a small set of forecast verification metrics for probabilistic and deterministic forecasts. We focus on the continuous ranked probability score (CRPS) and its skill score. The CRPS is the integral of the Brier score for all thresholds and the CRPS for a deterministic forecast is equivalent to the mean absolute error (H. Hersbach 2000).

Where

is the forecast probability for an arbitrary threshold and

is the corresponding observed probability (either zero or one). For each forecast, the CRPS hence measures the area between the cumulative distribution of the forecasts and the observations. The CRPS is thus sensitive to both calibration and sharpness of the forecast, where reliable and sharper forecasts result in lower CRPS on average. The corresponding skill score then compares the CRPS of the forecasting system with the CRPS of the reference. The continuous ranked probability skill score (CRPSS) is thus defined as

CRPSS ranges from minus infinity to 1 with CRPSS larger than zero indicating that the forecast performs better (in CRPS terms) than the reference and a CRPSS of one resulting for a perfect deterministic forecast. Throughout the manuscript we use different forecasts as the reference, if the reference forecast is not specified in particular, however, we will use a climatological forecast as the reference forecast. The climatological forecast is a probabilistic forecast for which each of the past years from the verifying observations is treated as equally likely for each forecast.

In addition to the CRPS and its skill score, we also use the spread to error ratio to assess the calibration of the ensemble forecasts. Ensemble forecasts are reliable if the mean squared error (of the ensemble mean) is identical to the average intra-ensemble variance (A. P. Weigel, Liniger, and Appenzeller 2009). In addition, we complement the analysis with verification metrics for deterministic forecasts including the correlation and the root mean square difference between the ensemble mean CII and the verifying observations.

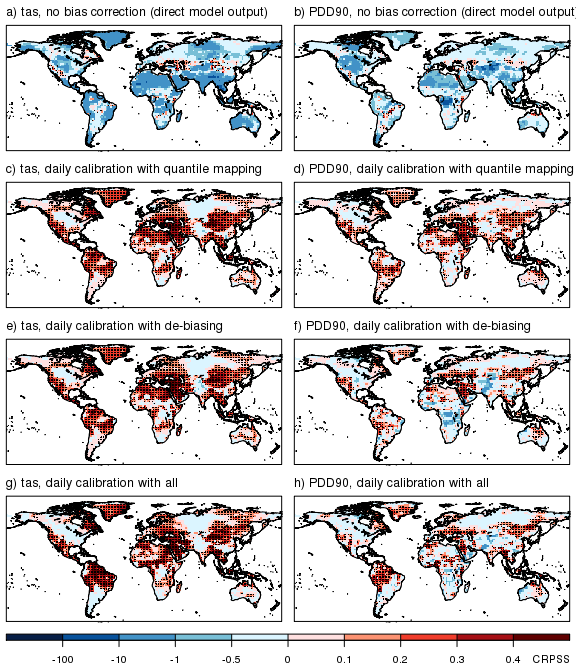
## Calibration and Verification Strategy

To avoid in-sample verification, we adopt a realistic forward calibration for which the calibration for forecast year only includes information from previous years (i.e. ). Using this forward calibration strategy, the calibration of the first years of the forecasts would be strongly influenced by the small sample of re-forecasts to estimate the calibration. Therefore we only calibrate years 18 to 34 in forward mode and adopt a backward calibration for the years 1 to 17 for which only years following the forecast are used for calibration. Such a verification strategy results in considerably lower skill compared to in-sample calibration (reflecting potential skill) and also results in slightly lower skill compared to calibration using cross-validation due to the reduced sample size for calibration. In contrast to a cross-validation approach, the forward calibration strategy adopted here allows us to assess the effect of a calibration of the time trend in the bias in a predictive, out-of-sample context.

# Results

## The Effect of Daily Calibration

To illustrate the effect of daily calibration, we show the CRPSS of seasonal mean temperature (tas) and the accumulated excceedance of the daily 90th percentile (PDD90) in Figure @ref(fig:daily-calibration-spatial). Skill of seasonal mean temperature vastly improves when the daily temperatures are calibrated (Figure @ref(fig:daily-calibration-spatial)c,e,g compared to a). The choice of daily calibration method is of minor importance: quantile mapping (Figure @ref(fig:daily-calibration-spatial)c) and the regression based calibration methods (de-biasing and all, Figure @ref(fig:daily-calibration-spatial)e,g) result in comparable CRPSS for seasonal mean temperature.

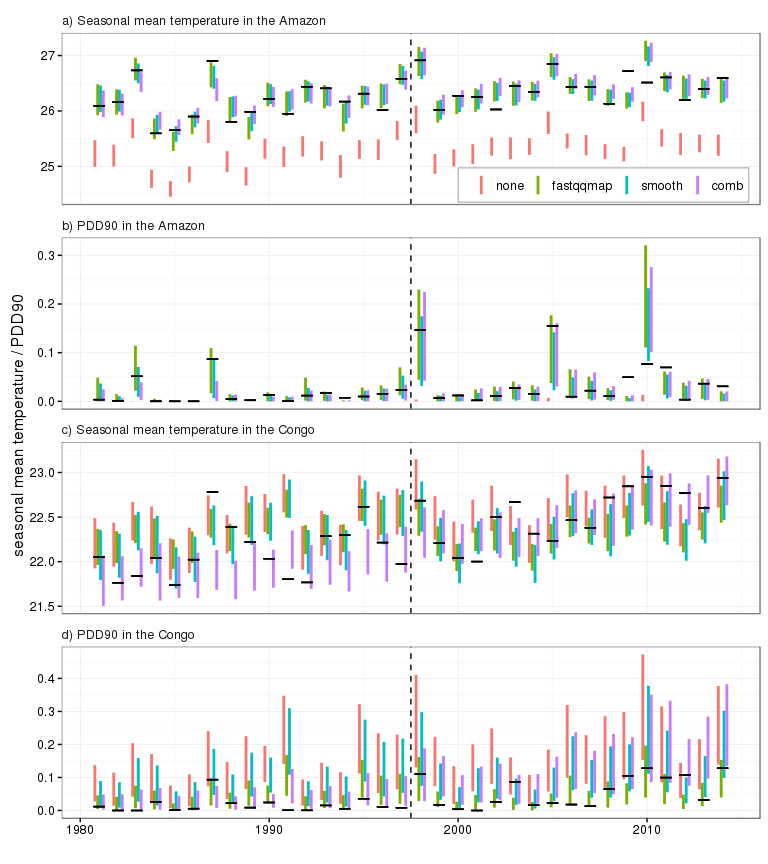


CRPSS of PDD90 in summer (JJA) from forecasts initialized in May derived from direct model output (a), and forecasts bias corrected with daily quantile mapping (b), mean de-biasing (c), and the all method (d, see text).

Daily calibration also improves the predictive skill (CRPSS) of forecasts of PDD90 (Figure @ref(fig:daily-calibration-spatial)d,f,h compared to b). In contrast to seasonal mean temperature, the choice of daily calibration method matters for skill of forecasts of PDD90. CRPSS of PDD90 forecasts calibrated using quantile mapping (Figure @ref(fig:daily-calibration-spatial)d) is generally higher than that of regression-based daily calibration methods (Figure @ref(fig:daily-calibration-spatial)f,h). Of the regression-based daily calibration methods, the more comprehensive method (all, Figure @ref(fig:daily-calibration-spatial)h) leads to slightly higher CRPSS than simple mean de-biasing (Figure @ref(fig:daily-calibration-spatial)f).

Also, we note that CRPSS for PDD90 is generally higher in regions that show positive CRPSS for seasonal mean temperature. For quantile mapping, the spatial pattern of CRPSS for seasonal mean temperature corresponds very well with that for PDD90 (Figure @ref(fig:daily-calibration-spatial)c,d). In most regions, CRPSS of PDD90 is comparable to the CRPSS of seasonal mean temperature. Elsewhere CRPSS is close to zero indicating that the calibrated PDD90 forecasts are as accurate a constant climatological forecast derived from the observations. For the regression-based calibration methods (Figure @ref(fig:daily-calibration-spatial)f,h), there are multiple regions where calibration of the ensemble mean appears to be insufficient and consequently the CRPSS of PDD90 is negative.

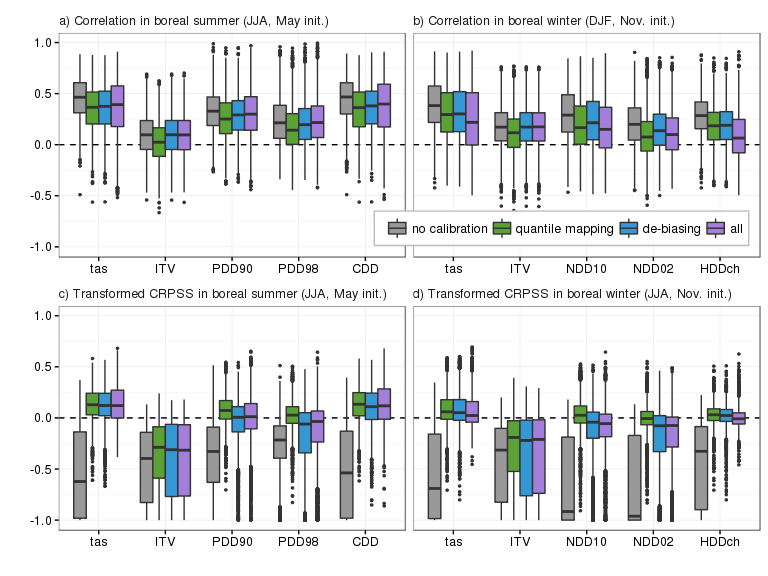
In Figure @ref(fig:example), we show the verifying observations and forecast time series for two locations in the tropics that illustrate some of the challenges when calibrating seasonal forecasts. The forecasts underestimate boreal summer (JJA) seasonal mean temperature in the Amazon at 6N, -56W by about 1 C (Figure @ref(fig:example)a) and consequently strongly underestimate PDD90 (Figure @ref(fig:example)b). All the daily bias correction methods reduce the mean bias and lead to significant improvements in CRPSS of seasonal mean temperature and PDD90 (see also Figure @ref(fig:daily-calibration-spatial)). While CRPSS of PDD90 with quantile mapping is the highest, the differences between the three bias correction methods are not significant.



Time series of JJA seasonal mean temperature and PDD90 forecasts from direct model output (red), and calibrated with quantile mapping (green), de-biasing (turquoise) and the all method (purple) for two points in the Amazon and the Congo respectively. The vertical bars show the 10th to the 90th percentile range of the 51-member forecasts, the horizontal black lines are the verifying observations (ERA Interim).

In contrast, the mean bias in JJA mean temperature in the Congo at 2N, 12W is small (Figure @ref(fig:example)c). The forecasts, however, considerably underestimate the warming trend over the past 34 years in that area. Consequently, both quantile mapping and mean de-biasing only marginally improve forecast quality (as measured by the CRPSS) compared to the direct model output. When a linear time trend in the bias is taken into account, the forecast skill is greatly improved (see also Figure @ref(fig:daily-calibration-spatial)h). The uncalibrated forecasts not only underestimate the time trend in seasonal mean temperature, they also overestimate daily variability (not shown). PDD90 is therefore overestimated. For PDD90, quantile mapping leads to the largest improvements in forecast skill, whereas PDD90 from calibrated daily temperature series using the mean de-biasing and all method is still generally overestimated due to the overestimation of daily variability in the uncalibrated forecasts.

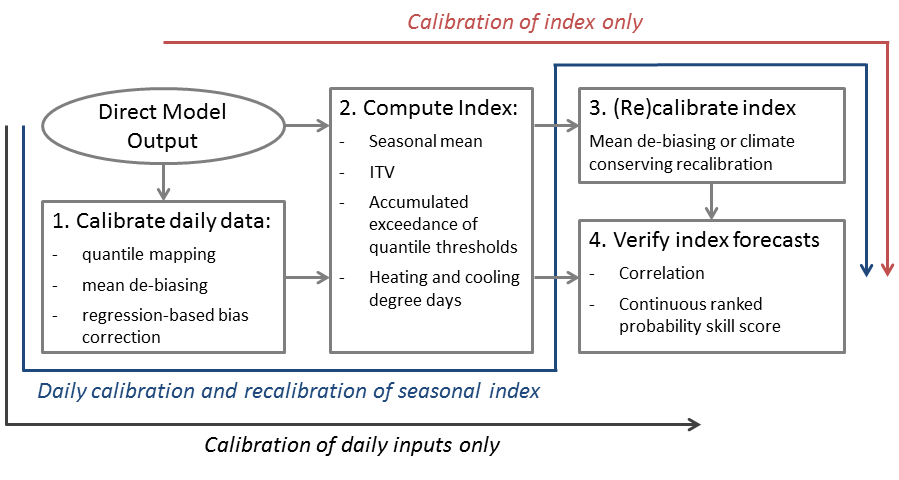
To get a more comprehensive overview of the advantages and disadvantages of the different daily calibration methods, we present the area-weighted distribution of grid point correlation and CRPSS in Figure @ref(fig:daily-calibration). Correlation of seasonal mean temperatures is generally reduced with daily calibration due to the out-of-sample calibration protocol and the corresponding information leakage (Storch and Zwiers 1999; Gangstø et al. 2013). The positive effect of the daily calibration compensates in some cases for the negative effect of the information leakage (e.g. for the all method in summer, purple box and whiskers in Figure @ref(fig:daily-calibration)a). For the various indices, this effect is even more pronounced. Whereas on average, correlation of forecasts calibrated using quantile mapping (green bars in Figure @ref(fig:daily-calibration)a,b) is always lower than the correlation of the forecasts from the uncorrected forecast (grey bars), in some cases, such as CDD and PDD90 in JJA for the all method (purple bars), the regression-based daily calibration methods compensate for the loss of correlation from the out-of-sample protocol.



Distribution of correlation (top) and CRPSS (bottom) for CII derived from seasonal forecasts for summer (JJA, initialized in May, left panels) and winter (DJF, initialized in Nov., right panels).

The continuous ranked probability skill score is negative for all indices at most grid boxes without daily calibration (grey box and whiskers in Figure @ref(fig:daily-calibration)c,d). All daily calibration lead to an improvement in CRPSS. For seasonal mean temperature, the improvements are largest for the all method in summer (JJA, Figure @ref(fig:daily-calibration)c) and the quantile mapping in winter (DJF, Figure @ref(fig:daily-calibration)d). For ITV and the degree day indices, daily calibration with quantile mapping generally leads to the most accurate forecasts as measured by CRPSS (green box and whiskers in Figure @ref(fig:daily-calibration)c,d). The difference between the daily calibration methods is most pronounced for the accumulated exceedances of percentile-based thresholds (i.e. PDD90/98 and NDD10/02).

## Daily Calibration Versus Seasonal Calibration of the Index Forecast

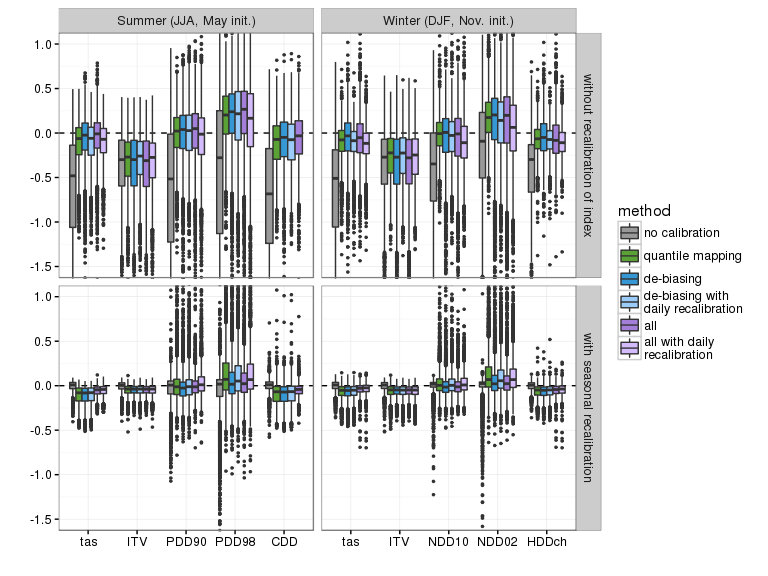


Flowchart of possible (re)calibration and verification strategies

So far, we have only presented forecasts derived from (un)calibrated daily temperature time series. Next, we assess the effect of an additional (re)calibration on the time series of seasonal indices. Such a recalibration may be necessary to correct biases in the CII time series that are not corrected for by the daily recalibration. Also, the re-calibration specifically corrects for under- or over-confidence in the forecasts by adjusting the ensemble spread of the target quantity (here the index).

In terms of root mean square error (Figure 2) or mean CRPS (Fig. XX), we find that the re-calibration does not reduce the RMSE or mean CRPS much beyond what daily calibration achieves. However, we also find that in RMSE and mean CRPS terms, re-calibration on the seasonal index alone is enough as the distribution of re-calibrated seasonal indices is largely independent of the daily calibration method used previous to computing seasonal indices and previous to re-calibrating these. This is somewhat surprising as one may expect threshold based indices to be strongly dependent on systematic errors (that are largely removed by the daily calibration). In contrast to such expectations, we find that systematic biases affecting the daily temperature series do not affect the computation of seasonal CII in a way that cannot be corrected for using a re-calibration on the seasonal CII.

With regard to correlation with the ensemble mean, on the other hand, we find that re-calibration of the seasonal CII degenerates the correlation in general. As for the daily calibration, the reduction in correlation is stronger for the more sophisticated daily calibration methods and stronger still, if these are further re-calibrated. This is an effect of using an out-of-sample approach for calibration which affects correlation negatively in a similar manner as discussed in Gangstø et al. (2013) for leave-one-out cross-validation.



Distribution of the spread to error ratio for CII derived from seasonal forecasts for summer (JJA, initialized in May, left panels) and winter (DJF, initialized in Nov., right panels). Colours denote the daily (re)calibration approach, the seasonal indices in the lower panels have been recalibrated in addition to the daily calibration of the input parameters.

The main advantage of a re-calibration on the seasonal CII, however, is the fact that re-calibration corrects the reliability of the forecasts. This is illustrated in Fig. XX using the spread to error ratio of the calibrated forecasts. When no re-calibration is applied, the calibrated forecasts are slightly over-dispersive (leftmost panels in Fig. XX) due to the reduction in systematic error without adjustment of the ensemble spread. An additional re-calibration (rightmost panels in Fig. XX), adjust also the forecast spread. This adjustment, however, tends to be too strong resulting in over-confident forecasts on average. This tendency to over-confident forecasts is again a consequence of the out-of-sample calibration approach. As shown in Tippett, DelSole, and Barnston (2014), estimation uncertainty in the parameters of regression-corrected forecasts leads to a positive bias in signal variance which in turn results in systematically over-confident forecasts. As the climate conserving re-calibration of A. P. Weigel, Liniger, and Appenzeller (2009) can be framed as a linear regression, the same systematic tendency to produce over-confident forecasts also applies for the CCR.

## Calibration strategies

In Figure 1, we illustrate the effect of different calibration strategies using seasonal mean temperature forecasts for land grid boxes for two regression-based calibration methods, namely the simple debias and the sophisticated all method (without recalibration of the daily data), and quantile mapping. For both calibration methods, we find that the in-sample calibration leads to the largest reduction in RMSE and to the highest improvement in CRPSS. Using the calibration methods in cross-validation or forward out-of-sample mode leads to slightly lower reductions in RMSE and increases in CRPSS compared with the in-sample calibration. CRPSS is lowest for the 10-year block cross-validation, and highest (among the out-of-sample approaches) for the leave-one-out cross-validation. It is important to note, that we do not adopt the corresponding out-of-sample protocol for the climatological reference forecast. Thereby, the CRPS of the reference forecast is biased low and thus the skill of the actual forecast is biased low. If one were to adopt the out-of-sample protocol also for the reference forecast, then generally the CRPSS improves beyond what is found for the in-sample calibrated forecast, as the negative bias in CRPS is stronger for the reference forecast than the actual forecast due to the smaller sample size of the reference (not shown). To facilitate comparison, however, we do not adopt an out-of-sample protocol for the reference, but note instead that our estimates of predictive skill are conservative and actual predictive skill is likely higher.

In contrast, correlation is unaffected when applying simple calibration methods in in-sample mode, but improvements in correlation may be found when more complex calibration methods (e.g. all in Figure 1) are used. Correlation, on the other hand, is reduced compared to using no calibration when calibration is applied in out-of-sample mode. This effect is strongest for the 10-year block cross-validation and weakest for the leave-one-out cross-validation with the forward calibration falling in between the two. The degradation of correlation skill with out-of-sample calibration is well known (H. M. van den Dool 1987) and is related to the leakage of information from the training data to the verification sample (Storch and Zwiers 1999).

# Conclusions

# Appendix: Climate conserving recalibration in the framework of linear regression

Here we demonstrate that the climate conserving recalibration (CCR) introduced by Weigel et al. (2009) can be formulated in a linear regression framework. To discuss this relationship we follow the notation used in Section 3, but drop reference to lead times and assume that both forecasts and observations are centered about zero. Thus the matrix contains the n forecast anomalies from m ensemble members, and the vector contains the verifying observation anomalies. Furthermore, denotes the series of ensemble mean forecasts and the ensemble mean forecast anomalies. The CCR is thus expressed using the adjustment parameter for the ensemble mean signal, and parameter for the intra ensemble spread

and

Where denotes the correlation between the ensemble mean forecast and the observations, the standard deviation of the observations, the standard deviation of the ensemble mean, and the square root of the time mean intra ensemble variance.

The adjustment parameter for the mean forecast, , is equivalent to the slope of the regression line of the observations on the ensemble mean forecast in an ordinary least squares framework, .

The adjustment on the intra ensemble spread, on the other hand, is equivalent to the ratio of the standard deviation of the residuals of the regression, , to the mean intra ensemble spread, .

Thus, the climate conserving recalibration can be rephrased as a prediction using ordinary least squares regression. In that respect, CCR is closely related to other MOS approaches (e.g. Gneiting et al. 2005). From this linear regression perspective on CCR follows a straightforward adjustment of the spread correction s to account for uncertainty in the signal calibration r. Following linear regression theory, the prediction interval scales with where is the forecast ensemble mean anomaly of the forecast the spread of which is to be corrected. Such an additional inflation has the property of increasing the spread for large forecast anomalies, that is for cases in which the signal calibration and its uncertainty have a large effect.

# References

Allen, M. R., and P. a. Stott. 2003. “Estimating Signal Amplitudes in Optimal Fingerprinting, Part I: theory.” *Climate Dynamics* 21 (5-6): 477–91. doi:[10.1007/s00382-003-0313-9](http://dx.doi.org/10.1007/s00382-003-0313-9). <http://link.springer.com/10.1007/s00382-003-0313-9>.

Barnston, Anthony G, and Huug M van den Dool. 1993. “A Degeneracy in Cross-Validated Skill in Regression-Based Forecasts.” *Journal of Climate* 6 (5). American Meteorological Society: 963–77. doi:[10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2). <http://dx.doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO 2>.

Cattiaux, J., H. Douville, R. Schoetter, S. Parey, and P. Yiou. 2015. “Projected Increase in Diurnal and Interdiurnal Variations of European Summer Temperatures.” *Geophysical Research Letters*, February, n/a–/a. doi:[10.1002/2014GL062531](http://dx.doi.org/10.1002/2014GL062531). <http://doi.wiley.com/10.1002/2014GL062531>.

Christenson, M., H. Manz, and D. Gyalistras. 2006. “Climate Warming Impact on Degree-Days and Building Energy Demand in Switzerland.” *Energy Conversion and Management* 47 (6): 671–86. doi:[10.1016/j.enconman.2005.06.009](http://dx.doi.org/10.1016/j.enconman.2005.06.009). <http://linkinghub.elsevier.com/retrieve/pii/S0196890405001603>.

Dee, D. P., S. M. Uppala, a. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, et al. 2011. “The ERA-Interim Reanalysis: configuration and Performance of the Data Assimilation System.” *Quarterly Journal of the Royal Meteorological Society* 137 (656): 553–97. doi:[10.1002/qj.828](http://dx.doi.org/10.1002/qj.828). <http://doi.wiley.com/10.1002/qj.828>.

Dool, H M van den. 1987. “A Bias in Skill in Forecasts Based on Analogues and Antilogues.” *Journal of Climate and Applied Meteorology* 26 (9). American Meteorological Society: 1278–81. doi:[10.1175/1520-0450(1987)026<1278:ABISIF>2.0.CO;2](http://dx.doi.org/10.1175/1520-0450(1987)026<1278:ABISIF>2.0.CO;2). <http://dx.doi.org/10.1175/1520-0450(1987)026<1278:ABISIF>2.0.CO 2>.

Gangstø, R, Ap Weigel, Ma Liniger, and C Appenzeller. 2013. “Methodological Aspects of the Validation of Decadal Predictions.” *Climate Research* 55 (3): 181–200. doi:[10.3354/cr01135](http://dx.doi.org/10.3354/cr01135). <http://www.int-res.com/abstracts/cr/v55/n3/p181-200/>.

Gneiting, Tilmann, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. 2005. “Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation.” *Monthly Weather Review* 133 (5): 1098–1118. doi:[10.1175/MWR2904.1](http://dx.doi.org/10.1175/MWR2904.1). <http://journals.ametsoc.org/doi/abs/10.1175/MWR2906.1 http://journals.ametsoc.org/doi/abs/10.1175/MWR2904.1>.

Hersbach, Hans. 2000. “Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems.” *Weather and Forecasting* 15 (5): 559–70. doi:[10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2](http://dx.doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2). <http://journals.ametsoc.org/doi/abs/10.1175/1520-0434{\%}282000{\%}29015{\%}3C0559{\%}3ADOTCRP{\%}3E2.0.CO{\%}3B2>.

Mahlstein, Irina, Christoph Spirig, Mark a. Liniger, and Christof Appenzeller. 2015. “Estimating Daily Climatologies for Climate Indices Derived from Climate Model Data and Observations.” *Journal of Geophysical Research: Atmospheres* 120 (7): 2808–18. doi:[10.1002/2014JD022327](http://dx.doi.org/10.1002/2014JD022327). <http://doi.wiley.com/10.1002/2014JD022327>.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, and T. Palmer. 2011. “The New ECMWF Seasonal Forecast System (System 4).”

Panofsky, H. A., and G. W. Brier. 1968. *Some Applications of Statistics to Meteorology*. University Park: The Pennsylvania State University.

Schlenker, Wolfram, W. Michael Hanemann, and Anthony C. Fisher. 2007. “Water Availability, Degree Days, and the Potential Impact of Climate Change on Irrigated Agriculture in California.” *Climatic Change* 81 (1): 19–38. doi:[10.1007/s10584-005-9008-z](http://dx.doi.org/10.1007/s10584-005-9008-z). <http://link.springer.com/10.1007/s10584-005-9008-z>.

Siegert, Stefan. 2015. “SpecsVerification: Forecast Verification Routines for the SPECS FP7 Project.” <http://cran.r-project.org/package=SpecsVerification>.

Storch, Hans von, and Francis W. Zwiers. 1999. *Statistical Analysis in Climate Research*. Cambridge University Press.

Tippett, Michael K., Timothy DelSole, and Anthony G. Barnston. 2014. “Reliability of Regression-Corrected Climate Forecasts.” *Journal of Climate* 27 (9): 3393–3404. doi:[10.1175/JCLI-D-13-00565.1](http://dx.doi.org/10.1175/JCLI-D-13-00565.1). <http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-13-00565.1>.

Weigel, Andreas P., Mark a. Liniger, and Christof Appenzeller. 2009. “Seasonal Ensemble Forecasts: Are Recalibrated Single Models Better Than Multimodels?” *Monthly Weather Review* 137 (4): 1460–79. doi:[10.1175/2008MWR2773.1](http://dx.doi.org/10.1175/2008MWR2773.1). <http://journals.ametsoc.org/doi/abs/10.1175/2008MWR2773.1>.

Wilcke, Renate Anna Irma, Thomas Mendlik, and Andreas Gobiet. 2013. “Multi-Variable Error Correction of Regional Climate Models.” *Climatic Change* 120 (4): 871–87. doi:[10.1007/s10584-013-0845-x](http://dx.doi.org/10.1007/s10584-013-0845-x). <http://link.springer.com/10.1007/s10584-013-0845-x>.