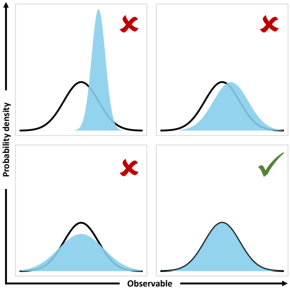


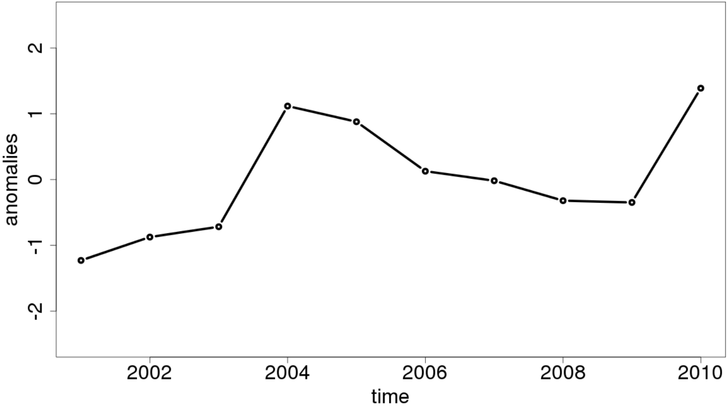
Calibration of decadal ensemble predictions

A. Pasternack, H. W. Rust, U. Ulbrich, M. A. Liniger, J. Bhend
Freie Universität Berlin

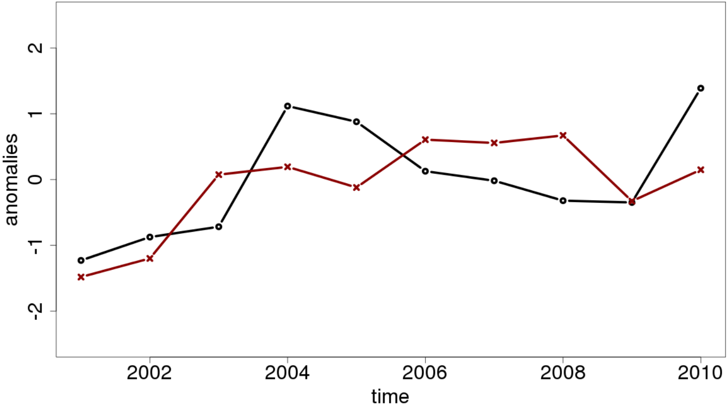
Berlin, October, 5th, 2016

- Probabilistic forecasts
- What is a good forecast
- Re-calibrating an example forecast
- Tailor re-calibration methods to decadal predictions
- Apply re-calibration methods to decadal predictions
  - Validation

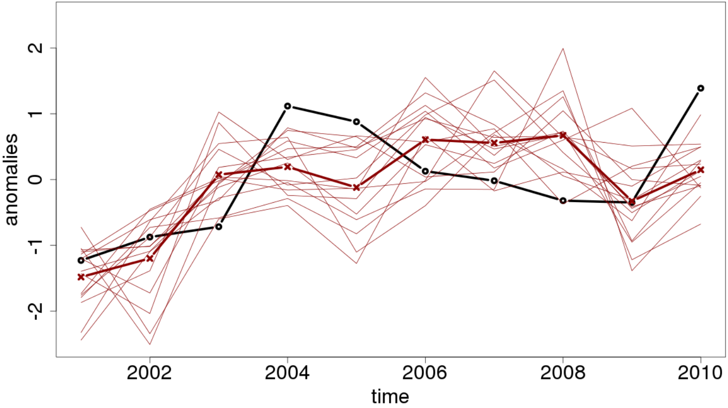
Probabilistic forecast



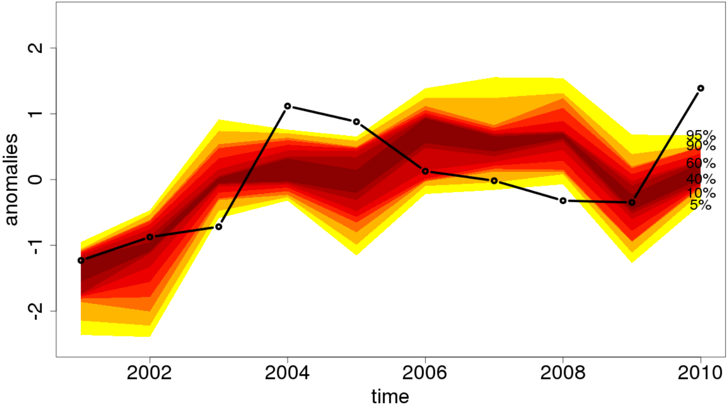
Probabilistic forecast



Probabilistic forecast



Probabilistic forecast



What is a good probabilistic forecast?

„... an important goal is to maximize <u>sharpness</u> without sacrificing <u>calibration</u>.“ (Wilks, 2011; Gneiting, 2007; Murphy and Winkler, 1987)

What is a good probabilistic forecast?

„... an important goal is to maximize sharpness without sacrificing calibration." (Wilks, 2011; Gneiting, 2007; Murphy and Winkler, 1987)

Sharpness

Forecasts take a risk, i.e. are frequently different from the climatological value?

„... an important goal is to maximize sharpness without sacrificing calibration.“ (Wilks, 2011; Gneiting, 2007; Murphy and Winkler, 1987)

Sharpness

Forecasts take a risk, i.e. are frequently different from the climatological value?

Calibration or reliability

Probabilistic forecasts „mean what they say“, e.g. for days with a forecast of 30% chance of rain, we expect a relative frequency of 30% rainy days.

What is a good probabilistic forecast?

„... an important goal is to maximize sharpness without sacrificing calibration." (Wilks, 2011; Gneiting, 2007; Murphy and Winkler, 1987)

Sharpness

Forecasts take a risk, i.e. are frequently different from the climatological value?

Calibration or reliability

Probabilistic forecasts „mean what they say", e.g. for days with a forecast of 30% chance of rain, we expect a relative frequency of 30% rainy days.

„Ensemble members are reliable if the MSE between the ensemble mean and observations is identical to the time mean intra-ensemble variance." (Palmer et al., 2006)

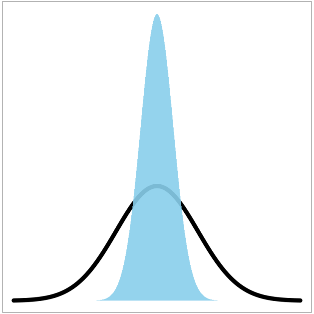
What About the Reliability of our Forecasts?

„... ensemble distributions typically underestimate the true forecast uncertainty and tend to be overconfident ...“ (e.g. Weigel, 2009)

What About the Reliability of our Forecasts?

„... ensemble distributions typically underestimate the true forecast uncertainty and tend to be overconfident ...“ (e.g. Weigel, 2009)

Observed relative frequency distribution is broader than forecasted distribution

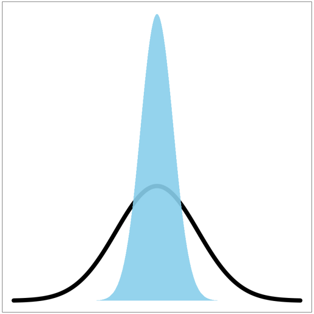


What About the Reliability of our Forecasts?

„... ensemble distributions typically underestimate the true forecast uncertainty and tend to be overconfident ...“ (e.g. Weigel, 2009)

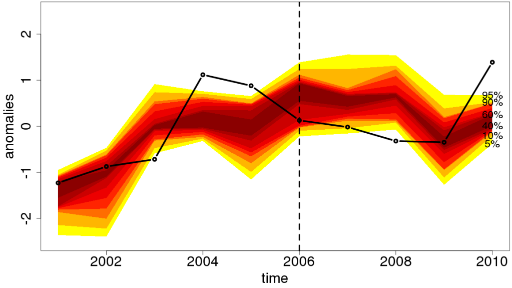
Observed relative frequency distribution is broader than forecasted distribution

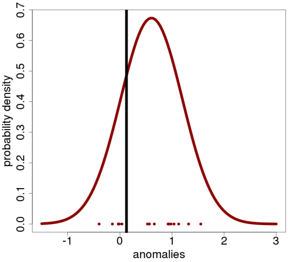
⟶ **adjust ensemble spread**



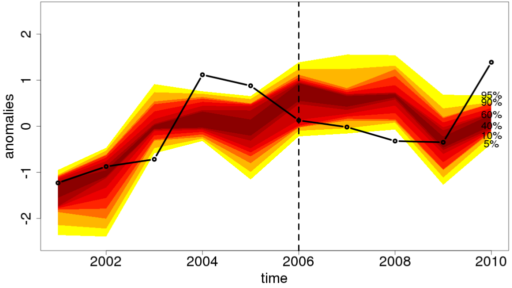
Re-calibrating an example forecast

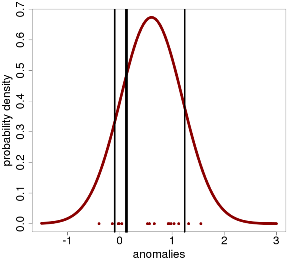
Probabilistic forecast



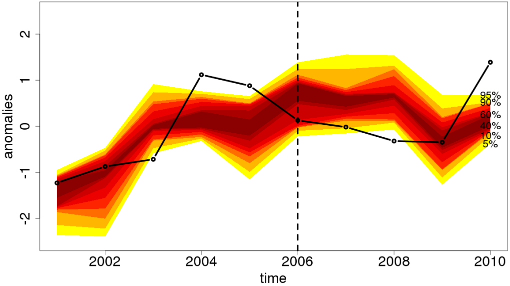


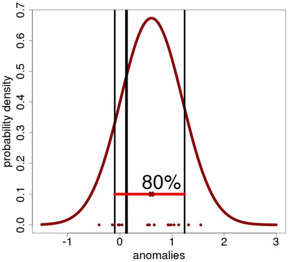
Probabilistic forecast



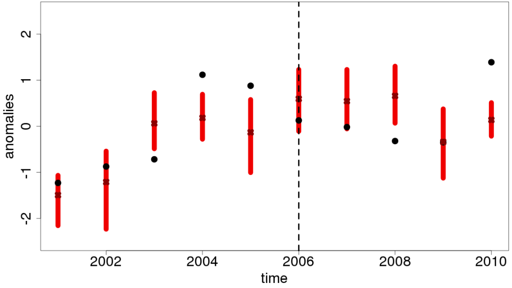


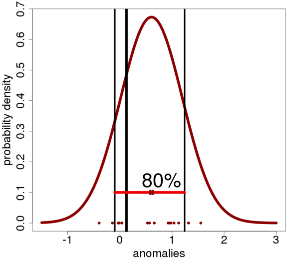
Probabilistic forecast





Probabilistic forecast



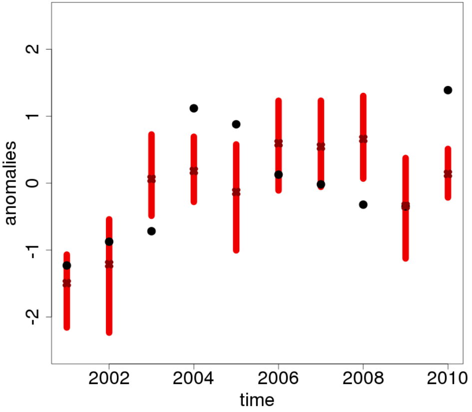


Reliable forecast:

- 80% of ens. spread should include 80% of observations
- only 50% are covered

What is wrong?

Re-calibrating an example forecast

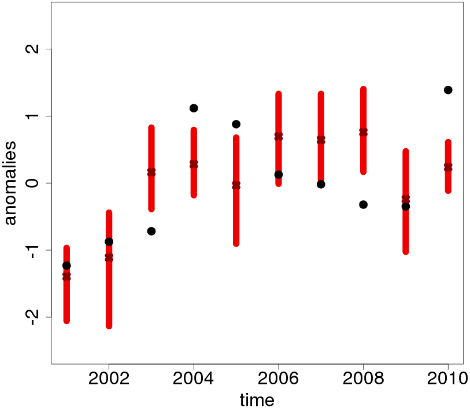


1. forecast is biased (uncond.)

Ensemble prediction:

$f_i(t) = \mu(t) + \epsilon_i(t) \qquad i = 1 \ldots M$

Re-calibrating an example forecast



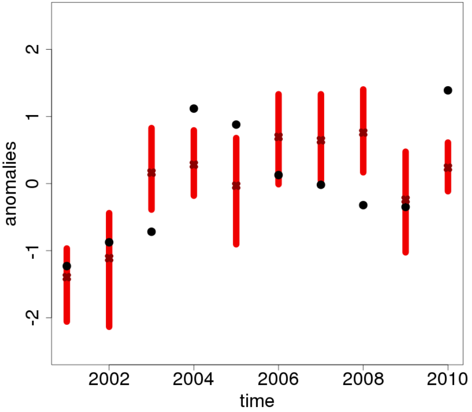
1. forecast is biased (uncond.) (shift ensemble mean)

2. forecast is conditionally biased

Ensemble prediction:

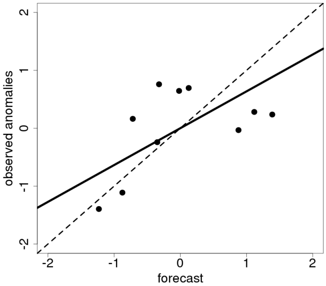
$f_i(t) = \alpha + \mu(t) + \epsilon_i(t) \qquad i = 1 \ldots M$

Re-calibrating an example forecast



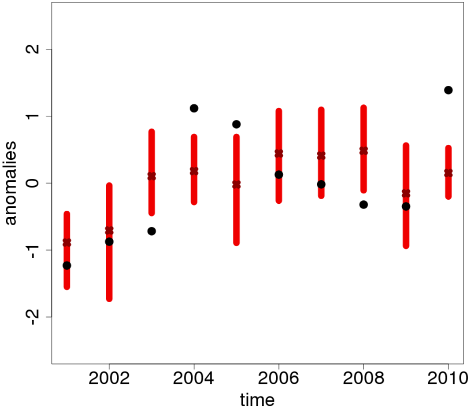
1. forecast is biased (uncond.)
(shift ensemble mean)

2. forecast is conditionally biased



Ensemble prediction:

$$f_i(t) = \alpha + \mu(t) + \epsilon_i(t) \qquad i = 1 \ldots M$$



1. forecast is biased (uncond.)
(shift ensemble mean)

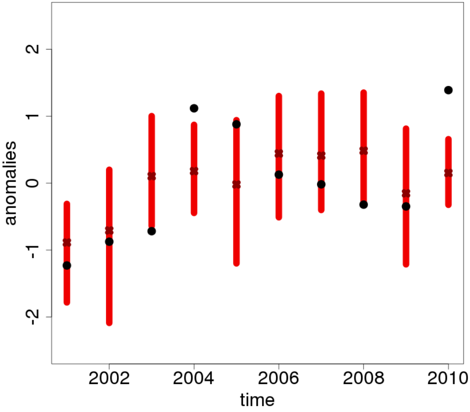
2. forecast is conditionally biased
(scale ensemble mean)

3. forecast is not reliable

Ensemble prediction:

$$f_i(t) = \alpha + \beta\mu(t) + \epsilon_i(t) \qquad i = 1 \ldots M$$

Re-calibrating an example forecast



1. forecast is biased (uncond.)
(shift ensemble mean)

2. forecast is conditionally biased
(scale ensemble mean)

3. forecast is not reliable
(scale ensemble spread)

4. re-calibrated forecast

Re-calibrated ensemble:

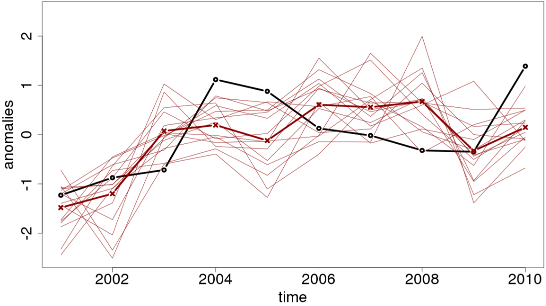
$$f_i^{Cal}(t) = \alpha + \beta\mu(t) + \gamma\epsilon_i(t) \qquad i = 1 \ldots M$$

Re-calibrating an example forecast

State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction



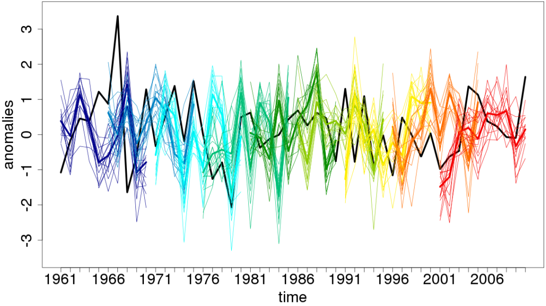
State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction

Tasks for CALIBRATION:

Tailor re-calibration methods to decadal predictions



State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction

Tasks for CALIBRATION:

Tailor re-calibration methods to decadal predictions

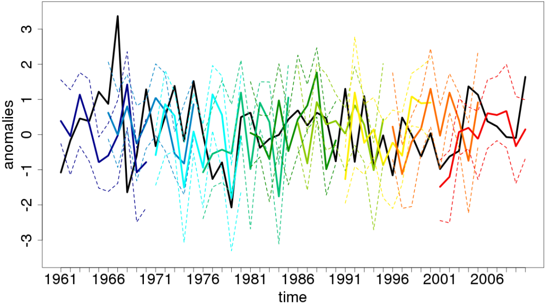
- limited number of hindcasts

Re-calibrating an example forecast

State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction



Tasks for CALIBRATION:

Tailor re-calibration methods to decadal predictions

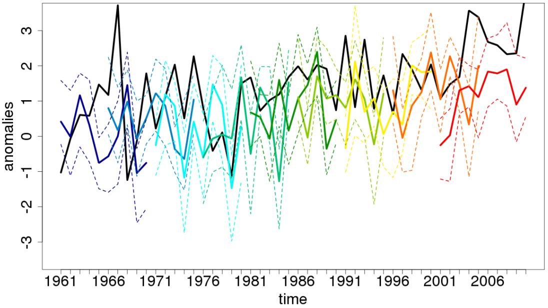
- limited number of hindcasts

Re-calibrating an example forecast

State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction



Tasks for CALIBRATION:

Tailor re-calibration methods to decadal predictions

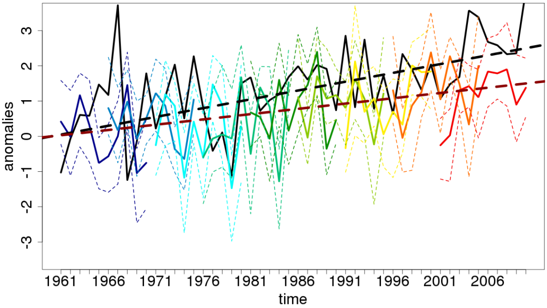
- limited number of hindcasts
- climate trend

Re-calibrating an example forecast

State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction



Tasks for CALIBRATION:

Tailor re-calibration methods to decadal predictions

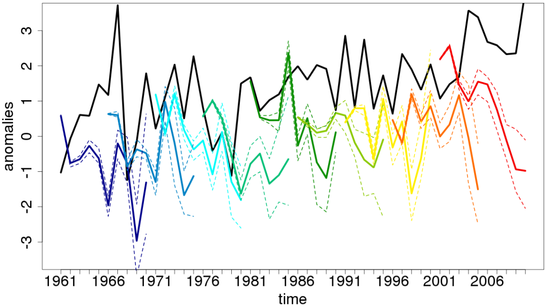
- limited number of hindcasts
- climate trend

Re-calibrating an example forecast

State of the art:

Re-calibration is used in

- weather prediction
- seasonal prediction



Tasks for CALIBRATION:

Tailor re-calibration methods to decadal predictions

- limited number of hindcasts
- climate trend
- dependence on lead years (drift)

Ensemble prediction

$$f_i(t,\tau) = \mu(t,\tau) + \epsilon_i(t,\tau)$$

$i = 1...M$ ensemble member, $t =$ start year, $\tau =$ lead year

with

$$\mu(t,\tau) = E(f_i(t,\tau))$$

Re-calibrated ensemble

$$f_i^{Cal}(t,\tau) = \alpha(t,\tau) + \beta(t,\tau)\mu(t,\tau) + \gamma(t,\tau)\epsilon_i(t,\tau)$$

find $\alpha(t,\tau)$, $\beta(t,\tau)$ and $\gamma(t,\tau)$ such that the ensemble is perfectly calibrated with maximum sharpness

1) α: bias and drift, 2) β: conditional bias, 3) γ: ensemble spread

General idea

Minimize continuous ranked probability score (crps) between model f^{Cal} and observation O (*Gneiting et al. 2005*):

General idea

Minimize continuous ranked probability score (crps) between model f^{Cal} and observation O (*Gneiting et al. 2005*):

$$crps(F, o) = \int_{-\infty}^{\infty} (F_{f^{Cal}}(y) - F_0(y))^2 dy,$$

General idea

Minimize continuous ranked probability score (crps) between model f^{Cal} and observation O (*Gneiting et al. 2005*):

$$crps(F, o) = \int_{-\infty}^{\infty} (F_{f^{Cal}}(y) - F_0(y))^2 dy,$$

$F_{f^{Cal}}(y)$: CDF derived from f^{Cal}

$F_0(y) = \begin{cases} 0 & y < o \\ 1 & y \geq o \end{cases}$: o is an arbitrary threshold (e.g. observation)

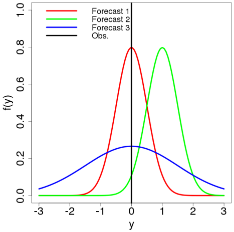
General idea

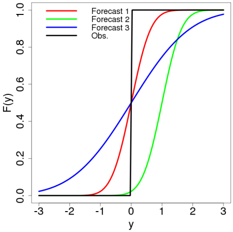
Minimize continuous ranked probability score (crps) between model f^{Cal} and observation O (*Gneiting et al. 2005*):

$$crps(F, o) = \int_{-\infty}^{\infty} (F_{f^{Cal}}(y) - F_0(y))^2 dy,$$

$F_{f^{Cal}}(y)$: CDF derived from f^{Cal}

$F_0(y) = \begin{cases} 0 & y < o \\ 1 & y \geq o \end{cases}$: o is an arbitrary threshold (e.g. observation)





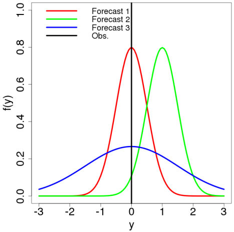
General idea

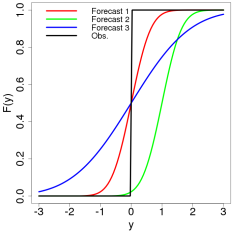
Minimize continuous ranked probability score (crps) between model f^{Cal} and observation O (*Gneiting et al. 2005*):

$$crps(F,o) = \int_{-\infty}^{\infty} (F_{f^{Cal}}(y) - F_0(y))^2 dy,$$

$F_{f^{Cal}}(y)$: CDF derived from f^{Cal}

$F_0(y) = \begin{cases} 0 & y < o \\ 1 & y \geq o \end{cases}$: o is an arbitrary threshold (e.g. observation)





$crps_1 = 0.12$, $crps_2 = 0.73$, $crps_3 = 0.35$

crps measures:

- reliability
- sharpness

General idea (2)

If the forecast distribution is Gaussian:

$$\longrightarrow f_i^{Cal}(t,\tau) \sim \mathcal{N}(\alpha(t,\tau) + \beta(t,\tau)\mu(t,\tau), \gamma(t,\tau)^2\sigma^2(t,\tau))$$

General idea (2)

If the forecast distribution is Gaussian:

$$\longrightarrow f_i^{Cal}(t,\tau) \sim \mathcal{N}(\alpha(t,\tau) + \beta(t,\tau)\mu(t,\tau), \gamma(t,\tau)^2\sigma^2(t,\tau))$$

...the crps simplifies to:

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

$\mu =$ens. mean, $\sigma =$ens. std., $o =$observation, $\Phi,\varphi =$ CDF and PDF of stand. norm. distr.

General idea (2)

If the forecast distribution is Gaussian:

$$\longrightarrow f_i^{Cal}(t,\tau) \sim \mathcal{N}(\alpha(t,\tau) + \beta(t,\tau)\mu(t,\tau), \gamma(t,\tau)^2\sigma^2(t,\tau))$$

...the crps simplifies to:

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

$\mu =$ens. mean, $\sigma =$ens. std., $o =$observation, $\Phi,\varphi =$ CDF and PDF of stand. norm. distr.

The average score over all k pairs of forecasts and observations is:

$$\Gamma(\mathcal{N}(\alpha + \beta\mu, \gamma^2\sigma^2), o) = \frac{1}{k}\sum_{j=1}^{k}\sqrt{\gamma^2\sigma_j^2}\{Z_j[2\Phi(Z_j) - 1] + 2\varphi(Z_j) - \frac{1}{\sqrt{\pi}}\},$$

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

General idea (2)

If the forecast distribution is Gaussian:

$$\longrightarrow f_i^{Cal}(t,\tau) \sim \mathcal{N}(\alpha(t,\tau) + \beta(t,\tau)\mu(t,\tau), \gamma(t,\tau)^2\sigma^2(t,\tau))$$

...the crps simplifies to:

$$crps(\mathcal{N}(\mu,\sigma^2),o) = \sigma\{\tfrac{o-\mu}{\sigma}[2\Phi(\tfrac{o-\mu}{\sigma}) - 1] + 2\varphi(\tfrac{o-\mu}{\sigma}) - \tfrac{1}{\sqrt{\pi}}\}$$

$\mu =$ens. mean, $\sigma =$ens. std., $o =$observation, Φ,$\varphi =$ CDF and PDF of stand. norm. distr.

The average score over all k pairs of forecasts and observations is:

$$\Gamma(\mathcal{N}(\alpha + \beta\mu, \gamma^2\sigma^2), o) = \frac{1}{k}\sum_{j=1}^{k}\sqrt{\gamma^2\sigma_j^2}\{Z_j[2\Phi(Z_j) - 1] + 2\varphi(Z_j) - \tfrac{1}{\sqrt{\pi}}\},$$

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

$\alpha = \alpha(t,\tau) = (a_0 + a_1 t) + (a_2 + a_3 t)\tau + (a_4 + a_5 t)\tau^2 + (a_6 + a_7 t)\tau^3 + \ldots$
$\beta = \beta(t,\tau) = (b_0 + b_1 t) + (b_2 + b_3 t)\tau + (b_4 + b_5 t)\tau^2 + (b_6 + b_7 t)\tau^3 + \ldots$
$\gamma = \gamma(t,\tau) = (c_0 + c_1 t) + (c_2 + c_3 t)\tau + (c_4 + c_5 t)\tau^2 + (c_6 + c_7 t)\tau^3 + \ldots$

$\longrightarrow$ find an $a_0, b_0, c_0, \ldots, a_7, b_7, c_7$ that minimize Γ

Example: Surface Temperature

Data:

- Surface temperature over the North Atlantic region
- Model: MPI-ESM-LR, Prototype (GECCO2)
- 15 ensemble members
- Initialisation years: 1961-2000
- Annual mean
- Reference: NCEP 20CR

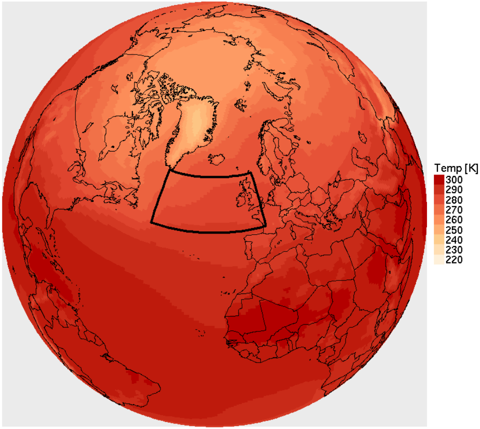
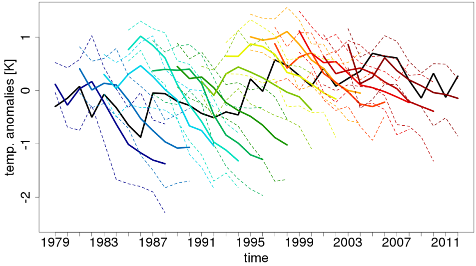


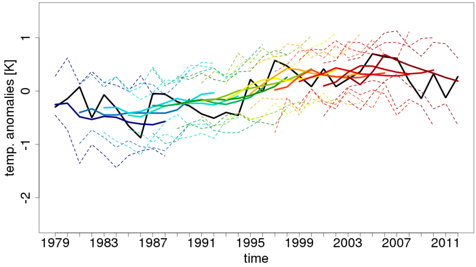
Figure: ST time mean for NCEP 20CR

Apply re-calibration method to decadal prediction

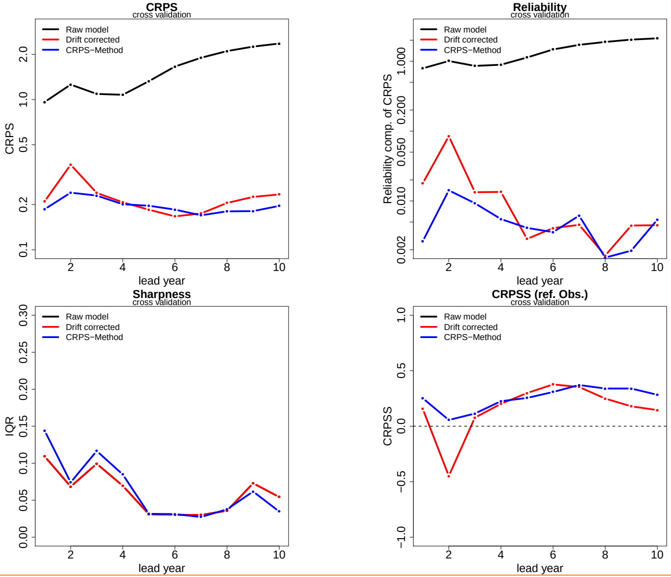
Before:



After:



Validation


CRPS
cross validation
Raw model
Drift corrected
CRPS−Method
CRPS
0.1
0.2
0.5
1.0
2.0
2
4
6
8
10
lead year
Reliability
cross validation
Raw model
Drift corrected
CRPS−Method
Reliability comp. of CRPS
0.002
0.010
0.050
0.200
1.000
lead year
Sharpness
cross validation
Raw model
Drift corrected
CRPS−Method
IQR
0.00
0.05
0.10
0.15
0.20
0.25
0.30
lead year
CRPSS (ref. Obs.)
cross validation
Raw model
Drift corrected
CRPS−Method
CRPSS
−1.0
−0.5
0.0
0.5
1.0
lead year

What is a good probabilistic forecast?

„... an important goal is to maximize sharpness without sacrificing reliability.“ (Wilks, 2011; Gneiting, 2007; Murphy and Winkler, 1987)

- CRPS minimization method by *Gneiting et al. 2005* addresses to reliability and sharpness for seasonal prediction.
- The developed extension to decadal predictions also includes a lead time dependent drift correction.

Validation

- CRPS method is mostly superior to drift correction and climatology w.r.t. predictive skill.
- Sharpness will be decreased to obtain good reliability.

Thank you for your attention!

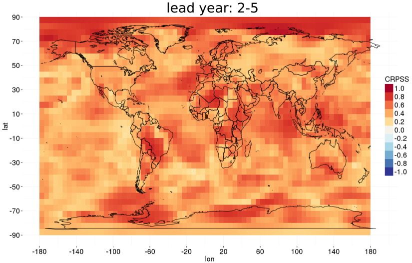
References

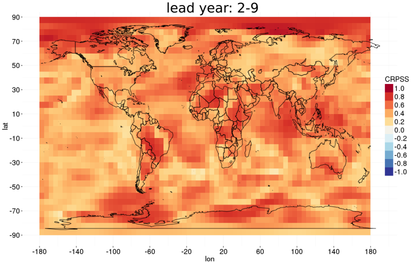
- Gneiting, T., A. E. Raftery, A. H.Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. Monthly Weather Review, 133, 1098–1118.
- Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102 (477), 359–378.
- Weigel, A., M. A. Liniger, and C. Appenzeller., 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? Mon. Weather Rev., 137(4), 1460–1479.
- Wilks,D.S.,2011:Statistical Methods in the Atmospheric Sciences.Academic Press.

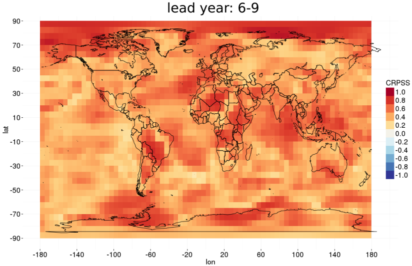
Appendix

Spatial Analyses (1)

CRPSS: Re-calibrated vs. Climatology (cross validation)







Spatial Analyses (2)

CRPSS: Re-calibrated vs. Drift Corrected (cross validation)