Forecast evaluation III

Thordis L. Thorarinsdottir
Norwegian Computing Center, Oslo, Norway
`www.nr.no/~thordis`

CUSO winter school 2021

Outline for this lecture

Assume we have a prediction $p \in \mathcal{P}$ and an observation $o \in \mathcal{O}$ where we wish to measure the skill of the prediction by applying a function

$$s : \mathcal{P} \times \mathcal{O} \longrightarrow \mathbb{R}$$

with a lower function value indicating a better skill.

- **What if we only care about a subset of the observations, e.g. the extremes?**
- **What if we are working in high dimensions?**
- **What if the observation is also given by a distribution?**

THE SPECTATOR

Forecast failure: how the Met Office lost touch with reality

Ideology has corrupted a valuable British institution

Rupert Darwall 13 July 2013 118 Comments

Verifying only the extremes erases propriety

Amisano and Giacomini (JBES, 2007) consider the restricted score

$$R^*(F, y) = -\mathbb{1}\{y \geq t\} \log f(y).$$

However, if $g(y) > f(y)$ for all $y \geq t$, then

$$\mathbb{E}R^*(G, y) < \mathbb{E}R^*(F, y)$$

independent of the true sampling density.

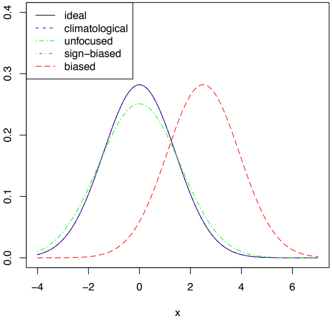
Indeed, if the forecaster's belief is F, his best prediction under R^* is

$$g(y) = \frac{f(y)}{\int_t^\infty f(x)dx} \mathbb{1}\{y \geq t\}$$

(Gneiting and Ranjan, JBES, 2011).

Demonstration by simulation

True data distribution: $G_t = N(\mu_t, 1)$ with $\mu_t \sim N(0, 1)$.

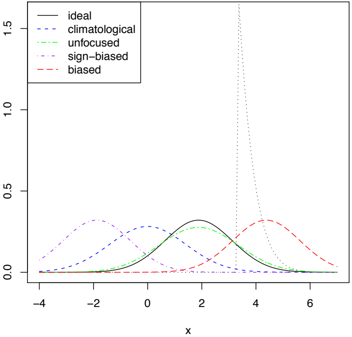


Forecaster	F_t
Ideal	$N(\mu_t, 1)$
Sign-biased	$N(-\mu_1, 1)$
Climatological	$N(0, 2)$
Unfocused	$\frac{1}{2}\{N(\mu_t, 1) + N(\mu_t + \tau_t, 1)\}$
Biased	$N(\mu_t + 2.5, 1)$

Here, $\tau_t = \pm 1$ with probability 1/2.

(Lerch et al., SS, 2015)

Results for y > 4.65 (99th precentile)



Forecaster	CRPS*	LogS*
Ideal	1.36	8.47
Sign-biased	5.01	16.87
Climatological	2.92	4.75
Unfocused	1.34	2.69
Biased	**0.55**	**1.38**

Better: Use threshold-weighted scoring rules

Diks *et al.* (JE, 2011) propose the **conditional likelihood score**

$$R(F, y) = -\omega(y) \log \left(\frac{f(y)}{\int \omega(x) f(x) dx} \right)$$

and the **cencored likelihood score**

$$R(F, y) = -\left[\omega(y) \log f(y) + (1 - \omega(y)) \log \left(1 - \int \omega(x) f(x) dx \right) \right].$$

Better: Use threshold-weighted scoring rules

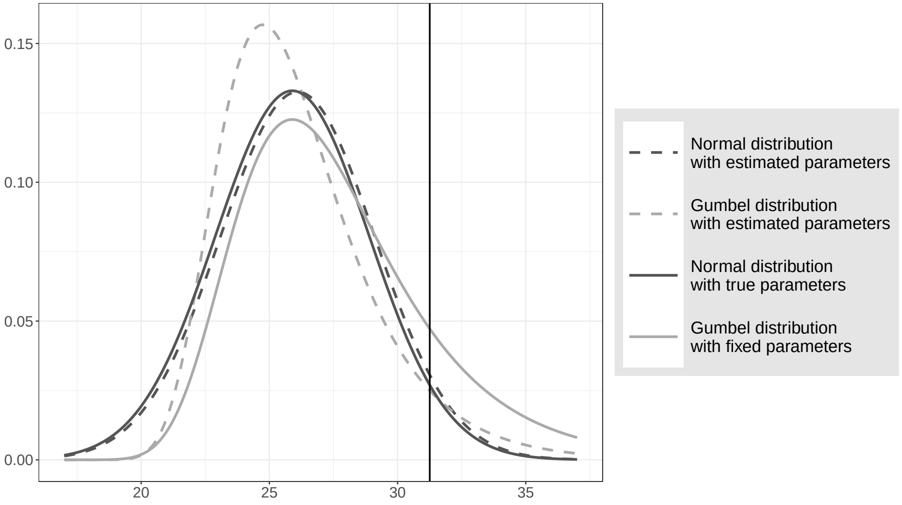
Gneiting and Ranjan (JBES, 2011) propose the **threshold weighted CRPS**

$$
\begin{aligned}
R(F,y) &= \int (F(x) - \mathbb{1}\{y \le x\})^2 \omega(x) dx \\
&= \int_0^1 \left(F^{-1}(\tau) - y\right)\left(\mathbb{1}\{y \le F^{-1}(\tau)\} - \tau\right)\omega(\tau)\mathrm{d}\tau.
\end{aligned}
$$

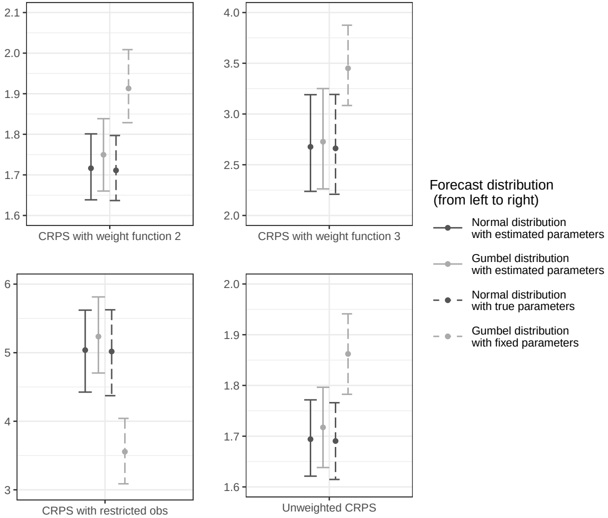
Here, we may e.g. set

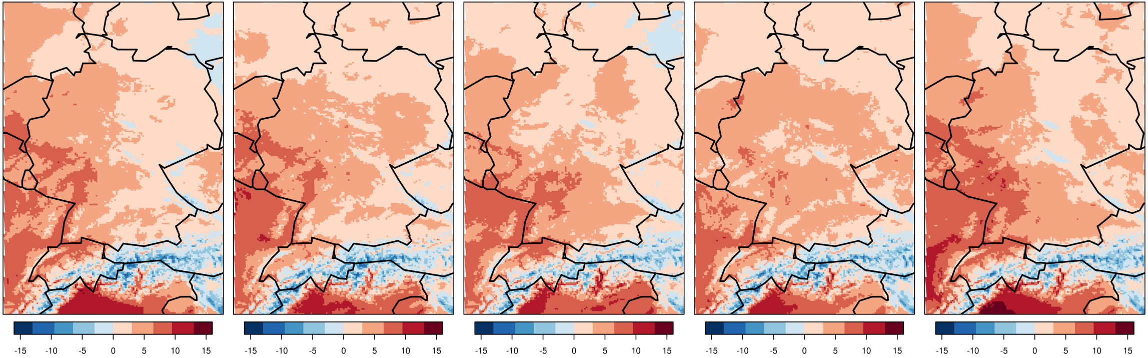
$$
\begin{aligned}
w_1(x) &= \mathbb{1}\{x \ge u\} \\
w_2(x) &= 1 + \mathbb{1}\{x \ge u\} \\
w_3(x) &= 1 + \mathbb{1}\{x \ge u\}\, u
\end{aligned}
$$

An extreme version of the example from last lecture



Results for 1 000 forecast-observation pairs




-15 -10 -5 0 5 10 15
-15 -10 -5 0 5 10 15
-15 -10 -5 0 5 10 15
-15 -10 -5 0 5 10 15
-15 -10 -5 0 5 10 15

Three scores for multivariate forecasts

1. The Dawid-Sebastiani (DS) score

$$\mathrm{R}(F, y) = \log\det\Sigma_F + (y - \mu_F)^\top \Sigma_F^{-1}(y - \mu_F)$$

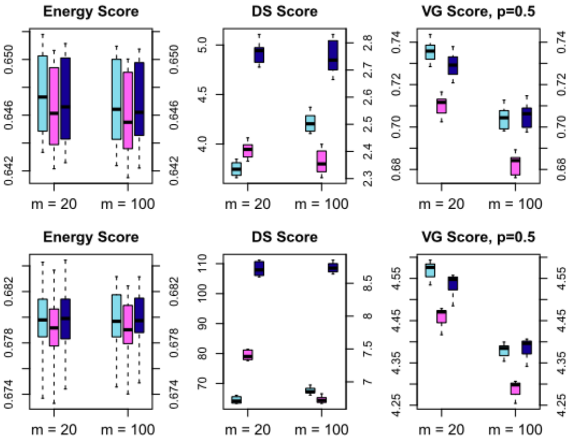
2. The energy score (ES)

$$\mathrm{R}(F, y) = \mathbb{E}_F\|X - y\| - \frac{1}{2}\mathbb{E}_F\|X - X'\|$$

3. The variogram score

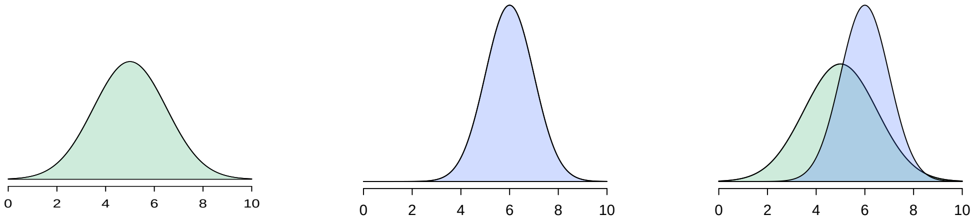
$$\mathbb{R}_p(F, y) = \sum_{i=1}^{d}\sum_{j=1}^{d}\omega_{ij}\left(|y_i - y_j|^p - \mathbb{E}_F|X_i - X_j|^p\right)^2$$

ES lacks discrimination; DS hard to estimate



Too weak (light blue), adequate (violet) and too strong (dark blue) correlation in 5 (top) and 15 (bottom) dimensions (Scheuerer and Hamill, MWR, 2015)

(a) Forecast (b) Observation (c) Comparison



Propriety condition for divergences

Two distributions may be compared using a **divergence function**,

$$d : \mathcal{F} \times \mathcal{F} \to [0, \infty], \qquad d(F, F) = 0 \quad \forall F \in \mathcal{F}.$$

Definition (Thorarinsdottir, Gneiting and Gissibl, 2013)

Let $Y_1, \ldots, Y_k \sim G$ and G_k be the corresponding empirical CDF. A divergence function d is *k-proper* if

$$\mathbb{E}\, d(G, G_k) \leq \mathbb{E}\, d(F, G_k).$$

Similarly, d is *asymptotically proper* if

$$\lim_{k \to \infty} \mathbb{E}\, d(G, G_k) \leq \lim_{k \to \infty} \mathbb{E}\, d(F, G_k),$$

for all $F, G \in \mathcal{F}$.

Many well known distances don't fulfill this condition

The area validation metric is given by

$$d(F, G) = \int |F(t) - G(t)| \mathrm{d}t$$

Let $G \sim \mathcal{U}([0, 1])$ and F_k discrete with probability mass $1/k$ in $x = i/(k + 1)$ for $i = 1, \ldots, k$. Then

$$\frac{1}{4} = \mathbb{E}_G d(F_1, \hat{G}_1) < \mathbb{E} d(G, \hat{G}_1) = \frac{1}{3}.$$

Similar example can be constructed for the Kolmogorov-Smirnov distance

$$d(F, G) = \sup_{t \in \mathbb{R}} |F(t) - G(t)|.$$

Every proper scoring rule defines a k-proper divergence function

Theorem (Thorarinsdottir, Gneiting and Gissibl, 2013)

Assume that $\mathrm{R}(G, G) \neq +\infty$ *and let*

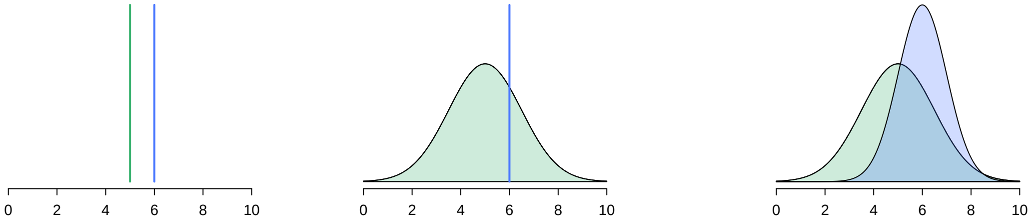
$$d(F, G) = \mathrm{R}(F, G) - \mathrm{R}(G, G),$$

where R *is a proper scoring rule. Then* d *is* k*-proper for all* $k = 1, 2, \ldots$.

Note that

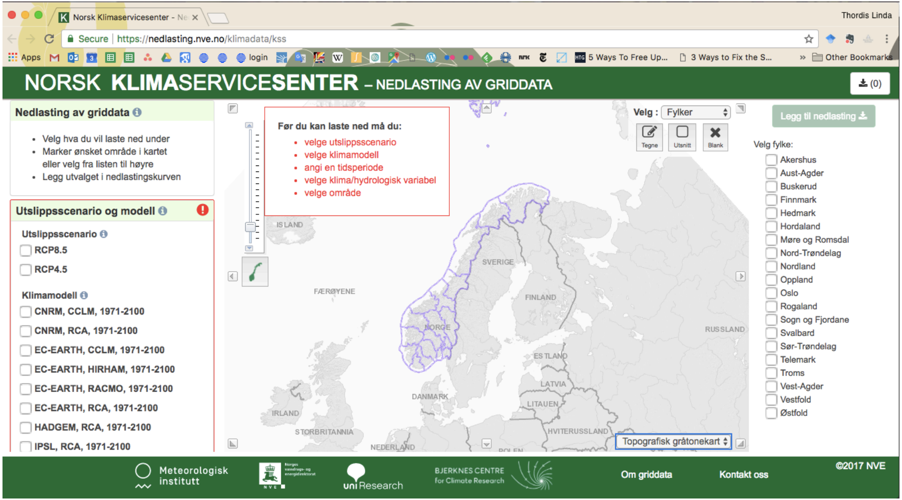
- $d(F_m, G_k)$ and $\frac{1}{k}\sum_i^k \mathrm{R}(F_m, y_i)$ will result in the same ranking of $F_1, \ldots, F_M$.
- it holds that $d(G, G) = 0$, while $\mathrm{R}(G, G)$ might depend on G.

Examples

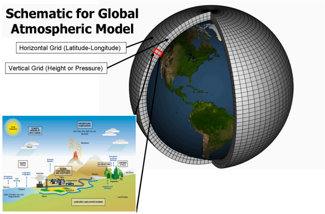


$(x - y)^2$	$(\mathrm{mean}(F) - y)^2$	$(\mathrm{mean}(F) - \mathrm{mean}(G))^2$
$\lvert x - y \rvert$	$\int [F(t) - \mathbb{1}\{t \geq y\}]^2 dt$	$\int [F(t) - G(t)]^2 dt$
	$-\log(f(y))$	$\int g(u) \log \frac{g(u)}{f(y)} d\lambda(u)$

A practical example: Climate services



Climate models and climate projections



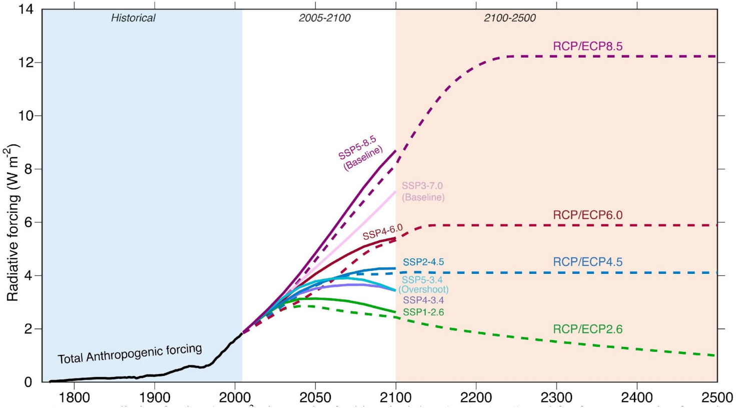


Figure on the right from IPCC.

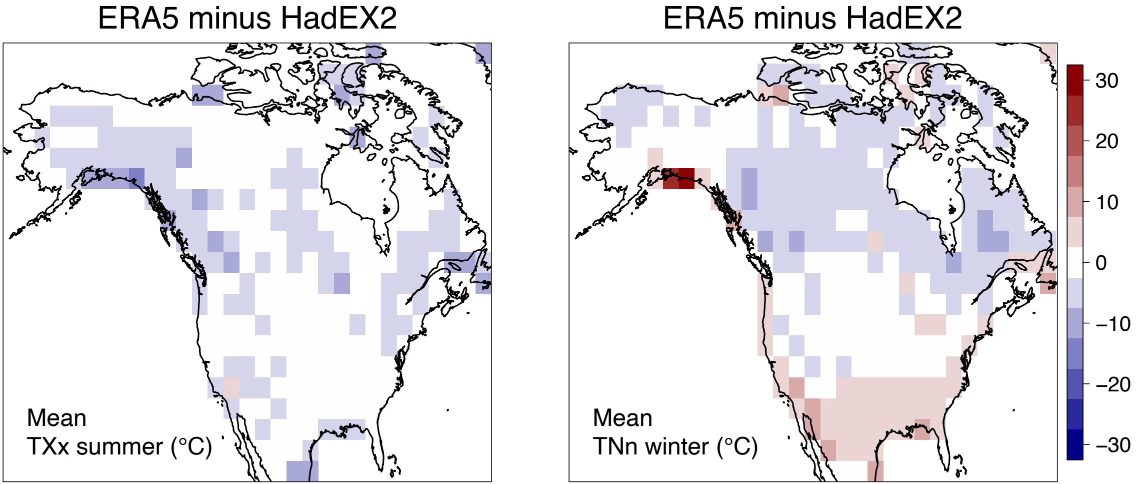
How to evalute climate predictions/projections?

Climate models are difficult to compare to data. Often climatologists compute some summary statistic (...) and compare climate models using observed (or rather estimated) forcings to the observed (or rather estimated) temperatures.

(...) it seems more appropriate to compare the distribution (over time and space) of climate model output to the corresponding distribution of observed data.

Guttorp (E, 2011)

Which is the better truth, model or data?

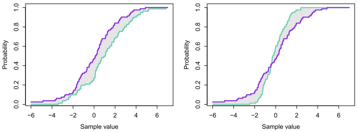


min. = −14.96, max. = 3.23,
mean = −2.42, MAE = 2.65

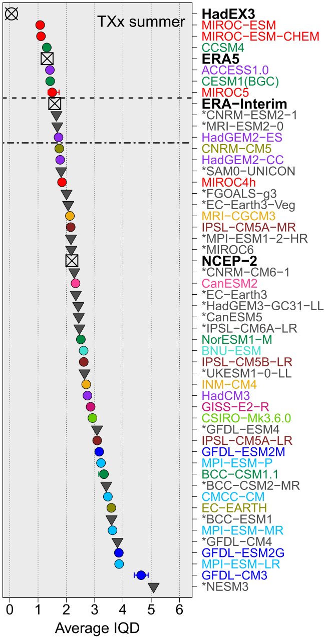
min. = −9.45, max. = 27.87,
mean = −0.85, MAE = 3.91

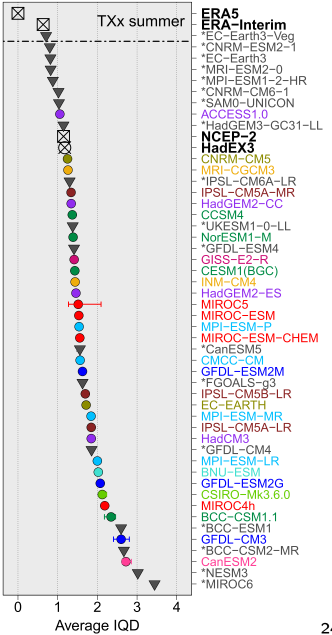
We use the IQD

$$d(F,G) = \int (F(t) - G(t))^2 \mathrm{d}t$$



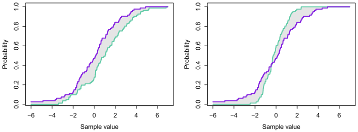
(T et al., ERL, 2020)



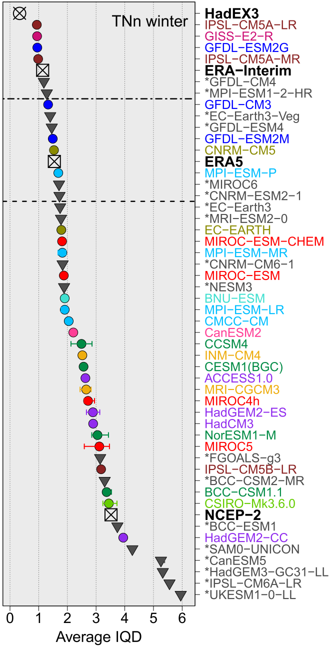


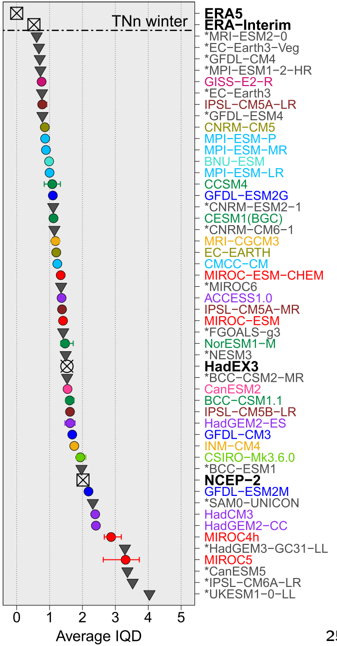
We use the IQD

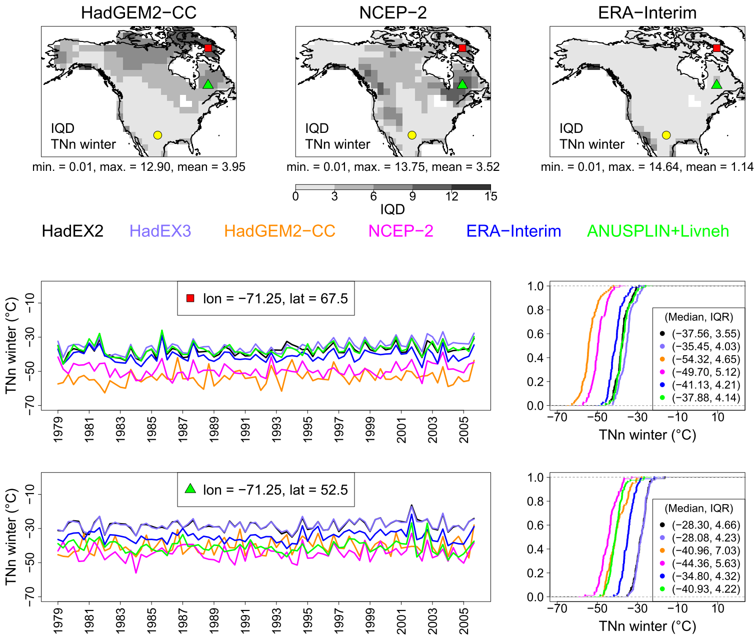
$$d(F,G) = \int (F(t) - G(t))^2 \mathrm{d}t$$



(T et al., ERL, 2020)






HadGEM2−CC
IQD
TNn winter
min. = 0.01, max. = 12.90, mean = 3.95
NCEP−2
IQD
TNn winter
min. = 0.01, max. = 13.75, mean = 3.52
ERA−Interim
IQD
TNn winter
min. = 0.01, max. = 14.64, mean = 1.14
0 3 6 9 12 15
IQD
HadEX2 HadEX3 HadGEM2−CC NCEP−2 ERA−Interim ANUSPLIN+Livneh
lon = −71.25, lat = 67.5
TNn winter (°C)
−10 −30 −50 −70
1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005
(Median, IQR)
(−37.56, 3.55)
(−35.45, 4.03)
(−54.32, 4.65)
(−49.70, 5.12)
(−41.13, 4.21)
(−37.88, 4.14)
1.0 0.8 0.6 0.4 0.2 0.0
−70 −50 −30 −10 10
TNn winter (°C)
lon = −71.25, lat = 52.5
TNn winter (°C)
−10 −30 −50 −70
1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005
(Median, IQR)
(−28.30, 4.66)
(−28.08, 4.23)
(−40.96, 7.03)
(−44.36, 5.63)
(−34.80, 4.32)
(−40.93, 4.22)
1.0 0.8 0.6 0.4 0.2 0.0
−70 −50 −30 −10 10
TNn winter (°C)

On the climate scale, we generally work with anomalies rather than absolute values

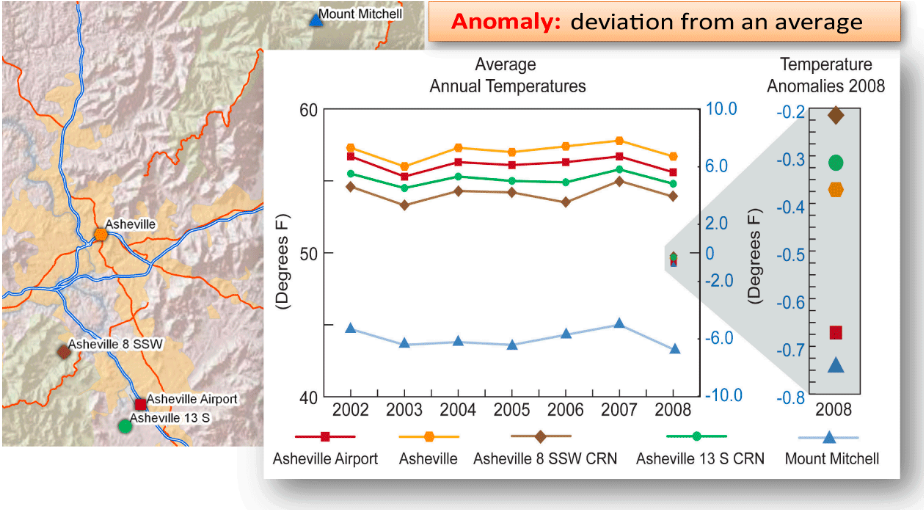
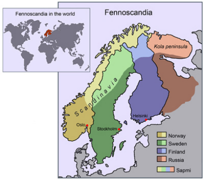
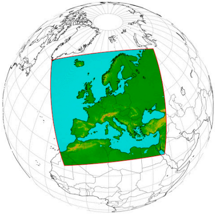
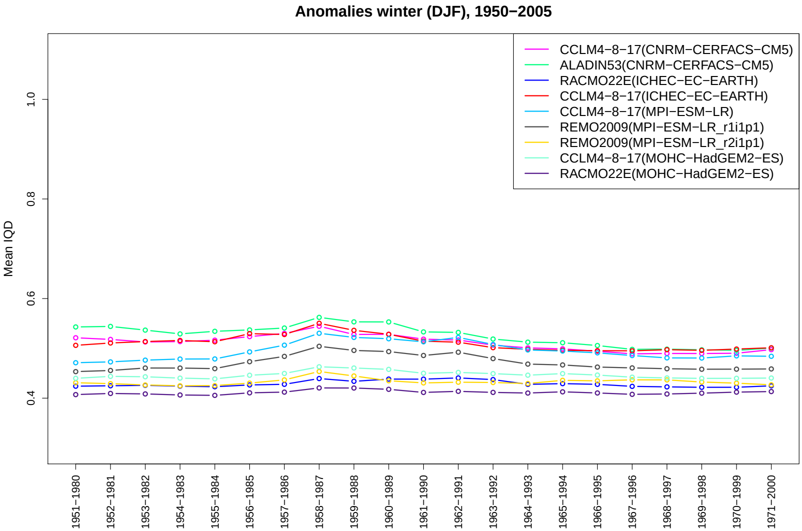


Figure from ncdc.noaa.gov.

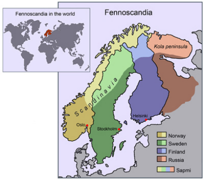
Standard reference periods are 30 years

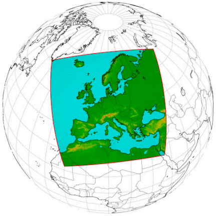


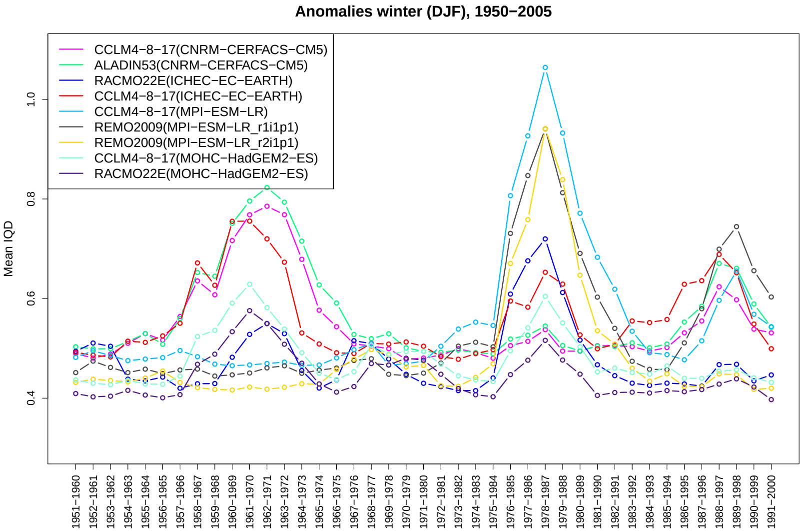




10 year reference periods result in unstable rankings

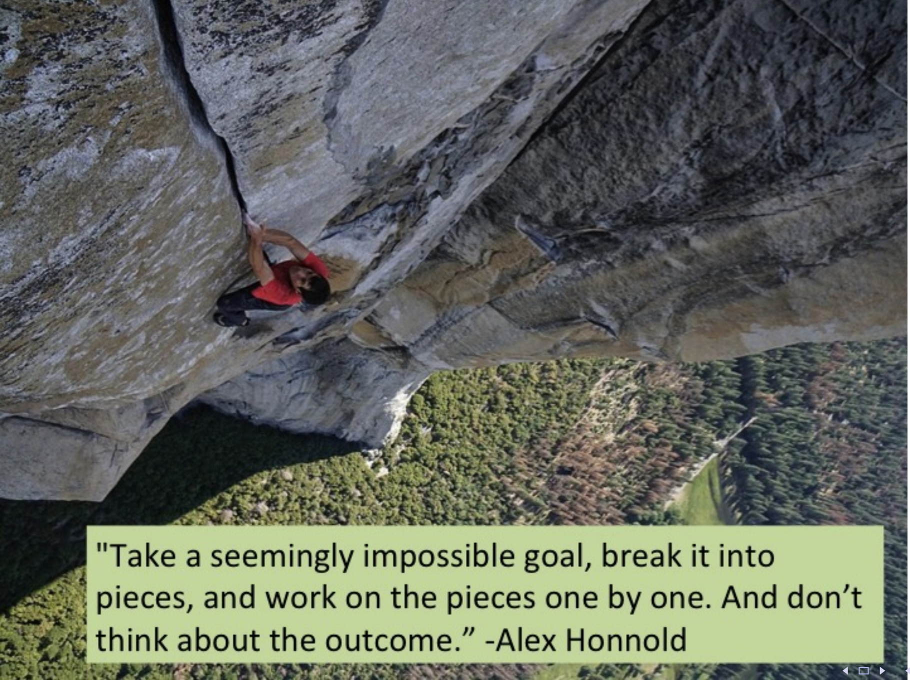






Conclusions

- From this morning: Performance measure should be selected with care, preferably used in groups.
- Forecaster's dilemma: Verification on extreme events only is bound to discredit skillful forecasters. The only remedy is to consider all available cases when evaluating the models.
- Careful application of weight functions can help interpreting prediction skill in certain regions of interest. In particular, the weighted versions of the CRPS share (almost all of) the desirable properties of the unweighted CRPS.
- Overall: The framework presented here provides a unified setting for comparing two values, a value and a distribution, or two distributions.


"Take a seemingly impossible goal, break it into pieces, and work on the pieces one by one. And don't think about the outcome." -Alex Honnold

References

1. A Amisano and R Giacomini (2007): Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, 25(2), 177-190.
2. T Gneiting and R Ranjan (2011): Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29(3), 411-422.
3. S Lerch, T Thorarinsdottir, F Ravazzolo and T Gneiting (2017): Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1), 106-127.
4. C Diks, V Panchenko and D van Dijk (2011): Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2), 215-230.
5. M Scheuerer and T Hamill (2015): Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4), 1321-1334.
6. T Thorarinsdottir, T Gneiting and N Gissibl (2013): Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1), 522-534.
7. T Thorarinsdottir, J Sillmann, M Haugen, N Gissibl and M Sandstad (2020): Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods. *Environmental Research Letters*, 15(12), 124041.