

## Answer to referee 1

Thank you very much for your informative and detailed comments.

### General comments

"This study proposes a probabilistic forecast recalibration scheme designed specifically to address the issues associated with decadal climate forecasts. Polynomial forms are estimated for the dependence of unconditional bias, conditional bias and ensemble spread on lead-time. The parameters of the polynomial forms are allowed to depend linearly on forecast-time, in order to account for any discrepancy between the long-term trends in the observations and the forecasts. The resulting recalibration is shown to outperform raw forecasts of surface temperature, and forecasts corrected only for lead-time dependent unconditional bias. The extension of lead-time dependent bias correction to the conditional bias and ensemble spread of parametric probability forecasts represents a valuable contribution to the decadal forecasting literature. Overall, the paper is clear and well written. The conclusions regarding the performance of the proposed recalibration are broadly justified by the results shown."

### Specific comments

1. "The review of existing recalibration methods is rather brief and would benefit from expansion, although it does cover the most relevant references. Other closely related methods for seasonal-to-decadal forecast recalibration include Eade et al. (2014, doi:10.1002/2014GL061146), Sansom et al. (2016, doi:10.1175/JCLI-D-15-0868.1) and references therein."

**Answer:** Thank you for that advice. An extended review of existing (decadal) recalibration approaches will be added to the manuscript.

2. "The DeFoReSt method addresses lead-time dependent unconditional and conditional biases in the ensemble mean, and unconditional bias in the ensemble spread. The authors cite the study by Fučkar et al. (2014) as a method specifically tailored to decadal forecasts. That study addressed corrections based on observed conditions at the time of forecast initialisation. Can the authors comment on the relevance of such corrections and why they chose to prioritise the biases addressed in DeFoReSt?"

**Answer:** The approach of Fučkar et al. (2014) is highly relevant as the model bias and forecast skill depend on the initial conditions. This approach seems to outperform the bias correction approach based on a linear start year dependency. But the success of this approach may be dependent on the analyzed variable and region. Thus, in this first description of DeFoReSt we have decided for the less sophisticated approach proposed by Kharin et al. (2012) and Kruschke et al. (2015) which assumes a linear start year dependency. However, combining the approach of Fučkar et al. (2014) with our approach is definitely an interesting idea for further studies.

-Fučkar, N. S., Volpi, D., Guemas, V., and Doblas-Reyes, F. J.: A posteriori adjustment of near-term climate predictions: Accounting for the drift dependence on the initial conditions, *Geophysical Research Letters*, 41, 5200–5207, 2014.

-Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F., and Lee, W.-S.: Statistical adjustment of decadal predictions in a changing climate, *Geophysical Research Letters*, 39, 2012.

-Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U.: Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, *Meteor. Z.*, 01, 2015

3. "The orders of the polynomial forms used to capture lead-time dependence are fixed a priori. The authors acknowledge the need for a systematic method of selecting the order of the polynomial forms. However, it would be useful if they could provide some insight or justification for the choices they made?"

**Answer:** For the proposed first version of DeFoReSt we follow the suggestion of Gangstø et al. (2013) and use a third order polynomial addressing the unconditional and conditional bias and a second order addressing the ensemble dispersion. With an increasing order of the polynomial the flexibility of the fitted curve increases, while the parameter uncertainty also increases. Here, Gangstø et al. (2013) suggested that a third order polynomial is a good compromise

between flexibility and parameter uncertainty. For the correction of the ensemble dispersion we assumed that a higher flexibility may not be necessary, because the MSE – which influences the dispersion – is already addressed by a third order polynomial of unconditional and conditional bias. Without any systematic model selection approach (which is to be developed) an answer to this question cannot be made more objective.

-Gangstø, R., A.P. Weigel, M.A. Liniger, C. Appenzeller, 2013: Methodological aspects of the validation of decadal predictions. – Climate Res. 55, 181–200, DOI: 10.3354/cr01135.

4. "A bootstrapping method is used to assess the uncertainty in the skill of the recalibrated forecasts. Can the authors clarify how the bootstrapping was performed?"

**Answer:** The scores have been calculated for a period from 1961 to 2005. For bootstrapping we draw a new pair of dummy time series with replacement from the original validation period and calculate these scores again. Here, this procedure has been repeated 1000 times. An explanation will be added in the manuscript.

5. "In Section 5.2, the authors state that “After applying DeFoReSt, [...] the ensemble spread is fairly constant for all lead times”. This statement is broadly supported by the results in Figures 8 and 9, but runs contrary to the intuition that our uncertainty about the future climate should increase with lead-time. Can the authors comment on this surprising result?"

**Answer:** Maybe, the declaration "fairly constant" is misleading and s.th. like "basically constant" is more appropriate. In Fig. 5b and 7b the ensemble spread appears to be constant for all lead years, but Fig. 6c and 8c show that the ensemble spread of the recalibrated forecast increases from lead year 1 to lead year 10. We admit that the ensemble spread of lead years 3-7 or 3-5 show an unexpected behavior by decreasing with lead years. Here, the ESS shows that the ensemble spread of the recalibrated forecast should be higher.

## 20 Technical corrections

1. Several examples: choose  $\text{Var}(x)$  or  $\text{var}(x)$ , there are examples of both

**Answer:** Will be corrected

2. Several examples:  $E[x] \rightarrow E(x)$  for consistency with  $\text{var}(x)$  and general bracket conventions

**Answer:** Will be corrected

3. Several examples: e.g.  $\rightarrow$  e.g.,

**Answer:** Will be corrected

4. Several examples: i.e.  $\rightarrow$  i.e.,

**Answer:** Will be corrected

5. Page 12, Lines 4 15: yields to

**Answer:** Will be corrected

6. Page 10, Line 26; Page 12, Line 6; Page 13, Line 23: yielding to

**Answer:** Will be corrected

7. Page 2, Line 31: verification calibration  $\rightarrow$  verification and calibration

**Answer:** Will be corrected

8. Page 3, Line 23: verification  $\rightarrow$  observations

**Answer:** Will be corrected

9. Page 3, Line 26: calibration → recalibration  
**Answer:** Will be corrected
10. Page 6, Line 11: the approach (9) → the second term in (9)  
**Answer:** Will be corrected
- 5 11. Page 6, Line 13: It might be worth explaining in words that the dependence on lead-time is cubic, while the correction for errors in time trends is only linear.  
**Answer:** A short explanation will be given
12. Page 6, Line 19: Additionally, → In addition  
**Answer:** Will be corrected
- 10 13. Page 6, Line 26: These assumption → These assumptions  
**Answer:** Will be corrected
14. Page 6, Line 27: order selection will be topic of → order selection will be a topic of  
**Answer:** Will be corrected
15. Page 7, Line 13: ensured → guaranteed  
15 **Answer:** Will be corrected
16. Page 7, Line 21: this observations → these observations  
**Answer:** Will be corrected
17. Page 8, Line 04:  $\mu_x$  and  $\varepsilon_x$  → The processes  $\mu_x$  and  $\varepsilon_x$   
**Answer:** Will be corrected
- 20 18. Page 8, Line 13: concrete → specific or exact  
**Answer:** Will be corrected
19. Page 8, Line 13: this variability → state which variability exactly  
**Answer:** Will be corrected
20. Page 8, Line 14: As for the recalibration strategy → For recalibration,  
25 **Answer:** Will be corrected
21. Page 8, Line 15: we use... → This sentence doesn't make sense.  
**Answer:** We will revise that sentence in the manuscript.
22. Page 8, Line 21: In general, ensemble mean and ensemble variance both can dependent on lead time  $\tau$  and start time  $t$ .  
→ In general, the ensemble mean and variance can both depend on lead time  $\tau$  and start time  $t$ .  
30 **Answer:** Will be corrected
23. Equation 21: It would be helpful to explain the motivation for non-linear form chosen here and the restriction in line 26.  
**Answer:** The motivation is that both conditional bias and  $\mu_x$  generally can depend on lead time and start time. The restriction is necessary to avoid negative values of the ensemble variance in Eq. 22. An explanation will be given in the manuscript.

24. Page 9, Lines 4-8: I understand from this paragraph that a trend is included in only the observations or the forecasts, but I am not clear on which. This needs to be clarified and possible linked explicitly to Eqn. 17 and Page 8, Lines 13-15.  
**Answer:** A linear trend will be imposed on the pseudo-forecast due to the unconditional bias. Indeed, this needs to be clarified in the manuscript.
- 5 25. Page 9, Line 26: 10 lead years  $\tau = 1, \dots, 10 \rightarrow 10$  lead years ( $\tau = 1, \dots, 10$ )  
**Answer:** Will be corrected
26. Page 10, Line 24: which ESS values are lower one  $\rightarrow$  whose ESS values are less than one  
**Answer:** Will be corrected
27. Page 10, Line 25: The reduced performance...  $\rightarrow$  The lower performance  
10 **Answer:** Will be corrected
28. Page 11, Lines 25: model data with a low potential predictability  
**Answer:** Will be corrected
29. Page 12, Line 01: the worse MSE performance  $\rightarrow$  the lower MSE performance  
**Answer:** Will be corrected
- 15 30. Page 12, Line 03: On the contrary  $\rightarrow$  In contrast  
**Answer:** Will be corrected
31. Page 12, Line 08: bias within the  $\rightarrow$  bias in the  
**Answer:** Will be corrected
32. Page 12, Line 09: w.r.t.  $\rightarrow$  compared to  
20 **Answer:** Will be corrected
33. Page 12, Line 28: applied on surface  $\rightarrow$  applied to surface  
**Answer:** Will be corrected
34. Page 12, Line 27: global mean and a spatial mean  
**Answer:** Will be corrected
- 25 35. Page 13, Line 11: On the contrary, the raw model's  $\rightarrow$  The raw model's  
**Answer:** Will be corrected
36. Page 13, Line 26: is small than  $\rightarrow$  is less than  
**Answer:** Will be corrected
37. Page 13, Line 27: DeFoReSt slightly performs better  $\rightarrow$  DeFoReSt performs slightly better  
30 **Answer:** Will be corrected
38. Page 13, Line 32: with the corresponding  $\rightarrow$  and the corresponding  
**Answer:** Will be corrected
39. Page 14, Line 10: is constant  $\rightarrow$  is almost constant  
**Answer:** Will be corrected

40. Page 14, Line 13: perform equally → perform similarly

**Answer:** Will be corrected

41. Page 14, Line 20: also outperforming → also outperform

**Answer:** Will be corrected

5 42. Page 14, Line 21: climatology is solely not significant → climatology is not significant

**Answer:** Will be corrected

43. Page 15, Line 05: Analog to → Following

**Answer:** Will be corrected

44. Page 15, Line 13: impose → imposed

10 **Answer:** Will be corrected

45. Page 15, Line 14: conditional bias and ensemble spread dispersion→ conditional bias or ensemble spread

**Answer:** Will be corrected

## **Answer to referee 2**

Thank you very much for your informative and detailed comments.

## 15 **General comments**

"This paper presents what may be a very important work in the study of decadal climate prediction. The authors presented the methodologically oriented post-processing model "Parametric Decadal Climate Forecast Recalibration (DeFoReSt)," to correct decadal climate prediction. The method uses earlier published approaches, and extend it to the long-term prediction by allowing the forecast errors to depend on forecast lead time. The performance of the presented approach is well established using different assessment measures."

## **Specific comments**

1. "It is mentioned in section 3.3 on lines 1-2 that the parameters are estimated by minimizing the average CRPS over the training period. Does this mean only the portion (the training) of the data was used for estimating the parameters? if yes, why the whole data was not used? it is expected that a training-validation grouping of data for checking the performance of DeFoReSt. But a parameter within DeFoReSt has to be primarily estimated from the whole data."

25 **Answer:** In case of a validation/comparison with a reference data set (e.g., climatology or raw model) the training data set is only a portion of the whole available data set, while the remaining data is used for validation. We aim at estimating a forecast error for a setting comparable to the operational forecast situation where no observations for the forecast period is available. DeFoReSt parameters can only be estimated from the available observataion period but the re-calibration is carried out on the forecast period, i.e. outside the period used for parameter estimation. Hence using the full hindcast period for estimating parameters and obtain a "forecast error" for hindcast from the same period would lead to overestimation of skill. Parameter estimation using the full available data set could be used once we use DeFoReSt for re-calibration decadal forecasts, for e.g. 2018-2027.

35 2. "DeFoReSt was defined based on ensemble mean and variance functions (by my reading on the paper), where 15 ensemble members were selected. However, systematic errors vary widely between ensemble members, a simple ensemble averaging limits the relevance of DeFoReSt for long-term prediction. Arisido et al (2017) and Tebaldi et al (2005), see below, demonstrated that the common ensemble averaging method where each ensemble member has the same weight

poses serious issues. I encourage the authors to discuss this issue supporting their argument with these papers and take into account the advisories in their revision.

- Arisido, M.W., Gaetan, C., Zanchettin, D. et al. Stoch Environ Res Risk Assess (2017). <https://doi.org/10.1007/s00477-017-1383-2> - Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. Journal of Climate 18:1524-1540"

**Answer:** It is true that the ensemble members of a multi-model ensemble cannot be treated equally because every corresponding model has different systematic errors. However, in this study we apply DeFoReSt to a single model ensemble with 15 members generated by lagged-day-initialization from MPI-ESM-LR; i.e., we do not expect that the single ensemble members have different systematic errors (due to the model). Nonetheless, for a recalibration of a multimodel ensemble DeFoReSt needs to be adapted. Which would be a topic of further research.

3. "How are the orders chosen for the polynomials used in equations such as (9) and (10). Without some cross-validation study it is not clear how a third order polynomial suffices for the drift along lead time. A sensitivity analysis for different order scenarios should guide for closer to the optimal choice needed for capturing the underlying features in a data."

**Answer:** We agree that there is need for a transparent model selection strategy! As already mentioned in section 6, this will be topic for future studies. For the first version of DeFoReSt we follow the suggestion of Gangstø et al. (2013) and use a third order polynomial addressing the unconditional and conditional bias and a second order addressing the ensemble dispersion. With an increasing order of the polynomial the flexibility of the fitted curve increases, while the parameter uncertainty also increases. Here, Gangstø et al. (2013) suggested that a third order polynomial is a good compromise between flexibility and parameter uncertainty. For the correction of the ensemble dispersion we assumed that a higher flexibility may not be necessary, because the MSE -which influences the dispersion- is addressed by a third order polynomial of unconditional and conditional bias.

-Gangstø, R., A.P. Weigel, M.A. Liniger, C. Appenzeller, 2013: Methodological aspects of the validation of decadal predictions. – Climate Res. 55, 181–200, DOI: 10.3354/cr01135.

## Technical corrections

1. "line 19 page 2, the acronym "PDF" should be defined on the first use"

**Answer:** Will be corrected

2. "line 1 Page 15, "...to change polynomially.." Remove "polynomially", then it is clear that the conditional bias and the ensemble dispersion change with lead time, while they change linearly with start time."

**Answer:** Will be corrected

3. " $v_5$  in (A9) page 17, the coefficient of  $\tau^2$ , is typos error?"

**Answer:** Indeed, it should be  $b_5$  instead of  $v_5$ .

## List of relevant changes made in the manuscript

1. Page 2, Lines 19 – 24: Added an extended review of existing recalibration approaches
2. Page 6, Lines 9 – 10: Added an explanation why we have decided for a third order polynomial approach.
3. Page 6, Lines 16 – 17: Added a statement for choosing an linear approach for start year dependency of the bias.
- 5 4. Page 7, Lines 2 – 3: Added an explanation for using a second order polynomial approach to account for ensemble dispersion.
5. Page 9, Lines 4 – 5: Motivation for the restriction of  $\sigma_{\epsilon_f}^2(t, \tau)$  added.
6. Page 9, Lines 12 – 13: Clarified that linear trend is included in the pseudo-forecasts.
7. Page 13, Lines 7 – 9: Added an detailed description of the applied bootstrapping approach.
- 10 8. Page 16, Lines 22 – 25: Added a statement that *DeFoReSt* needs to be modified if it should be applied to multimodel predictions.

Minor changes (e.g., typing errors, etc.) are not listed here.

# Parametric Decadal Climate Forecast Recalibration (DeFoReSt 1.0)

Alexander Pasternack<sup>1</sup>, Jonas Bhend<sup>2</sup>, Mark A. Liniger<sup>2</sup>, Henning W. Rust<sup>1</sup>, Wolfgang A. Müller<sup>3</sup>, and Uwe Ulbrich<sup>1</sup>

<sup>1</sup>Institute of Meteorology, Freie Universität Berlin, Berlin, Germany

<sup>2</sup>Federal Office of Meteorology and Climatology (MeteoSwiss), Zürich, Switzerland

<sup>3</sup>Max-Planck-Institute for Meteorology, Hamburg, Germany

Correspondence to: A. Pasternack (alexander.pasternack@met.fu-berlin.de)

**Abstract.** Near-term climate predictions such as decadal climate forecasts are increasingly being used to guide adaptation measures. For near-term probabilistic predictions to be useful, systematic errors of the forecasting systems have to be corrected. While methods for the calibration of probabilistic forecasts are readily available, these have to be adapted to the specifics of decadal climate forecasts including the long time horizon of decadal climate forecasts, lead time dependent systematic errors (drift), and the errors in the representation of long-term changes and variability. These features are compounded by small ensemble sizes to describe forecast uncertainty and a relatively short period for which typically pairs of re-forecasts and observations are available to estimate calibration parameters. We introduce the Decadal Climate Forecast Recalibration Strategy (*DeFoReSt*), a parametric approach to recalibrate decadal ensemble forecasts that takes the above specifics into account. *DeFoReSt* optimizes forecast quality as measured by the continuous ranked probability score (CRPS). Using a toy model to generate synthetic forecast observation pairs, we demonstrate the positive effect on forecast quality in situations with pronounced and limited predictability. Finally, we apply *DeFoReSt* to decadal surface temperature forecasts from the MiKlip Prototype system and find consistent and sometimes considerable improvements in forecast quality compared with a simple calibration of the lead time dependent systematic errors.

## 1 Introduction

Decadal climate predictions aim to characterize climatic conditions over the coming years. Recent advances in model development, data assimilation and climate observing systems together with the need for up-to-date and reliable information on near-term climate for adaptation planning have lead to considerable progress in decadal climate predictions. In this context, international and national projects like the German initiative Mittelfristige Klimaprognosen (MiKlip) have developed model systems to produce a skillful decadal climate prediction (Pohlmann et al., 2013a; Marotzke et al., 2016).

Despite the progress being made in decadal climate forecasting, such forecasts still suffer from considerable systematic biases. In particular, decadal climate forecasts are affected by lead time dependent biases (drift) and exhibit long-term trends that differ from the observed changes. To correct these biases in the expected mean climate, bias correction methods tailored to the specifics of decadal climate forecasts have been developed (Kharin et al., 2012; Fučkar et al., 2014; Kruschke et al., 2015).



Given the inherent uncertainties due to imperfectly known initial conditions and model errors, weather and climate predictions are framed probabilistically (Palmer et al., 2006). Such probabilistic forecasts are often affected by biases in forecast uncertainty (ensemble spread), i.e., they are not reliable. Forecasts are reliable if the forecast probability of a specific event equals the observed occurrence frequency on average (Palmer et al., 2008). Briefly said, if some event is declared with a certain probability, say 80%, it should also occur on average 80% of all times such a forecast is issued. Probabilistic forecasts, however, are often found to be underdispersive/overconfident (Hamill and Colucci, 1997; Eckel and Walters, 1998), i.e., the ensemble spread underestimates forecast uncertainty and events with a forecast probability of 80% occur on average less often.

Statistical post processing (Gneiting and Raftery, 2005) can be used to optimize – or recalibrate – the forecast, e.g., reducing systematic errors, such as bias and conditional bias, as well as adjusting ensemble spread. The goal of recalibrating probabilistic forecasts is to maximize sharpness without sacrificing reliability (Gneiting et al., 2003). A forecast is sharp if its distribution differs from the climatological distribution. For example, a constant climatological probability forecast is perfectly reliable, but exhibits small sharpness. Recalibration methods, have been developed for medium-range to seasonal forecasting; it is unclear to what extent lead time dependent biases (also called drift) and long-term trends of decadal climate forecasts can effectively be corrected. Here, we aim at adapting existing recalibration methods to deal with the specific problems found in decadal climate forecasting: lead time and start time dependent biases, conditional biases and inadequate ensemble spread.

The most prominent recalibration methods proposed in the context of medium-range weather forecasting are Bayesian model averaging (BMA, Raftery et al., 2005; SlUGHTER et al., 2007) and non-homogeneous Gaussian regression (NGR, Gneiting et al., 2005). In seasonal forecasting, the climate conserving recalibration (CCR, Doblas-Reyes et al., 2005; Weigel et al., 2009) is often applied, which is based on a scalar conditional adjustment of ensemble mean and ensemble spread. Here, Eade et al. (2014) applied this concept also to decadal predictions. BMA assigns a probability density function (PDF)PDF to every individual ensemble member and generates a weighted average of these densities where the weights represent the forecasting skill of the corresponding ensemble member. NGR extends traditional model output statistics (MOS, Glahn and Lowry, 1972)(MOS) by allowing the predictive uncertainty to depend on the ensemble spread. A further extension, proposed by Sansom et al. 2016, also accounts for a linear time-dependency of the mean bias. However, CCR is closely related to NGR in that the forecast mean error and forecast spread are jointly corrected to satisfy the necessary criterion for reliability that the time mean ensemble spread equals the forecast root mean square error.

We expand on NGR and CCR by introducing a parametric dependence of the forecast errors on forecast lead time and long-term time trends hereafter named decadal climate forecast recalibration strategy (*DeFoReSt*). To better understand the properties of *DeFoReSt*, we conduct experiments using a toy model to produce synthetic forecast observation pairs with known properties. We compare the decadal recalibration with the drift correction proposed by Kruschke et al. (2015) to illustrate its benefits and limitations.

The remainder of the paper is organized as follows. In Sec. 2 we introduce the MiKlip decadal climate prediction system and the corresponding reference data used. Moreover, we discuss how forecast quality of probabilistic forecasts is assessed. In Sec. 3, we motivate the extension of the NGR method named *DeFoReSt* and illustrate how verification and calibration can be linked by the way the calibration parameters are estimated. The toy model used to study *DeFoReSt* is introduced and assessed in

Sec. 4.2. In the following section, we apply the drift correction and *DeFoReSt* to decadal surface temperature predictions from the MiKlip system (Sec. 5). We assess global mean surface temperature and temperature over the North Atlantic subpolar gyre region (60°-10°W, 50°-65°N). The investigated North Atlantic region has been identified as a key region for decadal climate predictions with forecast skill for different parameters (e.g., Pohlmann et al., 2009; van Oldenborgh et al., 2010; Matei et al., 2012; Mueller et al., 2012). The paper closes with a discussion in Sec. 6.

## 2 Data and methods

### 2.1 Decadal climate forecasts

In this study we use retrospective forecasts (hereafter call hindcast) of surface temperature performed with the Max-Planck-Institute Earth System Model in a low-resolution configuration (MPI-ESM-LR). The atmospheric component of the coupled model is ECHAM6 run at a horizontal resolution of T63 with 47 vertical levels up to 0.1 hPa (Stevens et al., 2013). The ocean component is MPIOM with a nominal resolution of 1.58 and 40 vertical levels (Jungclaus et al., 2013).

We investigate one set of decadal hindcasts, namely from the MiKlip Prototype system, which consists 41 hindcasts, each with 15 ensemble members, yearly initialized at January 1<sup>st</sup> between 1961 and 2000 and then integrated for 10 years. The initialization of the atmospheric part was realized by full field initialization from fields of Era40 (Uppala et al., 2005) and Era-Interim (Dee et al., 2011), while the oceanic part was initialized with full fields from GECCO2 reanalysis (Köhl, 2015). Here, the full field initialization nudges the atmospheric or oceanic fields from the corresponding reanalysis to the MPI-ESM as full fields and not as anomalies. A detailed description of the Prototype system is given in Kröger et al. (2017, submitted).

### 2.2 Validation data

This study uses the 20th Century Reanalysis (20CR, Compo et al., 2011) for evaluation of the hindcasts. The reanalysis has been built by solely assimilating surface pressure observations, whereas the lower boundary forcing is given from HadISST1.1 sea surface temperatures and sea ice (Rayner et al., 2003). Moreover, 20CR is based on Ensemble-Kalman-filtering with 56 members and therefore also addresses observation and assimilation uncertainties. Additionally, 20CR covers the whole period of the investigated decadal hindcasts, which is a major benefit over other common reanalysis data sets.

### 2.3 Assessing reliability and sharpness

Calibration or reliability refers to the statistical consistency between the forecast PDFs and the verifying observations. Hence, it is a joint property of the predictions and the [observationsverification](#). A forecast is reliable if forecast probabilities correspond to observed frequencies on average. Alternatively, a necessary condition for forecasts to be reliable is given if the time mean intra-ensemble variance equals the mean squared error (MSE) between ensemble mean and observation (Palmer et al., 2006).

A common tool to evaluate the reliability and therefore the effect of a [recalibration](#) is the rank histogram or *Talagrand diagram* which were separately proposed by Anderson (1996); Talagrand et al. (1997); Hamill and Colucci (1997). For a

detailed understanding, the rank histogram has to be evaluated by visual inspection. Here, we have chosen to use the *Ensemble Spread Score* (ESS) as a summarizing measure. The ESS is the ratio between the time mean intra-ensemble variance  $\bar{\sigma}^2$  and the mean squared error between ensemble mean and observation,  $MSE(\mu, y)$  (Palmer et al., 2006; Keller and Hense, 2011):

$$ESS = \frac{\bar{\sigma}^2}{MSE(\mu, y)}, \quad (1)$$

5 with

$$\bar{\sigma}^2 = \frac{1}{k} \sum_{j=1}^k \sigma_j^2, \quad (2)$$

and

$$MSE(\mu, y) = \frac{1}{k} \sum_{j=1}^k (y_j - \mu_j)^2. \quad (3)$$

Here,  $\sigma_j^2$ ,  $\mu_j$  and  $y_j$  are the ensemble variance, the ensemble mean and the corresponding observation at time step  $j$ , with  
 10  $j = 1, \dots, k$ , where  $k$  is the number time steps.

Following Palmer et al. (2006),  $ESS = 1$  indicates perfect reliability. The forecast is overconfident when  $ESS < 1$ , i.e., the ensemble spread underestimates forecast error. If the ensemble spread is greater than the model error ( $ESS > 1$ ), the forecast is overdispersive and the forecast spread overestimates forecast error. To better understand the components of the  $ESS$ , we also analyze the mean squared error  $MSE$  of the forecast separately.

15 Sharpness, on the other hand, refers to the concentration or spread of a probabilistic forecast and is a property of the forecast only. A forecast is sharp, when it is taking a risk, i.e., when it is frequently different from the climatology. The smaller the forecast spread, the sharper the forecast. Sharpness is indicative of forecast performance for calibrated and thus reliable forecasts, as forecast uncertainty reduces with increasing sharpness (subject to calibration). To assess sharpness, we use properties of the width of prediction intervals as in Gneiting and Raftery (2007). In this study the time mean intra-ensemble  
 20 variance  $\bar{\sigma}^2$  is used to assess the prediction width.

Scoring rules, finally, assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, since they address reliability and sharpness simultaneously (Gneiting et al., 2005; Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014). These scores are generally taken as penalties, thus the forecasters seek to minimize them. A scoring rule is called proper, if its expected value is minimized when the observation is drawn from the same distribution as the  
 25 predictive distribution. If a scoring rule is not proper, it is possible to minimize its expected value by predicting an unrealistic probability of occurrence. In simple terms, a forecaster would be rewarded for not being honest. Moreover a proper scoring rule is called strictly proper if the minimum is unique. In this regard, the *Continuous Ranked Probability Score* (CRPS) is a suitable, strictly proper scoring rule for ensemble forecasts.

Given,  $F$  is the predictive cumulative distribution function (CDF) and  $o$  is the verifying observation, the CRPS is defined as

$$30 \quad CRPS(F, o) = \int_{-\infty}^{\infty} (F(y) - F_0(y))^2 dy, \quad (4)$$

where  $F_0(y)$  is the Heaviside function and takes the values 0 or 1 if  $y$  is less than or greater equal than the observed value  $o$ . Under the assumption that the predictive CDF is a normal distribution with mean  $\mu$  and variance  $\sigma^2$  Gneiting et al. (2005) showed that (4) can be written as

$$CRPS(\mathcal{N}(\mu, \sigma^2), o) = \sigma \left\{ \frac{o - \mu}{\sigma} [2\Phi\left(\frac{o - \mu}{\sigma}\right) - 1] + 2\varphi\left(\frac{o - \mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right\}, \quad (5)$$

5 where  $\Phi(\cdot)$  and  $\varphi(\cdot)$  denote the CDF and the PDF, respectively, of the standard normal distribution.

The CRPS is negatively oriented. A lower CRPS indicates more accurate forecasts; a CRPS of zero denotes a perfect (deterministic) forecast. Moreover, the average score over  $k$  pairs of forecasts  $F_j$  and observations  $y_j$

$$\overline{CRPS} = \frac{1}{k} \sum_{j=1}^k CRPS(F_j, y_j) \quad (6)$$

reduces to the Mean Absolute Error ( $MAE = \frac{1}{k} \sum_{j=1}^k |y_j - \mu_j|$ ) for deterministic forecasts (Gneiting and Raftery, 2004), i.e.,  $F_i$  in Eq. 6 would also be a step function. The CRPS can therefore be interpreted as a distance measure between the probabilistic forecast and the verifying observation (Siegert et al., 2015).

The Continuous Ranked Probability Skill Score ( $CRPSS$ ) is, as the name implies, the corresponding skill score. A skill score relates the accuracy of the prediction system to the accuracy of a reference prediction (e.g., climatology). Thus, with a given  $CRPS_F$  for the hindcast distribution and a given  $CRPS_R$  for the reference distribution the  $CRPSS$  can be defined as:

$$15 \quad CRPSS = 1 - \frac{CRPS_F}{CRPS_R}. \quad (7)$$

Positive values of the CRPSS imply that the prediction system outperforms the reference prediction. Furthermore, this skill score is unbounded for negative values (because hindcasts can be arbitrarily bad) but bounded by 1 for a perfect forecast.

### 3 DeFoReSt: Decadal Climate Forecast Recalibration Strategy

In the following paragraphs we discuss the decadal climate forecast recalibration strategy (*DeFoReSt*) and illustrate how forecast quality is used to estimate the parameters of the recalibration method.

We assume that the recalibrated predictive PDF  $f^{Cal}(X|t, \tau)$  for random variable  $X$  is a normal PDF with mean and variance being functions of ensemble mean  $\mu(t, \tau)$  and variance  $\sigma^2(t, \tau)$ , as well as start time  $t$  and lead year  $\tau$

$$f^{Cal}(X|t, \tau) \sim \mathcal{N}(\alpha(t, \tau) + \beta(t, \tau)\mu(t, \tau), \gamma(t, \tau)^2\sigma^2(t, \tau)). \quad (8)$$

The term  $\alpha(t, \tau)$  accounts for the mean or unconditional bias depending on lead year (i.e., the drift). Analogously  $\beta(t, \tau)$  accounts for the conditional bias. Thus, the expectation  $\underline{E(X) = \alpha(t, \tau) + \beta(t, \tau)\mu(t, \tau)}$   $\underline{E[X] = \alpha(t, \tau) + \beta(t, \tau)\mu(t, \tau)}$  could be a conditional and unconditional bias and drift adjusted deterministic forecast (we call a deterministic forecast a forecast

without specifying uncertainty). For now, we assume that the ensemble spread  $\sigma(t, \tau)$  is sufficiently well related to forecast uncertainty such that it can be adjusted simply by a multiplicative term  $\gamma(t, \tau)^2$ . We thus refrain from using the additive term suggested for NGR by Gneiting et al. (2005) to not end up with a too complex model, as the additive term should consequently be also a function of start time  $t$  and lead time  $\tau$ ; this term might be included in a future variant.

5 In the following, we motivate and develop linear parametric functions for  $\alpha(t, \tau)$ ,  $\beta(t, \tau)$  and  $\gamma(t, \tau)$ .

### 3.1 Addressing bias and drift: $\alpha(t, \tau)$

For bias and drift correction, we start with a parametric approach based on the studies of Kharin et al. (2012) and Kruschke et al. (2015). In their study, a third order polynomial captures the drift along lead time  $\tau$  (Gangstø et al., 2013; Kruschke et al., 2015). Here, Gangstø et al. (2013) suggested that a third order polynomial is a good compromise between flexibility

15 and parameter uncertainty. The drift corrected forecasts  $\hat{H}_{t,\tau,i} - \hat{H}_{h,\tau,i}$  is approximated with a linear function of the forecast  $\hat{H}_{t,\tau,i} - \hat{H}_{h,\tau,i}$  as

$$\hat{H}_{t,\tau,i} = H_{t,\tau,i} - (a_0 + a_1 t) - (a_2 + a_3 t)\tau - (a_4 + a_5 t)\tau^2 - (a_6 + a_7 t)\tau^3. \quad (9)$$

Here,  $H_{t,\tau,i}$  is the raw, i.e., uncorrected, hindcast for the start time  $t$ , ensemble member  $i$  and lead year  $\tau$ . In case the observations and model climatology have different climate trends, the bias between model and observations is non-stationary. Thus,

15 the second term in the approach (9) also accounts for the dependency of the bias on the start year and therefore corrects errors in time trends. Here, as suggested by Kharin et al. (2012) the dependency on start time is only linear to avoid a too complex model. The parameters  $a_0, \dots, a_7$  are estimated by standard least-squares using the differences between the ensemble mean of all available hindcasts and the reanalysis, corresponding to the given start and lead time (Kruschke et al., 2015).

This motivates the following functional form for  $\alpha(t, \tau)$  analogously to Eq. (9)

$$20 \quad \alpha(t, \tau) = \sum_{l=0}^3 (a_{2l} + a_{(2l+1)} t) \tau^l. \quad (10)$$

In principle arbitrary orders are possible for  $t$  and  $\tau$  as long as there is sufficient data to estimate the parameters.

### 3.2 Addressing conditional bias and ensemble spread:

$\beta(t, \tau)$  and  $\gamma(t, \tau)$

In addition Additionally to adjusting the unconditional lead year dependent bias, *DeFoReSt* aims at simultaneously adjusting

25 conditional bias and ensemble spread. As a first approach, we take the same functional form for  $\beta(t, \tau)$  and  $\gamma(t, \tau)$ :

$$\beta(t, \tau) = \sum_{l=0}^3 (b_{2l} + b_{(2l+1)} t) \tau^l, \quad (11)$$

$$\gamma(t, \tau) = \log \left( \sum_{l=0}^2 (c_{2l} + c_{(2l+1)} t) \tau^l \right). \quad (12)$$

The ensemble inflation  $\gamma(t, \tau)$  is, however, assumed to be quadratic at most and constrained to be greater zero by using a static logarithmic link function. We assumed that a higher flexibility may not be necessary, because the MSE – which influences the dispersion – is already addressed by a third order polynomial of unconditional and conditional bias.

These assumptions on model complexity are supported only by our experience; however, they remain subjective. A more transparent order selection will be a topic of future work.

### 3.3 Parameter estimation

The coefficients  $\alpha(t, \tau)$ ,  $\beta(t, \tau)$  and  $\gamma(t, \tau)$  are now expressed as parametric functions of  $t$  and  $\tau$ . The parameters are estimated by minimizing the average *CRPS* over the training period (Gneiting et al., 2005). The associated score function is

$$\Gamma(\mathcal{N}(\alpha(t, \tau) + \beta(t, \tau)\mu, \gamma(t, \tau)^2\sigma^2), o) = \overline{\text{CRPS}} = \frac{1}{k} \sum_{j=1}^k \sqrt{\gamma(t, \tau)^2\sigma_j^2} \left\{ Z_j [2\Phi(Z_j) - 1] + 2\varphi(Z_j) - \frac{1}{\sqrt{\pi}} \right\}, \quad (13)$$

where

$$Z_j = \frac{O_j - (\alpha(t, \tau) + \beta(t, \tau)\mu_j)}{\sqrt{\gamma(t, \tau)^2\sigma_j^2}} \quad (14)$$

is the standardized forecast error for the  $j$ th forecast in the training data set. In the present study, optimization is carried out using the algorithm of Nelder and Mead (1965) as implemented in R (R Core Team, 2016).

The initial guesses for optimization need to be carefully chosen to avoid local minima. Here, we obtain the  $a_i$  and  $b_j$  from linearly modelling the observations  $o$  with the forecast ensemble mean  $\mu$ ,  $t$  and  $\tau$

$$o \sim \sum_{l=0}^A (a_{2l} + a_{(2l+1)}t)\tau^l + \sum_{l=0}^B (b_{2l} + b_{(2l+1)}t)\tau^l \mu, \quad (15)$$

using the notation for linear models from McCullagh and Nelder (1989);  $c_0, c_1, c_2$  are set to zero which yields unity inflation ( $\exp(\gamma(t, \tau)) = 1$ ). However, convergence to a global minimum cannot be guaranteed.

## 4 Calibrating a toy model for decadal climate predictions

In this section, we apply *DeFoReSt* to a stochastic toy model, which is motivated from Weigel et al. (2009), but has been significantly altered to suit the needs of this study. Here, a detailed description of the toy models construction is given in the following subsection. Subsequently, we assess *DeFoReSt* for two exemplary toy model setups.

### 4.1 Toy model construction

The toy model consists of two parts which are detailed in the following two subsections: a) Pseudo-observations, the part generating a substitute  $x(t + \tau)$  for the observations, and b) Pseudo-forecasts, the second part deriving an associated ensemble prediction  $f(t, \tau)$  from thesethis observations. The third subsection motivates the choice of parameters for the toy model.

### 4.1.1 Pseudo-observations

We construct a toy model setup simulating ensemble predictions for the decadal time scale and associated pseudo-observations. Both are based on an arbitrary but predictable signal  $\mu_x$ . The pseudo-observations  $x$  (e.g., annual means of surface temperature over a given area) is the sum of this predictable signal  $\mu_x$  and an unpredictable noise term  $\epsilon_x$ ,

$$x(t + \tau) = \mu_x(t + \tau) + \epsilon_x(t + \tau). \quad (16)$$

Following Kharin et al. (2012)  $\mu_x$  can be interpreted as the atmospheric response to slowly varying and predictable boundary conditions, while  $\epsilon_x$  represents the unpredictable chaotic components of the observed dynamical system. The processes  $\mu_x$  and  $\epsilon_x$  are assumed to be stochastic Gaussian processes

$$\mu_x(t + \tau) \sim \mathcal{N}(0, \sigma_{\mu_x}^2) \quad \text{with} \quad \sigma_{\mu_x}^2 = \eta^2 \leq 1 \quad (17)$$

10 and

$$\epsilon_x(t + \tau) \sim \mathcal{N}(0, \sigma_{\epsilon_x}^2) \quad \text{with} \quad \sigma_{\epsilon_x}^2 = 1 - \eta^2. \quad (18)$$

The variation of  $\mu_x$  around a slowly varying climate signal can be interpreted as the predictable part of decadal variability, its amplitude is given by the variance  $\text{Var}(\mu_x(t + \tau)) = \sigma_{\mu_x}^2$   $\text{var}(\mu_x(t + \tau)) = \sigma_{\mu_x}^2$ . The total variance of the pseudo-observations is thus

$$15 \quad \text{uparrow} \text{arrowoperatornamename} Var(x) = \sigma_x^2 = \sigma_{\mu_x}^2 + \sigma_{\epsilon_x}^2. \text{ Here, the relation of the latter two is uniquely controlled by the parameter } \eta \in [0, 1], \text{ which can be interpreted as potential predictability } (\eta^2 = \sigma_{\mu_x}^2 / \sigma_x^2).$$

In this toy model setup, the specificeoncrete form of thethis variability of  $\mu_x$  and  $\epsilon_x$  is not considered and thus taken as random. A potential climate trend could be superimposed as a time varying mean  $\mu(t) = E(x(t))$   $\mu(t) = E[x(t)]$ . For recalibration As for the recalibration strategy only a difference in trends is important. Here, we use  $\mu(t) = 0$  and  $\alpha(t, \tau)$

20 addressing this difference in trends of forecast and observations.

### 4.1.2 Pseudo-forecasts

We now specify a model giving a potential ensemble forecast with ensemble members  $f_i(t, \tau)$  for observations  $x(t + \tau)$ :

$$f_i(t, \tau) = \mu_{ens}(t, \tau) + \epsilon_i(t, \tau), \quad (19)$$

where  $\mu_{ens}(t, \tau)$  is the ensemble mean and

$$25 \quad \epsilon_i(t, \tau) \sim \mathcal{N}(0, \sigma_{ens}^2(t, \tau)) \quad (20)$$

is the deviation of ensemble member  $i$  from the ensemble mean;  $\sigma_{ens}^2$  is the ensemble variance. In general, ensemble mean and ensemble variance both can both dependent on lead time  $\tau$  and start time  $t$ . We relate the ensemble mean  $\mu_{ens}(t, \tau)$  to the predictable signal in the observations  $\mu_x(t, \tau)$  by assuming a) a systematic deviation characterized by an unconditional

bias  $\chi(t, \tau)$  (accounting also for a drift and difference in climate trends), a conditional bias  $\psi(t, \tau)$  and b) a random deviation  $\epsilon(t, \tau)$ :

$$\mu_{ens}(t, \tau) = \chi(t, \tau) + \psi(t, \tau) (\mu_x(t, \tau) + \epsilon_f(t, \tau)), \quad (21)$$

with  $\epsilon_f(t, \tau) \sim \mathcal{N}(0, \sigma_{\epsilon_f}(t, \tau))$  being a random forecast error with variance  $\sigma_{\epsilon_f}^2(t, \tau) < \sigma_{\epsilon_x}^2$  in order to avoid negative values

5 of the ensemble variance  $\sigma_{ens}^2$ . Although the variance of the random forecast error can in principle be dependent on lead time  $\tau$  and start time  $t$ , we assume for simplicity a constant variance  $\sigma_{\epsilon_f}^2(t, \tau) = \sigma_{\epsilon_f}^2$ .

We further assume an ensemble dispersion related to the variability of the unpredictable noise term  $\epsilon_x$  with an inflation factor  $\omega(t, \tau)$

$$\sigma_{ens}^2(t, \tau) = \omega^2(t, \tau) (\sigma_{\epsilon_x}^2 - \sigma_{\epsilon_f}^2). \quad (22)$$

10 According to Eq. 21 the forecast ensemble mean  $\mu_{ens}$  is simply a function of the predictable signal  $\mu_x$ . In this toy model formulation, an explicit formulation of  $\mu_x$  is not required, hence a random signal might be used for simplicity and it would be legitimate to assume  $E(\mu_x) = \mu(t + \tau) = 0E[\mu_x] = \mu(t + \tau) = 0$  without restricting generality. Here, we propose a linear trend in time  $E[\mu_x] = \mu(t + \tau) = m_0 + m_1 t$  to the pseudo forecasts to emphasize a typical problem encountered in decadal climate prediction: different trends in observations and predictions (Kruschke et al., 2015).

### 15 4.1.3 Choosing the toy models' parameters

This toy model setup is controlled by four parameters: The first parameter  $\eta$  determines the ratio between the variances of the predictable signal and the unpredictable noise term (and thus characterizes potential predictability, see Sec. 4.1.2). Here, we investigate two cases: one with low ( $\eta = 0.2$ ) and one with high potential predictability ( $\eta = 0.8$ ).

The remaining three parameters are  $\chi(t, \tau)$ ,  $\psi(t, \tau)$  and  $\omega(t, \tau)$ , which control the unconditional and the conditional bias and 20 the dispersion of the ensemble spread. To have a toy model experiment related to observations,  $\chi(t, \tau)$  and  $\psi(t, \tau)$  are based on the correction parameters obtained from calibrating the MiKlip Prototype ensemble surface temperature over the North Atlantic against NCEP 20CR reanalyses;  $\chi(t, \tau)$  and  $\psi(t, \tau)$  are based on ratios of polynomials up to 3rd order (in lead years), Eqs. A1 and A2) with coefficients varying with start years (see Figs. 1a and 1b).

The ensemble inflation factor  $\omega(t, \tau)$  is chosen such that the forecast is overconfident for the first lead years and becoming 25 underconfident later; this effect intensifies with start years, see Fig. 1c. A more detailed explanation and numerical values used for the construction of  $\chi(t, \tau)$ ,  $\psi(t, \tau)$  and  $\omega(t, \tau)$  are given in Appendix A.

Given this setup, a choice of  $\chi(t, \tau) \equiv 0$ ,  $\psi(t, \tau) \equiv 1$  and  $\omega(t, \tau) \equiv 1$  would yield a perfectly calibrated ensemble forecast:

$$f^{\text{perf}}(t, \tau) \sim \mathcal{N}(\mu_x(t, \tau), \sigma_{\epsilon_x}^2(t, \tau)). \quad (23)$$

The ensemble mean  $\mu_x(t, \tau)$  of  $f^{\text{perf}}(t, \tau)$  is equal to the predictable signal of the pseudo-observations. The ensemble variance 30  $\sigma_{\epsilon_x}^2(t, \tau)$  is equal to the variance of the unpredictable noise term representing the error between the ensemble mean of  $f^{\text{perf}}(t, \tau)$  and the pseudo-observations. Hence,  $f^{\text{perf}}(t, \tau)$  is perfectly reliable.



Analogous to the MiKlip experiment, the toy model uses 50 start years ( $t = 0, \dots, 49$ ), each with 10 lead years ( $\tau = 1, \dots, 10$ ), and 15 ensemble members ( $i = 1, \dots, 15$ ). The corresponding pseudo-observations  $x(t + \tau)$  run over a period of 59 years in order to cover lead year 10 of start year 50.

## 4.2 Toy model verification

- 5 To assess *DeFoReSt* we consider two extreme toy model setups. The two setups are designed such that the predictable signal is stronger than the unpredictable noise for higher potential predictability (setup 1) and vice versa (setup 2, cf. Sec. 4.1). For each toy model setup we calculated the *ESS*, the *MSE*, time mean intra-ensemble variance and the *CRPSS* with respect to climatology for the corresponding recalibrated toy model.

In addition to the recalibrated pseudo-forecast, we compare

- 10
- a *raw* pseudo-forecast (no correction of unconditional, conditional bias and spread),
  - a *drift-corrected* pseudo-forecast (no correction of conditional bias and spread), and
  - a *perfect* pseudo-forecast (Eq.23, available only in this toy model setup)

All scores have been calculated using cross-validation with a yearly moving calibration window with a width of 10 years. A detailed description of this procedure is given in appendix B.

- 15 The *CRPSS* and reliability values of the perfect forecast could be interpreted as optimum performance within the associated toy model setup, due to the missing bias and ensemble dispersion. For instance, the perfect model's *CRPSS* with respect to climatology would be 1 for a toy model setup with perfect potential predictability ( $\eta = 1$ ) and zero for a setup with no potential predictability ( $\eta = 0$ ). Hence, the climatology could not be outperformed by any prediction model when no predictable signal is existing.

### 20 4.2.1 A toy model setup with high potential predictability

- Figures 2a and 2c show the temporal evolution of the toy model data before and after recalibration with *DeFoReSt* together with the corresponding pseudo-observations. Before recalibration, the pseudo-forecast apparently exhibits the characteristic problems of a decadal ensemble prediction: unconditional bias (drift), conditional bias and underdispersion, which are lead and start time dependent. Additionally, the pseudo-observations and the pseudo-forecast have different trends. After recalibration,
- 25 the lead and start time dependent biases are corrected, such that the temporal evolution of the pseudo-observations is mostly represented by the pseudo-forecast.

Moreover, the pseudo-forecast is almost perfectly reliable after recalibration (not underdispersive), which could be shown with the *ESS* (Fig. 3a). Here, the recalibrated model is nearly identical to the perfect model for all lead years with reliability values close to 1.

- 30 The recalibrated forecast outperforms the raw model output and the drift corrected forecast, replaced whose which *ESS* values are lower one and thus underdispersive. The lower reduced performance of the raw models and the drift correction is

a result of the toy model design, [leading to](#) a higher ensemble mean variance combined with a decreased ensemble spread. In addition, the increased variance of the ensemble mean also results in an increased influence of the conditional bias. The problem is, the raw model forecast and the drift correction could not account for that conditional bias, because neither the ensemble mean nor the ensemble spread were corrected by these forecasts. Therefore, the influence of the conditional bias also becomes noticeable for the reliability of the raw model and the drift corrected forecast; one can see that the minimum and maximum of the conditional bias (see Fig. 1) is reproduced by the reliability values of these forecasts.

Regarding the differences between raw model and the drift corrected forecast, it is visible that the latter outperforms the raw model. The explanation is that the drift correction accounts for the unconditional bias, while the raw model does not correct this type of error. Here, one can see the impact of the unconditional bias on the raw model. Nonetheless, the influence of the unconditional bias is rather small, compared to the conditional bias.

The effect of unconditional and conditional bias is illustrated in Fig. 3b), which shows the  $MSE$  of the different forecasts to the pseudo observations. Here, the drift corrected forecast outperforms the raw model. These forecasts are outperformed by the recalibrated forecast, which simultaneously corrects the unconditional and conditional bias. In this regard, both biases are corrected properly because the  $MSE$  of the recalibrated forecast is almost equal to the perfect models  $MSE$ .

The sharpness of the different forecasts are compared by calculating the time mean intra-ensemble variance (see Fig. 3c). For all lead years, the raw model and the drift corrected forecast exhibit the same sharpness, because the ensemble spread is unmodified for both forecasts.

Another notable aspect is that the raw and drift corrected forecast have a higher sharpness (i.e., lower ensemble variance) than the perfect model for lead years 1 to 4 and vice versa for lead years 5 to 10. This is due to the toy models incorporated underdispersion for the first lead years and an overdispersion for later lead years. Therefore the sharpness of the perfect model could be interpreted as the maximum sharpness of the model without being unreliable.

The sharpness of the recalibrated forecast is very similar to the sharpness of the perfect model for all lead years. The recalibration therefore performs well in correcting under- and overdispersion in the toy model forecasts.

A joint measure for sharpness and reliability is the  $CRPS$  and consequently the  $CRPSS$  with respect to climatology, where the latter is shown in Fig. 3d. The relatively low  $CRPSS$  values of the raw and drift corrected forecast are mainly affected by their reliability; i.e., the unconditional and conditional bias influences are also noticeable for this skill score. Thus, both models exhibit a maximum at lead year 2 and a minimum at lead year 7, where the drift corrected forecast performs better. However, the raw model and the drift corrected forecast are inferior to climatology (the  $CRPSS$  is below zero) for all lead years.

In contrast, the recalibrated forecast approaches  $CRPSS$  values around 0.5 for all lead years and performs nearly identical to the perfect model. This illustrates that the unconditional bias, conditional bias and ensemble dispersion can be corrected with this method.

#### 4.2.2 A toy model setup with low potential predictability

Figures 2b and 2d show the temporal evolution of the toy model data with a low potential predictability before and after recalibration with *DeFoReSt* together with the corresponding pseudo-observations. Before recalibration, the pseudo-forecast

is underdispersive for the first lead years, whereas the ensemble spread increases for later lead years. Moreover, the pseudo-forecast exhibits lead and start time dependent (unconditional) bias (drift) and conditional bias.

After recalibration, the lead and start time dependent biases are corrected, such that the recalibrated forecast mostly describes the trend of the pseudo-observations.

- 5 The recalibrated forecast is also reliable (Fig. 4a); it performs as well as the the perfect model. Here, the value of the  $ESS$  is close to one for both forecasts. Thus, comparing the reliability of the setups with low and high potential predictability, no differences are recognizable. The reason is, the ratio between  $MSE$  and ensemble variance, characterizing the  $ESS$ , does not change much; the lower  $MSE$  performance of the recalibrated forecast (Fig. 4b) is compensated with a higher ensemble variance (Fig. 4c).
- 10 In contrast ~~On the contrary~~, one can see a general improvement of the raw and drift corrected forecasts' reliability compared to the model setup with high potential predictability. The reason is, that the low potential predictability  $\eta$  of this toy model setup leads to smaller variance of the ensemble mean, i.e., the conditional bias has a minor effect. Another aspect for the comparatively good performance of the raw model, is the increased ensemble spread, leading to an enhanced representation of the unconditional bias.
- 15 The minor effect of the conditional bias in ~~within~~ the low potential predictability setup is also represented by the  $MSE$  (Fig. 4b). Here, the difference between drift corrected and recalibrated forecast has decreased compared to ~~the~~ the high potential predictability setup. Comparing both toy model setups, it is also apparent that, for a setup with  $\eta = 0.2$ , the  $MSE$  generally has increased for all forecasts. The reason is, that the predictable signal decreases for a lower  $\eta$ . Therefore, even the perfect models  $MSE$  has increased.
- 20 Figure 4c shows the time mean intra-ensemble variance for the toy model setup with low potential predictability. It is notable that the ensemble variance for this low potential predictability setup is generally greater than for a high  $\eta$  (Fig. 3c). This is due to the fact that the total variance in the toy model is constrained to one and a lower  $\eta$  therefore leads to a greater ensemble spread.
- Nonetheless, the raw model and drift corrected forecast also still have a higher sharpness (i.e., lower ensemble variance) than
- 25 the perfect model for lead years 1 to 4 and vice versa for lead years 5 to 10. Here, the reason for this is again the construction of the toy model, with an underdispersion for the first lead years and an overdispersion for later lead years.
- The recalibrated forecast reproduces the perfect models sharpness also quite well for the potential predictability setup.
- Figure 4d shows the  $CRPSS$  with respect to climatology. Firstly, it is apparent that the weak predictable signal of this toy model setup shifted the  $CRPSS$  of all models closer to zero or the climatological skill. Nevertheless, please note that the
- 30 recalibrated forecast is almost as good as the perfect model and that it is slightly superior to the drift corrected forecast. We conclude that the recalibration works well also in situations with limited predictability.

## 5 Calibrating decadal climate surface temperature forecasts

While in Sec. 4.2 *DeFoReSt* was applied to toy model data, in this section *DeFoReSt* will be applied ~~to~~ surface temperature of MiKlip Prototype runs with MPI-ESM-LR. Here, global mean and a spatial mean values over the North Atlantic subpolar gyre (60°-10°W, 50°-65°N) region will be analyzed.

- 5 Analogous to the previous section we compute the *ESS*, the *MSE* the intra-ensemble variance and the *CRPSS* with respect to climatology. The scores have been calculated for a period from 1961 to 2005. In this section, a 95% confidence interval was additionally calculated for these metrics using a bootstrapping approach with 1000 replicates. For bootstrapping we draw a new pair of dummy time series with replacement from the original validation period and calculate these scores again. This procedure has been repeated 1000 times. Furthermore, all scores have been calculated using cross-validation with  
10 a yearly moving calibration window with a width of 10 years (see appendix B).

### 5.1 North Atlantic mean surface temperature

Figures 5a and 5b show the temporal evolution of North Atlantic mean surface temperature before and after recalibration with the corresponding NCEP 20CR reference. Before recalibration, the MiKlip Prototype hindcasts exhibit a lead time dependent bias (drift) and a lead time dependent ensemble spread. Here, lead time dependent bias of Prototype is a consequence of an  
15 initialization shock due to a full-field initialization (Meehl et al., 2014; Kruschke et al., 2015; Kröger et al., 2017, submitted). After recalibration with *DeFoReSt* the drift of the MiKlip Prototype was corrected and the ensemble spread is also modified.

Regarding the reliability, Fig. 6a shows the *ESS*. The recalibrated forecast is almost perfectly reliable for all lead years because all *ESS* values of this model are close to one. Moreover, the recalibrated forecast is more skillful than the drift corrected forecast for years 3 to 10, where the improvement is only significant for lead year 4 to 8. It is also apparent that the  
20 drift corrected forecast is significantly overdispersive for lead years 3 to 10. For lead years 1 and 2 both post processing methods perform equally well. ~~The~~On the contrary, the raw model's reliability is obviously inferior to the post processed models and significantly underdispersive for all lead years. This implies that the unconditional bias induces most of the systematic error of the MiKlip Prototype runs.

Regarding the *MSE*, one can see that the recalibrated forecast outperforms the drift corrected forecast for lead years 1 and  
25 2 and 8 to 10 (Fig. 6b). Although this improvement of the recalibrated forecast is not significant, it may be still attributed to its correction of the conditional bias. Here, the raw model performs obviously worse compared to the post processed models, because neither the unconditional nor the conditional bias were corrected.

Figure 6c shows the spread as measured by the time mean intra-ensemble variance for the North Atlantic mean surface temperature. The ensemble variance of the raw model and the drift corrected forecast is equal, since the ensemble spread of  
30 the drift corrected forecast was not corrected. Here, the ensemble variance of both models is increasing with lead times. The ensemble variance of the recalibrated forecast is lower than the variance of the raw and drift corrected forecast for the first lead years 2 to 10, i.e., the recalibrated forecast has a higher sharpness than the other two forecasts. The combination of increasing

ensemble variance and almost constant  $MSE$  leadingyielding to the identified increasing underconfidence (see Fig. 6a) of the drift corrected forecast for that period.

Figure 6d shows that in terms of  $CRPSS$  both the drift corrected forecast and the recalibrated forecast outperform the raw model. Here, the  $CRPSS$  of the raw model is lesssmaller than -1 for all lead year, thus the corresponding graph lies below the plotted range. *DeFoReSt* slightly performs slightly better (but not significantly better) than the drift corrected forecast for almost all lead years, except lead year 3 and 4. Additionally, the  $CRPSS$  with respect to climatology shows that the recalibrated forecast outperforms a constant climatological forecast for all lead times and is significantly better for lead years 1 and 3 to 10.

## 5.2 Global mean surface temperature

Figures 7a and 7b show the temporal evolution of global mean surface temperature before (see equation 19) and after recalibration andwith the corresponding NCEP 20CR reference. Before recalibration with *DeFoReSt*, the MiKlip Prototype hindcasts exhibit a lead time dependent bias (drift) and a lead time dependent ensemble spread. The drift of the global mean surface temperature is even stronger than the North Atlantic counterpart. After applying *DeFoReSt*, the drift of the MiKlip Prototype was corrected and the ensemble spread is basicallyfairly constant for all lead times.

The  $ESS$  for a global mean surface temperature is shown in Fig. 8a. It can be seen that the recalibrated forecast is also perfectly reliable for the global mean surface temperature. Here, all  $ESS$  values are near one. Additionally, the recalibrated forecast is more skillful than the drift corrected forecast for all lead years. Here, only lead year 1 and 10 are significant. The reliability values of the drift corrected forecast indicate a significantly overconfidence for almost every lead year. As for the North Atlantic mean, the raw model's reliability for a global mean temperature is inferior to the post processed models.

Figure 8b shows the  $MSE$ . It is apparent that the recalibrated forecast outperforms the drift corrected forecast for all lead years, where the improvement for lead years 5 to 6 and 8 to 10 is significant. Moreover, the  $MSE$  of the drift corrected forecast increases with lead years, while the  $MSE$  of the recalibrated forecast is almost constant. Thus, this increasing difference between these forecasts is an effect of a lead year dependency of the conditional bias.

Figure 8c shows the time mean intra-ensemble variance for the global mean surface temperature. Regarding sharpness, the drift corrected and the recalibrated forecast perform similarlyequally for lead years 2 and 3. Hence, the improved reliability of the recalibrated forecast could not attributed to a modified ensemble spread. The explanation is that the recalibration method also accounts for conditional and unconditional bias, while the drift correction method only addresses to the unconditional bias. Thus, the error between observation and ensemble mean of the recalibrated forecast is lower than the error of the drift corrected forecast (see Fig. 8b). Consequently, the drift corrected forecast is overconfident for this period (see Fig. 8a), due to a greater error combined with an equal sharpness.

Regarding the  $CRPSS$ , Fig. 8d shows that *DeFoReSt* performs significantly better than the drift corrected forecast for lead years 1, and 8 to 10. Furthermore, the  $CRPSS$  shows that these forecasts also outperforming the climatology, where the improvement of the drift corrected forecast against climatology is solely not significant for lead years 8 to 9. The  $CRPSS$  of the raw model is smaller than -1 for all lead years and therefore out of the shown range.

All in all, the better *CRPSS* performance of *DeFoReSt* model could be explained due to a superior reliability for all lead years (see Fig. 8a).

## 6 Summary and conclusions

There are many studies describing recalibration methods for weather and seasonal forecasts (e.g., Gneiting et al., 2005; Weigel et al., 2009). Regarding decadal climate forecasts, those methods cannot be applied easily, because decadal climate prediction systems on that time scale exhibit characteristic problems including model drift (lead time dependent unconditional bias) and climate trends which could differ from observations. In this regard Kruschke et al. (2015); Kharin et al. (2012) proposed methods to account for lead and start time dependent unconditional biases of decadal climate predictions.

In addition to unconditional biases, probabilistic forecasts could show lead and start year dependent conditional biases and under- or overdispersion. Therefore, we proposed the post processing method *DeFoReSt* which accounts for the three above mentioned issues. Following the suggestion for the unconditional bias (Kruschke et al., 2015), we allow for the conditional bias and the ensemble dispersion to change [polynomially](#) with lead time and linearly with start time. Two advantages of a polynomial fit over the common exponential fit (e.g., as proposed by Kharin et al. (2012)) are stated by Gangstø et al. (2013): First, for a small sample size (this is given for decadal climate predictions) the fit of an exponential with offset is relatively difficult and unreliable. Second, a polynomial approach can capture a local maximum/minimum of the above mentioned errors at a specific lead time; the evolution of these errors may be non-monotonous. [Following Analog to](#) Kruschke et al. (2015), we chose a third order polynomial approach for the correction parameter of the unconditional bias and the conditional bias. A second order polynomial approach is chosen for the correction parameter of the ensemble dispersion. Note that these choices might influence the resulting forecast skill. It might be worth using a transparent model selection strategy, this is topic of future research. The associated *DeFoReSt* parameters are estimated by minimization of the *CRPS* (Gneiting et al., 2005). The *CRPSS*, the *ESS*, the time mean intra-ensemble variance (as measure for sharpness) and the *MSE* assess the performance of *DeFoReSt*. All scores were calculated with 10 year block-wise cross-validation.

We investigated *DeFoReSt* using toy model simulations with high ( $\eta = 0.8$ ) and low potential predictability ( $\eta = 0.2$ ). Errors based on the same polynomial structure as used for the recalibration method were imposed. *DeFoReSt* is compared to a conventional drift correction and a perfect toy model without unconditional bias, conditional bias [or](#) ensemble spread [dispersion](#) was used as a benchmark. Here, the recalibration and drift correction benefits from the fact that the structure of errors imposed is known. Although the model for the error structure is flexible, the gain in skill is an upper limit to other applications where the structure of errors is unknown. Conclusions on the relative advantage of *DeFoReSt* over the drift correction for different potential predictability setups, however, should be largely unaffected by the choice of toy model errors.

A recalibrated forecast shows (almost) perfect reliability ( $ESS = 1$ ). Sharpness can be improved due to the correction of conditional and unconditional biases. Thus, given a high potential predictability ( $\eta = 0.8$ ), recalibration leads to major improvements in skill (*CRPSS*) over a climatological forecast. Forecasts with low potential predictability ( $\eta = 0.2$ ) improve also but the gain in skill (*CRPSS*) over a climatological forecast is limited. In both cases, reliability, sharpness and thus

*CRPSS* of the recalibrated model are almost equal to the perfect model. *DeFoReSt* outperforms the drift corrected forecast with respect to *CRPSS*, reliability and *MSE*, due to additional correction of the conditional bias and the ensemble dispersion. The differences between these two post processed forecasts are, however, smaller for the low potential predictability setup.

We also applied *DeFoReSt* to surface temperature data of the MiKlip Prototype decadal climate forecasts, spatially averaged over the North Atlantic subpolar gyre region and a global mean. Pronounced predictability for these cases has been identified by previous studies (e.g., Pohlmann et al., 2009; van Oldenborgh et al., 2010; Matei et al., 2012; Mueller et al., 2012). Nonetheless, both regions are also affected by model drift (Kröger et al., 2017, submitted). The North Atlantic region shows an overconfident forecasts for all lead years for the raw model output. The drift corrected forecast is underconfident for lead years 8 to 10. The recalibrated forecast is almost perfectly reliable for all lead years ( $ESS = 1$ ) and outperforms the drift correction method with respect to *CRPSS* for lead years 1 and 2 and 5 to 10. For the global mean surface temperature *DeFoReSt* significantly outperforms the drift corrected forecast for several lead years with respect to *CRPSS*. The *CRPSS* in for the global case is generally higher than for the North Atlantic region. The recalibrated global forecast is perfectly reliable; the drift corrected forecast, however, tends to be overconfident for all lead years. This is in accordance to other studies suggesting that ensemble forecasts typically underestimate the true uncertainty and tend to be overconfident (Weigel et al., 2009; Hamill and Colucci, 1997; Eckel and Walters, 1998). *DeFoReSt* thus accounts for both, underdispersive and overdispersive forecasts.

*DeFoReSt* with third/second order polynomials is quite successful. However, it is worthwhile investigating the use of order selection strategies, such as LASSO (Tibshirani, 1996) or information criteria. Furthermore parameter uncertainty due to a small training size may result in forecast that are still underdispersive after recalibration. For the seasonal scale, this has been discussed by Siegert et al. (2015). However, for decadal climate forecasts, this aspect should be further considered in future studies. Recalibration based on *CRPS*-minimization is computationally expensive which might become problematic if not regional means but individual grid points are considered. As an alternative to *CRPS*-minimization Vector Generalized Linear Model (VGLM, Yee, 2008) might be considered which have been implemented in an efficient way. We proposed *DeFoReSt* to recalibrate ensemble predictions of a single model. However, *DeFoReSt* needs to be modified if it should be applied to a multimodel prediction. This is necessary because single ensemble members of a multimodel may differ in their systematic errors (Tebaldi et al,2005; Arisodo et al. 2017).

Based on simulations from a toy model and the MiKlip decadal climate forecast system we could show that *DeFoReSt* is a consistent recalibration strategy for decadal forecast leading to reliable forecast with increased sharpness due to simultaneous adjustment of conditional and unconditional biases depending no lead-time.

*Code and data availability.* The NCEP 20CR reanalysis used in this study are freely accessible through NCAR (National Center for Atmospheric Research) after a simple registration process. The MiKlip Prototype data used for this paper are from the BMBF-funded project MiKlip and are available on request. The post-processing, toy model and cross-validation algorithms are implemented using GNU licensed free software from the R Project for Statistical Computing (<http://www.r-project.org>). Our implementations are available on request.



## Appendix A: Construction of the toy model's free parameters

For this toy model setup,  $\chi(t, \tau)$  and  $\psi(t, \tau)$  are obtained from  $\alpha(t, \tau)$  and  $\beta(t, \tau)$  as follows:

$$\chi(t, \tau) = -\frac{\alpha(t, \tau)}{\beta(t, \tau)} \quad (\text{A1})$$

$$\psi(t, \tau) = \frac{1}{\beta(t, \tau)} \quad (\text{A2})$$

$$5 \quad \omega(t, \tau) = \frac{1}{\gamma(t, \tau)}. \quad (\text{A3})$$

The parameters  $\chi(t, \tau)$ ,  $\psi(t, \tau)$  and  $\omega(t, \tau)$  are defined such that a perfectly recalibrated toy model forecast  $f^{\text{Cal}}$  would have the following form:

$$f_i^{\text{Cal}}(t, \tau) = \alpha(t, \tau) + \beta(t, \tau)\mu_{\text{ens}}(t, \tau) + \gamma(t, \tau)\epsilon_i, \quad (\text{A4})$$

where  $\epsilon_i$  is the deviation of each ensemble member  $i$  from the ensemble mean  $\mu_{\text{ens}}(t, \tau)$ . Here,  $\sigma_{\text{ens}}^2$  is the ensemble variance.

10 Writing (A4) as Gaussian distribution and applying the definitions of  $\mu_{\text{ens}}$  (Eq. 21) and  $\sigma_{\text{ens}}$  (Eq. 22) leads to

$$f_i^{\text{Cal}}(t, \tau) \sim \mathcal{N}(\alpha(t, \tau) + \beta(t, \tau)(\chi(t, \tau) + \psi(t, \tau)\mu_x(t, \tau)), \gamma(t, \tau)\omega(t, \tau)\sigma_{\epsilon_x}^2(t, \tau)), \quad (\text{A5})$$

and applying the definitions of  $\chi(t, \tau)$ ,  $\psi(t, \tau)$  and  $\omega(t, \tau)$  (Eqs. A1-A3) to (A5) would further lead to:

$$f_i^{\text{Cal}}(t, \tau) \sim \mathcal{N}(\alpha(t, \tau) - \beta(t, \tau)\frac{\alpha(t, \tau)}{\beta(t, \tau)} + \frac{\beta(t, \tau)}{\beta(t, \tau)}\mu_x(t, \tau), \frac{\gamma(t, \tau)}{\gamma(t, \tau)}\sigma_{\epsilon_x}^2(t, \tau)), \quad (\text{A6})$$

This shows that  $f^{\text{Cal}}$  is equal to the perfect toy model  $f^{\text{Perf}}(t, \tau)$  (Eq. 23):

$$15 \quad f^{\text{Cal}}(t, \tau) \sim \mathcal{N}(\mu_x(t, \tau), \sigma_{\epsilon_x}^2(t, \tau)). \quad (\text{A7})$$

This setting has the advantage that the perfect estimation of  $\alpha(t, \tau)$ ,  $\beta(t, \tau)$  and  $\gamma(t, \tau)$  is already known prior to calibration with *CRPS*-minimization.

Following the suggestion of Kruschke et al. (2015), a third order polynomial approach was chosen for unconditional  $\alpha(t, \tau)$  and conditional bias  $\beta(t, \tau)$  as well as for the inflation factor  $\omega(t, \tau)$ , yielding

$$20 \quad \alpha(t, \tau) = (a_0 + a_1t) + (a_2 + a_3t)\tau + (a_4 + a_5t)\tau^2 + (a_6 + a_7t)\tau^3, \quad (\text{A8})$$

$$\beta(t, \tau) = (b_0 + b_1t) + (b_2 + b_3t)\tau + (b_4 + b_5t)\tau^2 + (b_6 + b_7t)\tau^3 \text{ and} \quad (\text{A9})$$

$$\omega(t, \tau) = (w_0 + w_1t) + (w_2 + w_3t)\tau + (w_4 + w_5t)\tau^2 + (w_6 + w_7t)\tau^3. \quad (\text{A10})$$

For the current toy model experiment, we exemplarily specify values for  $u_i$  and  $v_i$  as obtained from calibrating the ensemble mean of MiKlip Prototype GECCO2 ( $\bar{f}_{\text{Prot}}$ ) surface temperature over the North Atlantic against NCEP 20CR reanalyses

25 ( $T_{\text{obs}}$ ):

$$\begin{aligned} \underline{E(T_{\text{obs}})} \underline{E(T_{\text{obs}})} &\sim (a_0 + a_1t) + (a_2 + a_3t)\tau + (a_4 + a_5t)\tau^2 + (a_6 + a_7t)\tau^3 + \\ &((b_0 + b_1t) + (b_2 + b_3t)\tau + (b_4 + b_5t)\tau^2 + (b_6 + b_7t)\tau^3)\bar{f}_{\text{Prot}}. \end{aligned} \quad (\text{A11})$$



The values of the coefficients are given in Tab. A1 (upper and middle row). The last row of Tab. A1 gives the values of  $w_i$ , i.e., the series expansion of the inflation factor  $\omega(t, \tau)$ . These are chosen such that the forecast is overconfident for the first lead years and becoming underconfident for later lead years (see Fig. 1c).

## Appendix B: Cross-validation procedure for decadal climate predictions

- 5 We propose a cross-validation setting for decadal climate predictions to ensure fair conditions for assessing the benefit of a post processing method over a raw model without any post processing. All scores are calculated with a yearly moving validation period with a length of 10 years. This means that one start year including 10 lead years was left out for validation. The remaining start years and the corresponding lead years were used for estimating the correction parameters for the prediction within the validation period; start years within the validation period were not taken into account. This procedure was repeated
- 10 for a start year wise shifted validation period.

This setting is illustrated in Fig. A1 for an exemplary validation period from 1964 to 1973, i.e., the correction parameters are estimated for all hindcasts which are initialized outside the validation period (1962; 1963; 1974; 1975,...).

*Competing interests.* The authors declare that they have no conflict of interest.

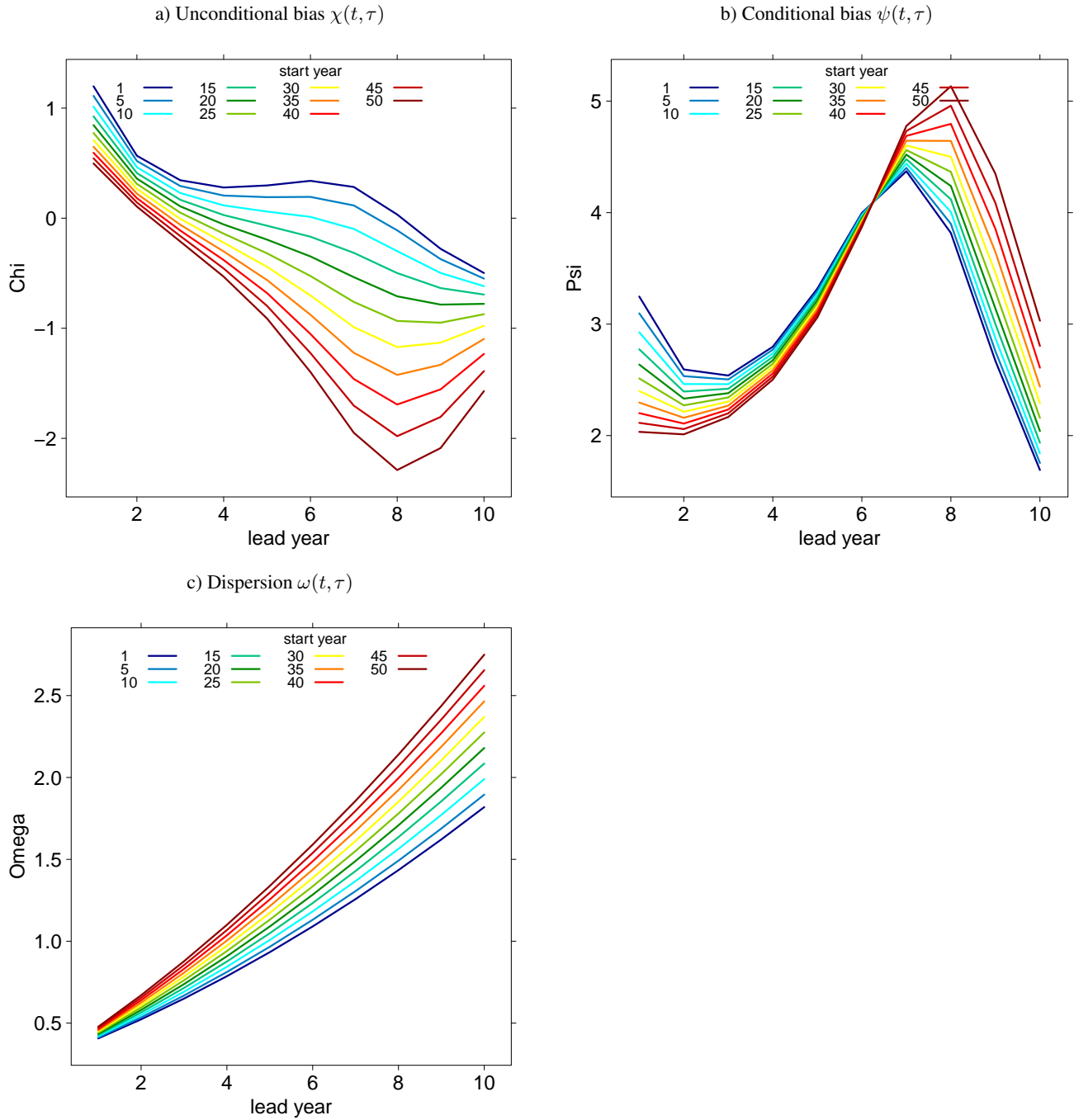
- 15 *Acknowledgements.* This study was funded by the German Federal Ministry for Education and Research (BMBF) project MiKlip (sub-projects CALIBRATION Förderkennzeichen FKZ 01LP1520A and FLEXFORDEC Förderkennzeichen: FKZ 01LP1519A). Mark Liniger and Jonas Bend have received funding from the European Union's Seventh Framework Programme [FP7/2007- 2013] under Grant Agreement no 308291.

## References

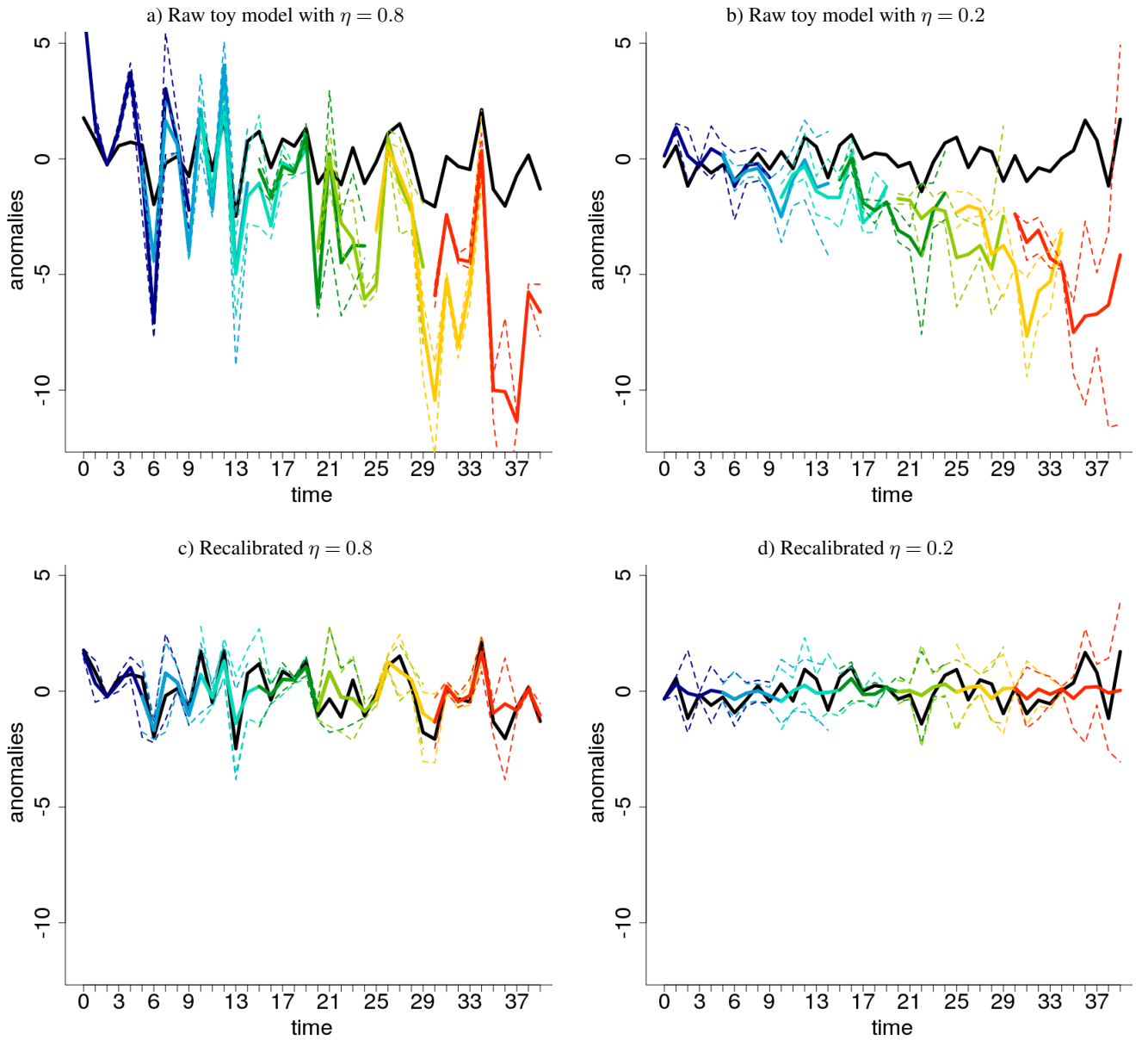
- Anderson, J. L.: A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Climate*, 9, 1518–1530, 1996.
- Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., and others.: SThe twentieth century reanalysis project., *Q. J. Roy. Meteorol. Soc.*, 137(654), 1–28, 2011.
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., et al.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system., *Quart. J. Roy. Meteor. Soc.*, 137(656), 553–597, <https://doi.org/doi: 10.1002/qj.828>, 2011.
- Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting–II. Calibration and combination, *Tellus A*, 57, 234–252, 2005.
- Eckel, F. A. and Walters, M. K.: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble., *Wea. Forecasting*, 13, 1132–1147, 1998.
- Fučkar, N. S., Volpi, D., Guemas, V., and Doblas-Reyes, F. J.: A posteriori adjustment of near-term climate predictions: Accounting for the drift dependence on the initial conditions, *Geophysical Research Letters*, 41, 5200–5207, 2014.
- Gangstø, R., Weigel, A. P., Liniger, M. A., , and Appenzeller, C.: Methodological aspects of the validation of decadal predictions., *Climate Res.*, 55(3), 181–200, <https://doi.org/doi: 10.3354/cr01135>, 2013.
- Gneiting, T. and Katzfuss, M.: Probabilistic forecasting., *Annual Review of Statistics and Its Application*, 1, 125–151, 2014.
- Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation., *Tech. Rep. 463*, Department of Statistics, University of Washington, 29 pp. [Available online at [www.stat.washington.edu/tech.reports/](http://www.stat.washington.edu/tech.reports/)], 2004.
- Gneiting, T. and Raftery, A. E.: Weather forecasting with ensemble methods., *Science*, 310, 248, 2005.
- Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation., *Journal of the American Statistical Association*, 102 (477), 359–378, 2007.
- Gneiting, T., Raftery, A. E., Balabdaoui, F., and Westveld, A. H.: Verifying probabilistic forecasts: Calibration and sharpness., *Proc. Workshop on Ensemble Weather Forecasting in the Short to Medium Range*. Val-Morin, QC, Canada., 2003.
- Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Monthly Weather Review*, 133, 1098–1118, 2005.
- Hamill, T. M. and Colucci, S. J.: Verification of Eta-RSM short-range ensemble forecasts., *Mon. Wea. Rev.*, 125, 1312–1327, 1997.
- Jungclaus, J. H., Fischer, N., Haak, H., Lohmann, K., Marotzke, J., Mikolajewicz, D. M. U., Notz, D., and von Storch, J. S.: Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model, *J. Adv. Model. Earth Syst.*, 5(2), 422–446, 2013.
- Keller, J. D. and Hense, A.: A new non-Gaussian evaluation method for ensemble forecasts based on analysis rank histograms, *Meteorologische Zeitschrift*, 20, 107–117, 2011.
- Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F., and Lee, W.-S.: Statistical adjustment of decadal predictions in a changing climate, *Geophysical Research Letters*, 39, 2012.
- Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U.: Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, *Meteor. Z.*, 01, –, <https://doi.org/10.1127/metz/2015/0641>, 2015.

- Kröger, J., Pohlmann, H., Sienz, F., Marotzke, J., Baehr, J., Köhl, A., Modali, K., Polkova, I., Stammer, D., Vamborg, F., and Müller, W. A.: Full-Field initialized decadal predictions with the MPI Earth System Model: An initial shock in the North Atlantic., *Clim. Dyn.*, 2017, submitted.
- Köhl, A.: Evaluation of the GECCO2 ocean synthesis: transports of volume, heat and freshwater in the Atlantic., *Quart. J. Roy. Meteor. Soc.*, 5 141(686), 166–181, 2015.
- Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., Kaspar, F., Kottmeier, C., Marini, C., Polkova, I., et al.: Miklip – a national research project on decadal climate prediction, *Bull. Amer. Meteorol. Soc.*, 97, 2379–2394, 2016.
- Matei, D., Baehr, J., Jungclaus, J. H., Haak, H., Müller, W. A., and Marotzke, J.: Multiyear prediction of monthly mean Atlantic meridional overturning circulation at 26.5 N, *Science*, 335, 76–79, 2012.
- 10 McCullagh, P. and Nelder, J.: *Generalized Linear Models*, CRC Press, Boca Raton, Fla, 2 edn., 1989.
- Meehl, G. A., Goddard, L., Boer, G. J., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., et al.: Decadal Climate Prediction: An Update from the Trenches, *B. Am. Meteorol. Soc.*, 95, 243–267, doi: 10.1175, Tech. rep., BAMS-D-12-00241.1, 2014.
- Mueller, W. A., Baehr, J., Haak, H., Jungclaus, J. H., Kröger, J., Matei, D., Notz, D., Pohlmann, H., Storch, J., and Marotzke, J.: Forecast 15 skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology, *Geophysical Research Letters*, 39, 2012.
- Nelder, J. A. and Mead, R.: A simplex method for function minimization, *The computer journal*, 7, 308–313, 1965.
- Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., and Smith, L.: Ensemble prediction: a pedagogical perspective, *ECMWF newsletter*, 106, 10–17, 2006.
- 20 Palmer, T. N., Doblas-Reyes, F. J., Weisheimer, A., and Rodwell, M. J.: Toward seamless prediction: Calibration of climate change projections using seasonal forecasts, *Bulletin of the American Meteorological Society*, 89, 459–470, 2008.
- Pohlmann, H., Jungclaus, J. H., Köhl, A., Stammer, D., and Marotzke, J.: Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic, *Journal of Climate*, 22, 3926–3938, 2009.
- Pohlmann, H., Mueller, W. A., Kulkarni, K., Kameswarrao, M., Matei, D., Vamborg, F., Kadow, C., Illing, S., and Marotzke, J.: Improved 25 forecast skill in the tropics in the new MiKlip decadal climate predictions, *Geophysical Research Letters*, 40, 5798–5802, 2013a.
- R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>, 2016.
- Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, 133, 1155–1174, 2005.
- 30 Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., and others.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century., *J. Geophys. Res. Atmos.*, 108(D14), 4407, 2003.
- Siebert, S., Sansom, P. G., and Williams, R.: Parameter uncertainty in forecast recalibration, *Quarterly Journal of the Royal Meteorological Society*, 2015.
- Sloughter, J. M., Raftery, A. E., Gneiting, T., and Fraley, C.: Probabilistic quantitative precipitation forecasting using Bayesian model aver- 35 aging, *Monthly Weather Review*, 135, 3209–3220, 2007.
- Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., et al.: Atmospheric component of the MPI-M Earth System Model: ECHAM6, *J. Adv. Model. Earth Syst*, 5(2), 146–172, <https://doi.org/10.1002/jame.20015>, 2013.

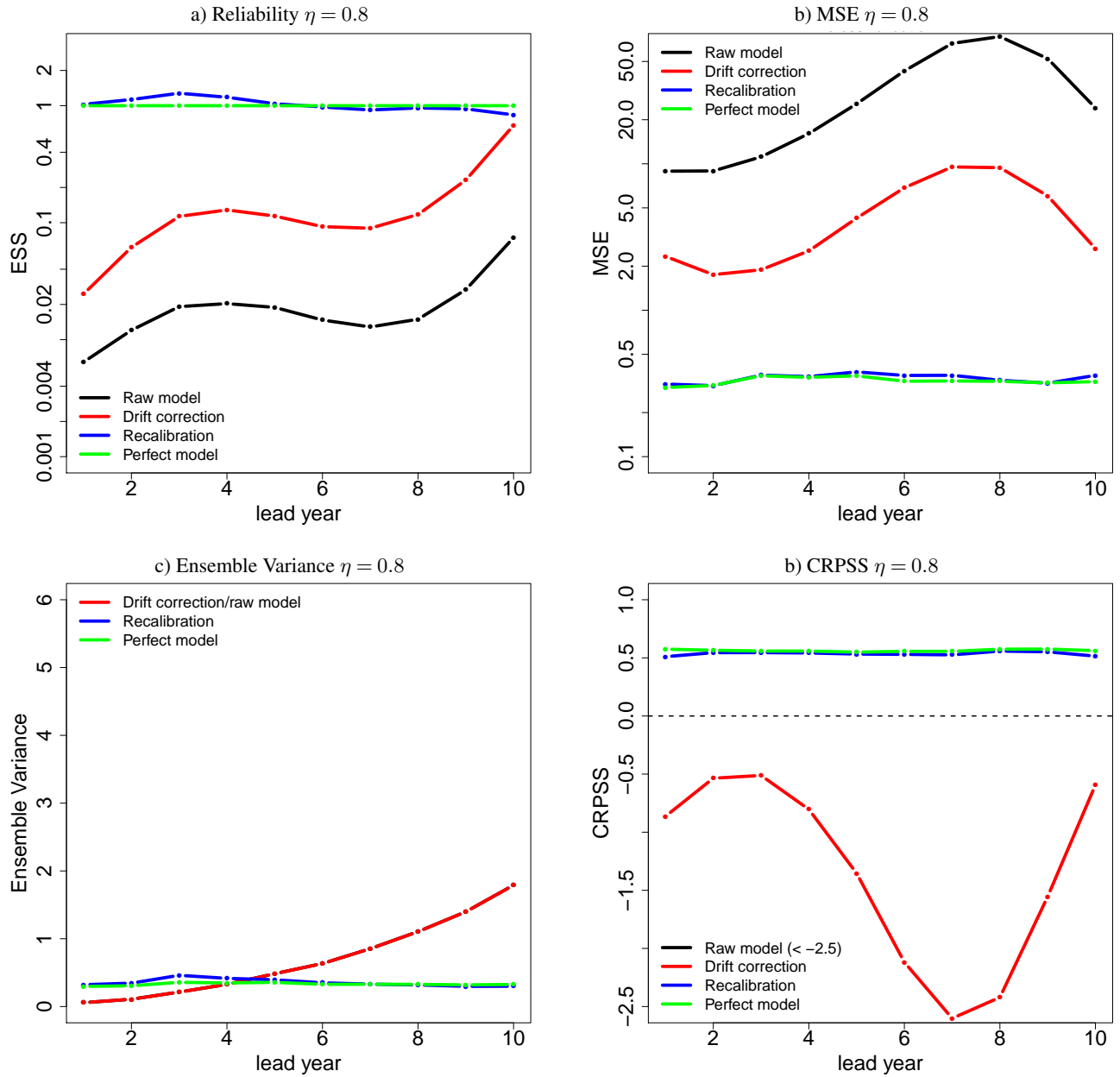
- Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems., Proc. Workshop on Predictability, Reading, United Kingdom, European Centre for Medium- Range Weather Forecasts., pp. 1–25, 1997.
- Tibshirani, R.: Regression shrinkage and selection via the lasso, J. Royal Statist. Soc. B, pp. 267–288, 1996.
- Uppala, S., Kallberg, P., Simmons, A., Andrae, U., et al.: The ERA-40 re-analysis, Quart. J. Roy. Meteor. Soc., 131, 2961–3012, 2005.
- 5 van Oldenborgh, G. J., Doblas Reyes, F., Wouters, B., and Hazeleger, W.: Skill in the trend and internal variability in a multi-model decadal prediction ensemble, in: EGU General Assembly Conference Abstracts, vol. 12, p. 9946, 2010.
- Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels?, Mon. Weather Rev., 137(4), 1460–1479, 2009.
- Yee, T. W.: VGAM: Vector generalized linear and additive models, R package version 0.7-7, URL [http://CRAN.R-project.org/package=](http://CRAN.R-project.org/package=VGAM)
- 10 VGAM, 2008.



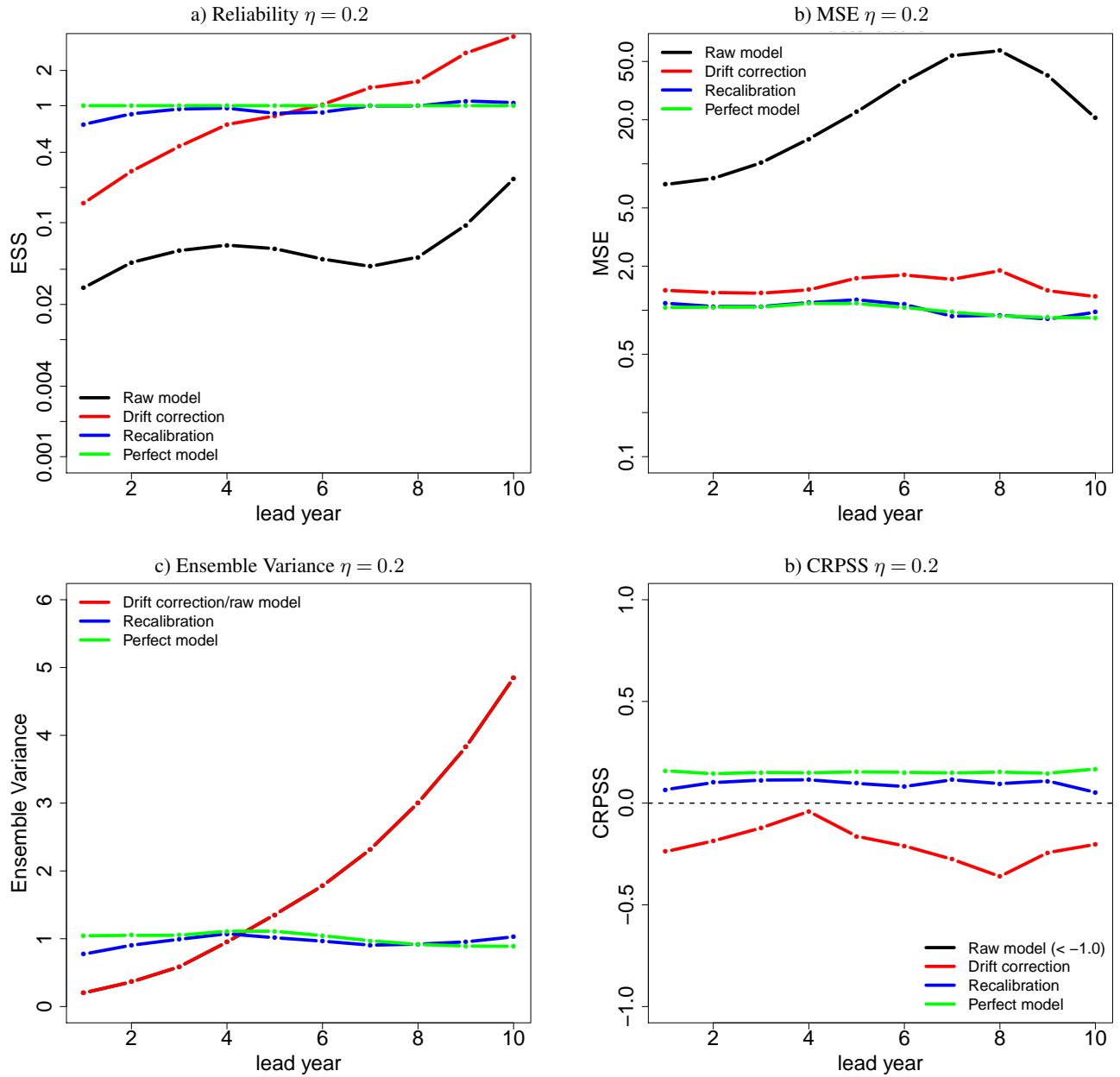
**Figure 1.** Unconditional bias (a,  $\chi(t, \tau)$ ), conditional bias (b,  $\psi(t, \tau)$ ), and dispersion of the ensemble spread (c,  $\omega(t, \tau)$ ) as a function of lead year  $\tau$  with respect to different start years  $t$ .



**Figure 2.** Temporal evolution of the raw (a, b) and with *DeFoReSt* recalibrated (c, d) pseudo-forecast for different start years (colored lines) with potential predictability  $\eta = 0.8$  (a, c) and  $\eta = 0.2$  (b, d). Each pseudo-forecast runs over 10 lead years. The black line represents the associated pseudo-observation.

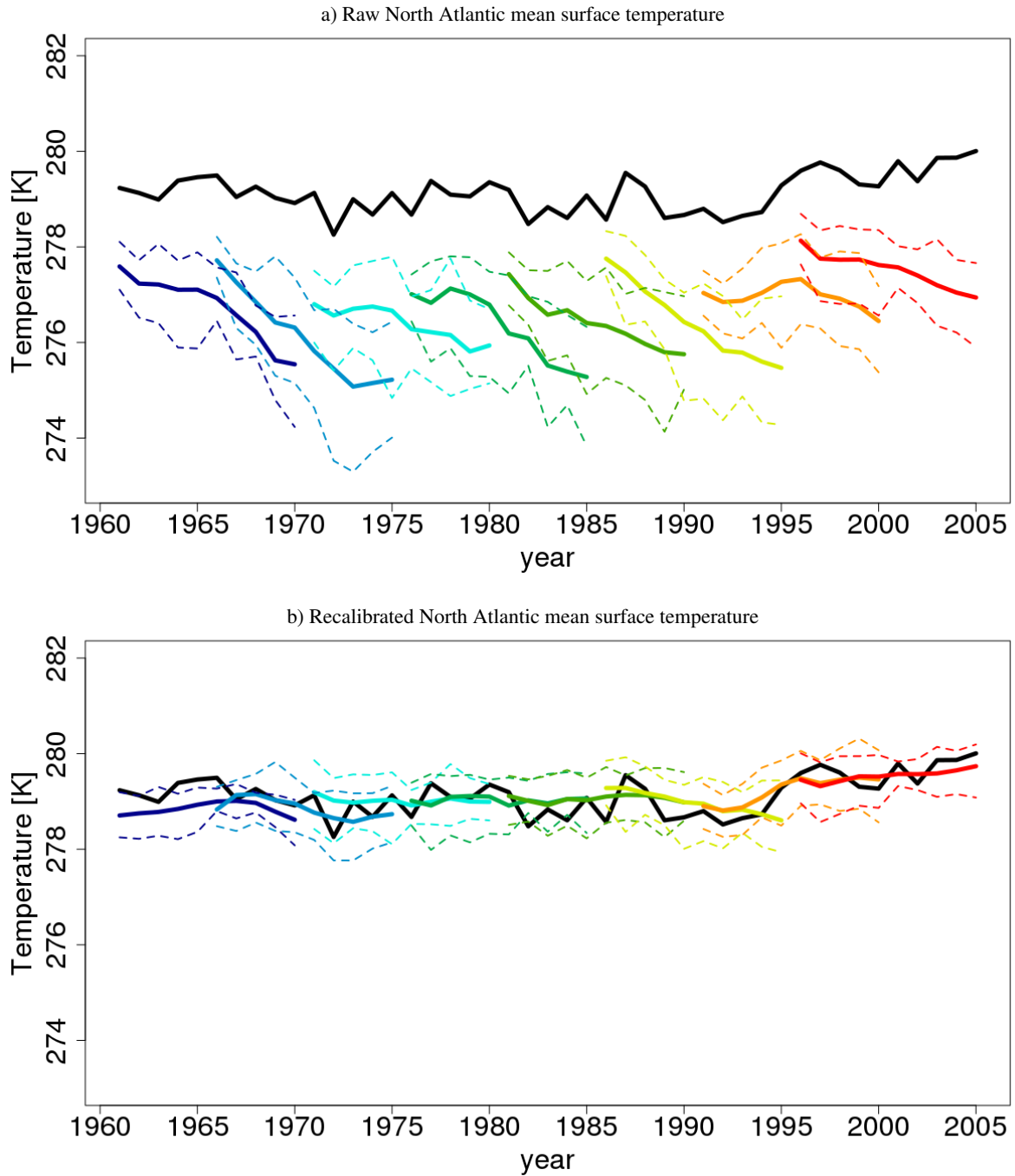


**Figure 3.** Reliability (a), MSE (b), Ensemble Variance (c) and CRPSS (d) of the raw toy model (black line), the drift corrected toy model forecast (red line), recalibrated (*DeFoReSt*) toy model forecast (blue line) and the perfect toy model (green line) for  $\eta = 0.8$ . The drift correction method does not account for the ensemble spread, thus the ensemble variance of the raw model and the drift corrected forecast are equal. For reasons of clarity the raw models CRPSS with values between -5 and -9 are not shown here.

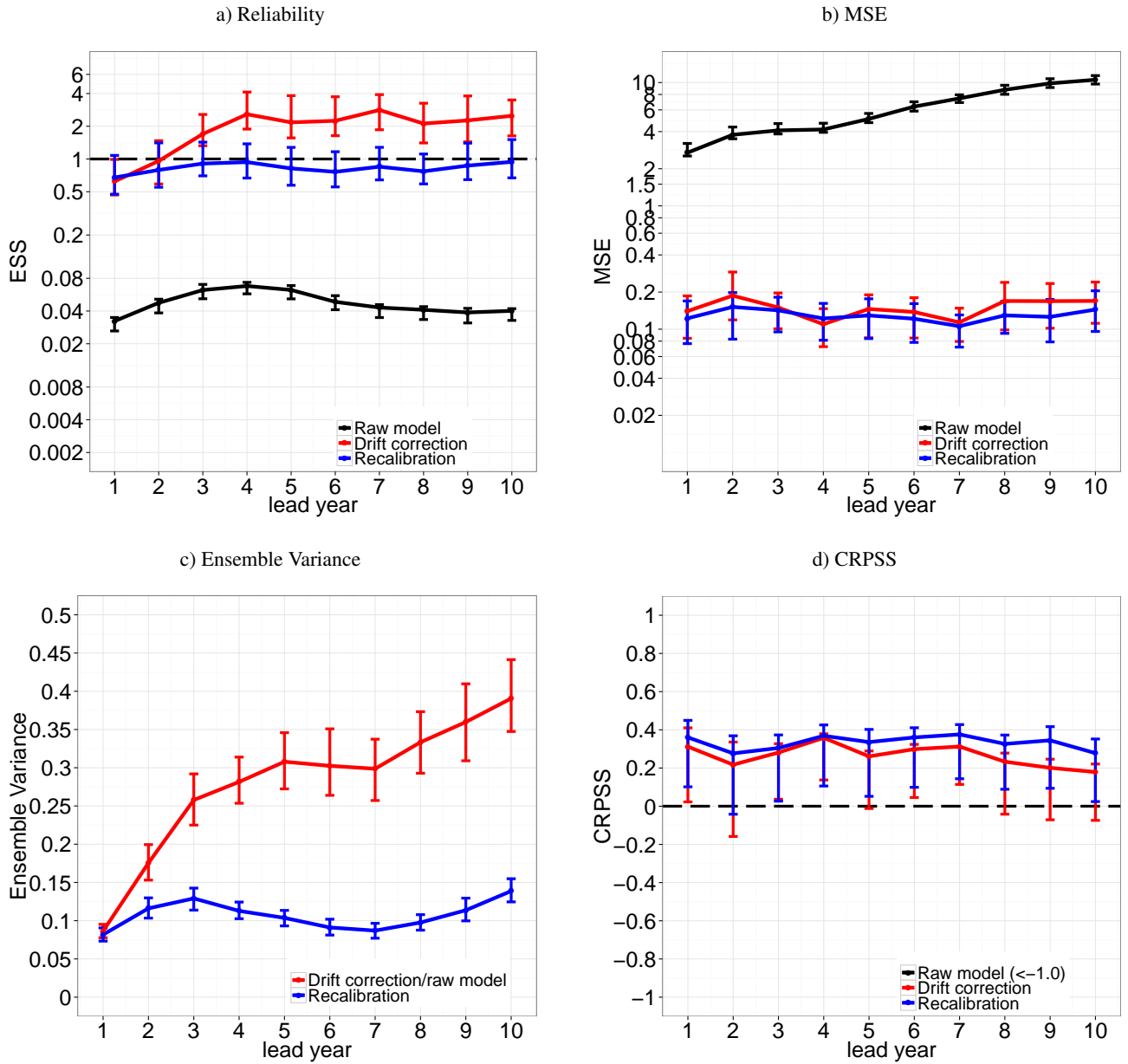


**Figure 4.** Reliability (a), MSE (b), Ensemble Variance (c) and CRPSS (d) of the raw toy model (black line), the drift corrected toy model (red line), recalibrated (*DeFoReSt*) toy model (blue line) and the perfect toy model (green line) for  $\eta = 0.2$ . The drift correction method does not account for the ensemble spread, thus the ensemble variance of the raw model and the drift corrected forecast are equal. For reasons of clarity the raw models CRPSS with values between  $-3$  and  $-9$  are not shown here.

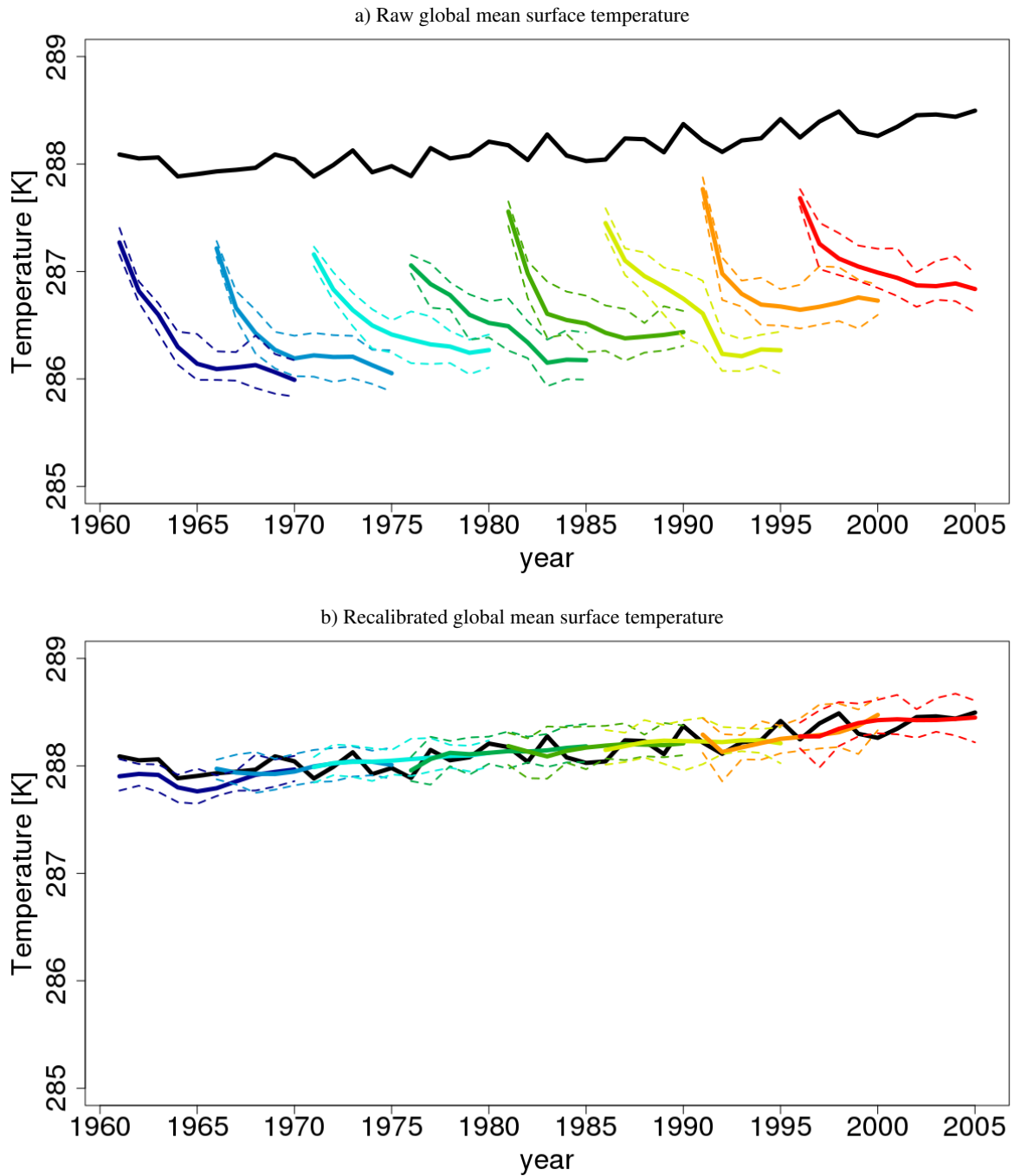




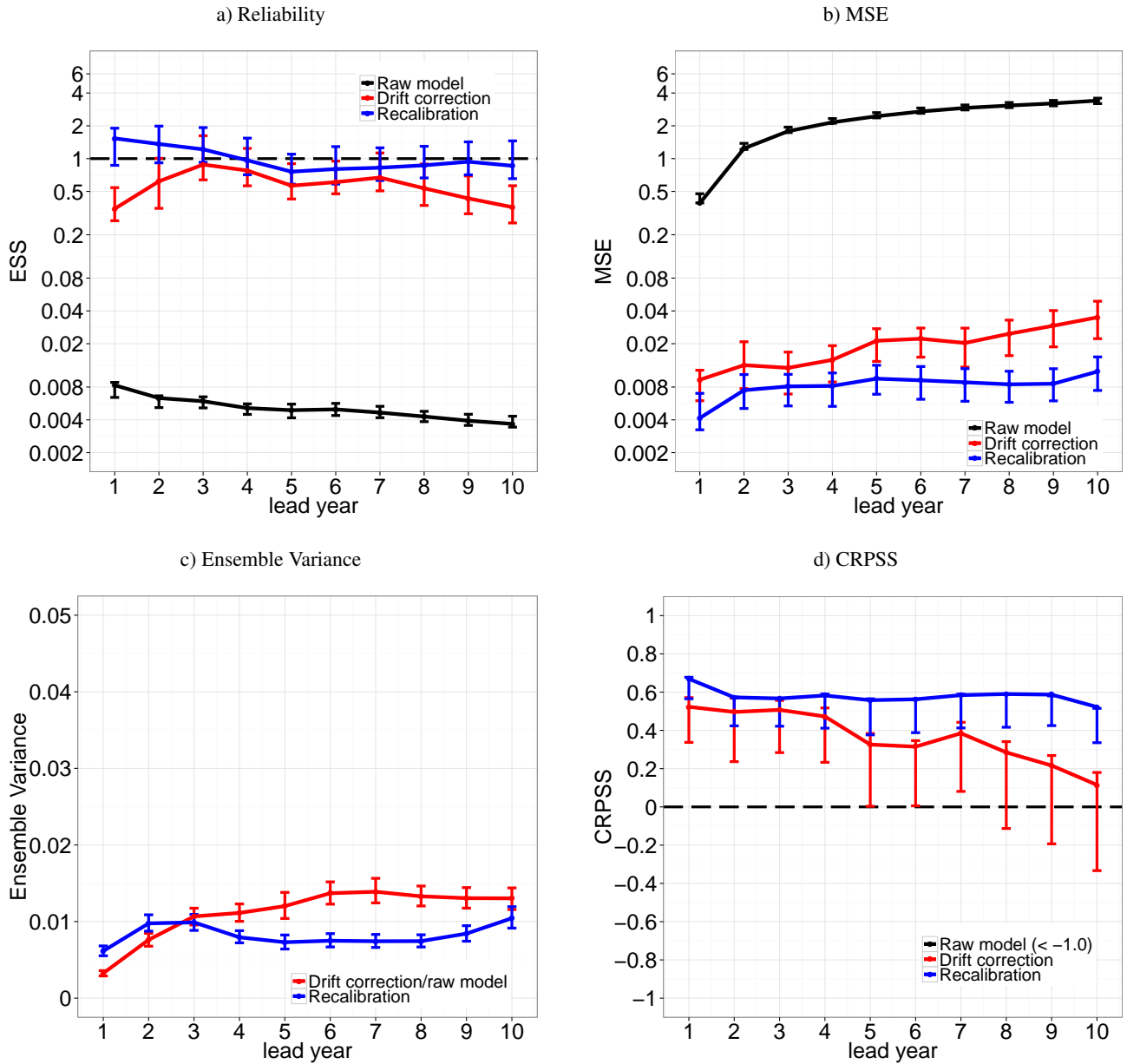
**Figure 5.** Temporal evolution of North Atlantic yearly mean surface temperature from MiKlip Prototype before (a) and (b) after recalibration with *DeFoReSt*. Shown are different start years with 5 year intervals (colored lines). The black line represents the surface temperature anomalies of NCEP 20CR. Units are in Kelvin [K].



**Figure 6.** Reliability (a), MSE (b), Ensemble Variance (c), and CRPSS (d) of surface temperature over the North Atlantic without any correction (black line), after drift correction (red line) and recalibration with *DeFoReSt* (blue line). The CRPSS for the raw forecasts (black line) is smaller than -1 and therefore not shown. As the drift correction method does not account for the ensemble spread, the ensemble variance of the raw model and the drift corrected forecast are equal. The vertical bars show the 95% confidence interval due 1000-wise bootstrapping.



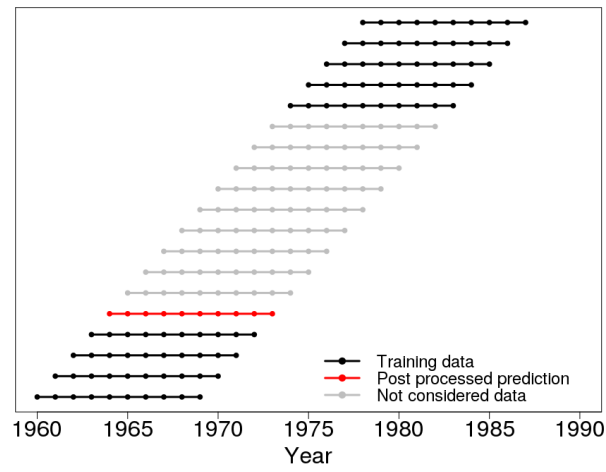
**Figure 7.** Temporal evolution of global yearly mean surface temperature from MiKlip Prototype (a) before and (b) after recalibration with *DeFoReSt*. Shown are different start years with 5 year intervals (colored lines). The black line represents the surface temperature anomalies of NCEP 20CR. Units are in Kelvin [K].



**Figure 8.** Reliability (a), MSE (b), Ensemble Variance (c), and CRPSS (d) of global mean surface temperature without any correction (black line), after drift correction (red line) and recalibration with *DeFoReSt* (blue line). The CRPSS for the raw forecasts (black line) is smaller than -1 and therefore not shown. As the drift correction method does not account for the ensemble spread, the ensemble variance of the raw model and the drift corrected forecast are equal. The vertical bars show the 95% confidence interval due 1000-wise bootstrapping.

	l=0	l=1	l=2	l=3	l=4	l=5	l=6	l=7
$a_l$	-0.61	0.0025	0.29	-0.00046	-0.11	0.0011	0.021	-0.00029
$b_l$	0.13	0.006	0.23	-0.0027	-0.12	0.00097	0.025	-0.000197
$w_l$	0.3	0	0.1	0.0014	0.01	0.0001	0	0

**Table A1.** Overview of the values coefficients  $a_l$ ,  $b_l$  and  $w_l$ .



**Figure A1.** Schematic overview of the applied cross-validation procedure for a decadal climate prediction, initialized in 1964 (red dotted line). All hindcasts which are initialized outside the prediction period are used as training data (black dotted lines). A hindcast which is initialized inside the prediction period is not used for training (gray dotted lines).